

# Global burden of vaccine-associated myocarditis and pericarditis, 1967-2023: a comprehensive analysis of the international pharmacovigilance database

Hyesu Jo<sup>1</sup>, Selin Woo<sup>1</sup>, Dong Keon Yon,<sup>1\*</sup>

<sup>1</sup>Center for Digital Health, Medical Science Research Institute, Kyung Hee University Medical Center, Kyung Hee University College of Medicine, Seoul, South Korea

\*Correspondence: Dong Keon Yon (yonkkang@gmail.com)

**Abstract—** This study examined the worldwide prevalence of vaccine-associated pericarditis and myocarditis using the WHO database, focusing on 19 vaccines across 156 countries. Of the 73,590 total reports, 49,096 were vaccine-related. Reports have increased significantly, especially before 2020, due to COVID-19 mRNA vaccines. Smallpox vaccines were most associated (ROR: 73.68; IC0.25: 5.91). Reports were more frequent in males and older age groups, with most reactions occurring within one day and 0.44% fatality rate. Our analysis indicates increased reports linked to vaccines, especially live ones like smallpox, requiring caution.

## I. INTRODUCTION

Vaccines are considered a key aspect of public health, reducing preventable deaths and disease prevalence and promoting herd immunity with the potential to eradicate diseases globally. However, it's crucial to identify and address potential side effects, specifically myocarditis and pericarditis, which can lead to cardiac fibrosis.

During the COVID-19 pandemic, concerns arose about myocarditis and pericarditis due to vaccination. Previous reports of these conditions following other vaccines were rare, so various efforts were made to determine the cause. Despite global vaccine hesitancy observed due to concerns about cardiac side effects from COVID-19 vaccines, data analysis for other vaccines remains limited.

Our research analyses all types of vaccines to investigate the frequency of adverse events based on the unique characteristics and mechanisms of each vaccine. The objective of this study is to objectively present the potential adverse effects following various vaccination regimes, an area scarcely addressed in existing literature. This endeavor is not merely an academic pursuit but a pivotal step in enhancing awareness within the medical and scientific communities.

## II. METHODS

### A. Patient selection and data collection

VigiBase is a global database managed by the WHO for Drug Monitoring Cooperation in Sweden. Data have been collected from its establishment in 1967 until June 2023, and the database is developed and maintained by the Uppsala Monitoring Center. This extensive repository comprises over 131,255,418 individual safety reports (ICSRs) related to

suspected adverse drug reactions (ADRs) from more than 156 countries actively participating in WHO international drug monitoring programs. Additionally, it encompasses a wide range of over 25,000 drugs, providing a comprehensive dataset for drug and vaccine safety testing. Despite potential heterogeneity in the data regarding the relationship between medications/drugs and reported ADRs, conducting extensive quantitative testing based on big data is crucial for efficient drug compliance. It is essential to acknowledge that the likelihood of suspected adverse events attributed to drugs varies across different cases. The Institutional Review Boards of Kyung Hee University Medical Center and the Uppsala Monitoring Center (WHO Collaborating Center) approved the use of confidential and electronically processed patient data.

### B. Main outcome

Vaccine-related pericarditis and myocarditis data were compiled from the beginning of 1969 until 2023, and vaccines were organized into 18 distinct categories: (1) anthrax vaccines; (2) diphtheria, tetanus toxoids, pertussis, polio, and Hemophilus influenza type b [DTaP-IPV-Hib] vaccines; (3) meningococcal vaccines; (4) pneumococcal vaccines; (5) typhoid vaccines; (6) encephalitis vaccines; (7) influenza vaccines; (8) hepatitis A vaccines; (9) hepatitis B vaccines; (10) measles, mumps, and rubella [MMR] vaccines; (11) rotavirus diarrhea vaccines; (12) zoster vaccines; (13) papillomavirus vaccines; (14) smallpox; (15) COVID-19 mRNA vaccines; (16) Ad5-vectored COVID-19 vaccines; (17) inactivated whole-virus COVID-19 vaccines; and (18) others (tuberculosis, brucellosis, plague, typhus, leptospirosis, rabies, yellow fever, Ebola, and dengue vaccines). Using the Medical Dictionary for Regulatory Activities (MedDRA) 25.0, we collected all the adverse events of deduplicated vaccinators around the world regarding the preferred terms: "Pericarditis" and "Myocarditis". Following the WHO causality assessment recommendations, all the vaccines were only considered "suspected" to compute the disproportional association with pericarditis and myocarditis.

### C. Covariates

In this study, we meticulously documented cases of suspected vaccine-related pericarditis and myocarditis, focusing on a comprehensive investigation. Our study primarily relied on ICSRs collected from diverse sources, including patients,

healthcare professionals, and pharmaceutical companies, within the post-market setting. The dataset encompassed patient demographics (i.e., age [0-11, 12-17, 18-44, 45-64,  $\geq$  65 years, and unknown] and sex), administrative information (i.e., reporting regions [African, America, South-East Asia, Europe, Eastern Mediterranean, and Western Pacific], reporting years [1967-2019 and 2020-2023], reporter qualifications [health professionals, non-health professionals, and unknown], and study categories [study-related, non-study-related, and unknown]). Information regarding vaccines (i.e., vaccine type and single suspected vaccine), and adverse drug reaction information (i.e., time to onset [TTO] of reaction and fatal outcomes [recovered/recovering, fatal, and unknown]). All voluntary reports indicated at least one suspected vaccine in pericarditis and myocarditis adverse events.

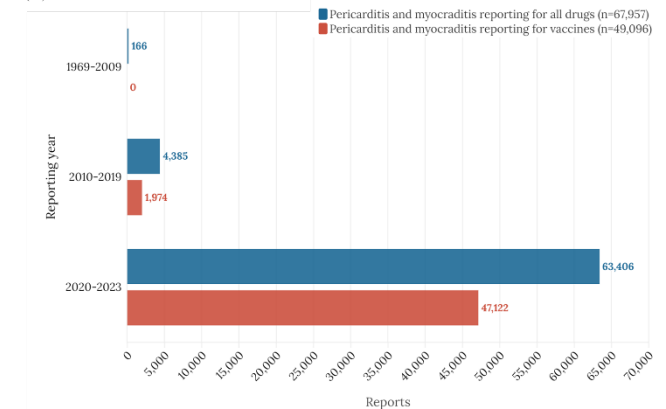
#### D. Statistical analyses

From the dataset, report and non-report groups were created, and each vaccine in VigiBase was subjected to disproportionality analysis to identify significant associations with pericarditis and myocarditis reports. Two common pharmacovigilance indicators, the information component (IC) and reporting odds ratio (ROR), were utilized for this analysis. The IC was calculated using a Bayesian method, comparing the event rates of pericarditis and myocarditis for a given vaccine with those of all other vaccines. When the entire database was not used as a comparator, sensitivity analyses were conducted, and ROR was used as the measure of disproportionality. The ROR is a frequentist measure derived from the number of adverse events and the contingency table of the vaccine. Statistical comparisons involved the unpaired Kruskal-Wallis test for continuous variables and the chi-squared test or Fisher's exact test for categorical variables. Statistical significance was determined with a two-sided p-value  $< 0.05$ . All analyses were conducted using SAS (version 9.4; SAS Inc., Cary, NC, USA).

### III. RESULTS

Of the 11,516,395 reports in the entire database, 49,096 reports (male,  $n=30,013$ ) of myocarditis and pericarditis were identified in the Vigibase from 1967 to 2023. The world is divided into six unique geographical regions, with the American region accounting for nearly half of the reports, followed by Europe, Western Pacific, Eastern Mediterranean, Africa, and Southeast Asia. The COVID-19 mRNA vaccine (90.42%) was the most associated with pericarditis and myocarditis reports, followed by the Ad5-vectored COVID-19 vaccine (4.44%) and smallpox vaccine (1.18%). An analysis of reports on vaccine-associated pericarditis and myocarditis revealed several vaccines implicated in these adverse incidents. Smallpox vaccines were most associated with pericarditis and myocarditis reports, followed by the

COVID-19 mRNA vaccine, anthrax vaccine, typhoid vaccine, encephalitis vaccine, influenza vaccine, and the Ad5-vectored COVID-19 vaccine. When considering the risk by age groups, a significant sex disproportion was observed in the age groups of 12-17 years with males being more associated with vaccine-associated pericarditis and myocarditis in every age group. An overall disproportion between the sexes was also observed. Aside from the COVID-19 vaccine, which makes up the majority of reports, the age group of 18-44 years accounts for 97.3% of anthrax vaccine, 91.7% of typhoid vaccine, and 91.2% of smallpox vaccine. In the case of the COVID-19 mRNA vaccine and Ad5-vectored COVID-19 vaccine, the age group of 18-44 years takes a proportion of 42.4% and 30.5% respectively. In respect of individual vaccines, anthrax and smallpox vaccines were more likely to be associated with pericarditis and myocarditis reports in the age group of 18-44 years. In the case of the COVID-19 mRNA vaccine and Ad5-vectored COVID-19 vaccine, they were more likely to be associated with the age group of 0-11 years and the age group over 65 years. The cumulative number of vaccine-associated pericarditis and myocarditis is shown in Figure 2. There was an extremely limited number of reports until the 2010s. However, since 2020, there has been a significant rise in reports of pericarditis and myocarditis associated with the COVID-19 vaccine, with the largest percentage of the COVID-19 mRNA vaccine.



**Figure 1.** Temporal trends of vaccine-associated pericarditis and myocarditis adverse events by continent.

### IV. CONCLUSIONS

Our study identified long-term trends in the prevalence of RA and OA over a 24-year period from 1998 to 2021, with a particular focus on the impact of the COVID-19 pandemic. The results showed a consistent decline in the prevalence of both RA and OA until the year 2020, followed by an increase in 2021. Notably, OA exhibited a higher prevalence among specific vulnerable groups, such as individuals over 60 years of age, urban residents, and those with a high education level, whereas RA did not show a particularly vulnerable population.

It would be beneficial for government policy researchers to devise personalized policies targeted at the vulnerable groups affected by OA. While our study did not find any evidence of a relationship between the COVID-19 pandemic and the prevalence of RA has yet been identified, additional follow-up studies based on our study findings would be helpful in further exploring this topic.

#### ACKNOWLEDGMENTS

This research was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT; RS-2023-00248157). The funders had no role in study design, data collection, data analysis, data interpretation, or writing of the report.

#### REFERENCES

1. Benn CS, Fisker AB, Rieckmann A, Sørup S, Aaby P. Vaccinology: time to change the paradigm? *Lancet Infect Dis* 2020;20:e274-e283.
2. Lee K, Lee H, Kwon R et al. Global burden of vaccine-associated anaphylaxis and their related vaccines, 1967-2023: A comprehensive analysis of the international pharmacovigilance database. *Allergy* 2023.
3. Nguyen LS, Cooper LT, Kerneis M et al. Systematic analysis of drug-associated myocarditis reported in the World Health Organization pharmacovigilance database. *Nat Commun* 2022;13:25.
4. Calcaterra G, Mehta JL, de Gregorio C et al. COVID 19 Vaccine for Adolescents. Concern about Myocarditis and Pericarditis. *Pediatr Rep* 2021;13:530-533.
5. Diaz GA, Parsons GT, Gering SK, Meier AR, Hutchinson IV, Robicsek A. Myocarditis and Pericarditis After Vaccination for COVID-19. *Jama* 2021;326:1210-1212.
6. Altman NL, Berning AA, Mann SC et al. Vaccination-Associated Myocarditis and Myocardial Injury. *Circ Res* 2023;132:1338-1357.
7. Ling RR, Ramanathan K, Tan FL et al. Myopericarditis following COVID-19 vaccination and non-COVID-19 vaccination: a systematic review and meta-analysis. *Lancet Respir Med* 2022;10:679-688.
8. Smith L, Shin JI, Hwang SY et al. Global Burden of Disease study at the World Health Organization: research methods for the most comprehensive global study of disease and underlying health policies. *Life Cycle* 2022;2:e8.
9. Lindquist M. VigiBase, the WHO Global ICSR Database System: Basic Facts. *Drug information journal : DIJ / Drug Information Association* 2008;42:409-419.
10. Kim MS, Jung SY, Ahn JG et al. Comparative safety of mRNA COVID-19 vaccines to influenza vaccines: A pharmacovigilance analysis using WHO international database. *J Med Virol* 2022;94:1085-1095.
11. Jung SY, Kim MS, Li H et al. Cardiovascular events and safety outcomes associated with remdesivir using a World Health Organization international pharmacovigilance database. *Clin Transl Sci* 2022;15:501-513.
12. Lee SW. Methods for testing statistical differences between groups in medical research: statistical standard and guideline of Life Cycle Committee. *Life Cycle* 2022;2:e1.
13. Ferchichi K, Aouinti I, Zaiem A et al. Myocarditis following Coronavirus vaccination. *Clin Immunol Commun* 2022;2:162-164.
14. Fairweather D, Cooper LT, Jr., Blauwet LA. Sex and gender differences in myocarditis and dilated cardiomyopathy. *Curr Probl Cardiol* 2013;38:7-46.
15. Heidecker B, Dagan N, Balicer R et al. Myocarditis following COVID-19 vaccine: incidence, presentation, diagnosis, pathophysiology, therapy, and outcomes put into perspective. A clinical consensus document supported by the Heart Failure Association of the European Society of Cardiology (ESC) and the ESC Working Group on Myocardial and Pericardial Diseases. *Eur J Heart Fail* 2022;24:2000-2018.
16. Hajjo R, Sabbah DA, Bardaweel SK, Tropsha A. Shedding the Light on Post-Vaccine Myocarditis and Pericarditis in COVID-19 and Non-COVID-19 Vaccine Recipients. *Vaccines (Basel)* 2021;9.
17. Szalay G, Meiners S, Voigt A et al. Ongoing coxsackievirus myocarditis is associated with increased formation and activity of myocardial immunoproteasomes. *Am J Pathol* 2006;168:1542-52.
18. Cosper PF, Harvey PA, Leinwand LA. Interferon- $\gamma$  causes cardiac myocyte atrophy via selective degradation of myosin heavy chain in a model of chronic myocarditis. *Am J Pathol* 2012;181:2038-46.
19. West K, Petrie L, Haines DM et al. The effect of formalin-inactivated vaccine on respiratory disease associated with bovine respiratory syncytial virus infection in calves. *Vaccine* 1999;17:809-20.
20. Wang WH, Wei KC, Huang YT, Huang KH, Tsai TH, Chang YC. The Incidence of Myocarditis Following an Influenza Vaccination: A Population-Based Observational Study. *Drugs Aging* 2023;40:145-151.
21. Awadalla M, Golden DLA, Mahmood SS et al. Influenza vaccination and myocarditis among patients receiving immune checkpoint inhibitors. *J Immunother Cancer* 2019;7:53.
22. Gatti M, Raschi E, Moretti U, Ardizzoni A, Poluzzi E, Diemberger I. Influenza Vaccination and Myo-Pericarditis in Patients Receiving Immune Checkpoint Inhibitors: Investigating the Likelihood of Interaction through the Vaccine Adverse Event Reporting System and VigiBase. *Vaccines (Basel)* 2021;9.
23. Witberg G, Barda N, Hoss S et al. Myocarditis after Covid-19 Vaccination in a Large Health Care Organization. *N Engl J Med* 2021;385:2132-2139.

24. Pasupathy S, Air T, Dreyer RP, Tavella R, Beltrame JF. Systematic review of patients presenting with suspected myocardial infarction and nonobstructive coronary arteries. *Circulation* 2015;131:861-70.
25. Abukhalil AD, Shatat SS, Abushehadeh RR, Al-Shami N, Naseef HA, Rabba A. Side effects of Pfizer/BioNTech (BNT162b2) COVID-19 vaccine reported by the Birzeit University community. *BMC Infect Dis* 2023;23:5.
26. The L. Urbanisation, inequality, and health in Asia and the Pacific. *Lancet* 2017;389:1370.



**Table 1.** Baseline characteristics of reports on vaccine-associated pericarditis and myocarditis adverse events. (n= 49,096)

Variables		Number (%)
Region reporting	African Region	43 (0.09)
	Region of the Americas	27,247 (55.50)
	South-East Asia Region	41 (0.08)
	European Region	16,172 (32.94)
	Eastern Mediterranean Region	48 (0.10)
	Western Pacific Region	5,545 (11.29)
Reporting year	1967 to 2019	1,974 (4.02)
	2020 to 2023	47,122 (95.98)
Reporter qualification	Health Professional	10,258 (20.89)
	Non-Health Professional	10,472 (21.33)
	Unknown	28,366 (57.78)
Studies	Study-related	48,872 (99.54)
	Non-study related	223 (0.45)
	Unknown	1 (0.00)
Sex	Male	30,013 (61.13)
	Female	18,779 (38.25)
	Unknown	304 (0.62)
Age, years	0 to 11	456 (0.93)
	12 to 17	3,620 (7.37)
	18 to 44	20,906 (42.58)
	45 to 64	8,499 (17.31)
	≥65	3,347 (6.82)
	Unknown	12,268 (24.99)
TTO, days	Median days (IQR)	1 (1-1)
Drug class	Anthrax vaccines	223 (0.45)
	DTaP-IPV-Hib vaccines	292 (0.59)
	Meningococcal vaccines	96 (0.20)
	Pneumococcal vaccines	137 (0.28)
	Typhoid vaccines	121 (0.25)
	Encephalitis vaccines	42 (0.09)
	Influenza vaccines	500 (1.02)

	Hepatitis A vaccines	65 (0.13)
	Hepatitis B vaccines	118 (0.24)
	MMR vaccines	69 (0.14)
	Rotavirus diarrhea vaccines	20 (0.04)
	Zoster vaccines	81 (0.16)
	Papillomavirus vaccines	87 (0.18)
	Smallpox vaccines	579 (1.18)
	COVID-19 mRNA vaccines	44,391 (90.42)
	Ad5-vectored COVID-19 vaccines	2,178 (4.44)
	Inactivated whole-virus COVID-19 vaccines	48 (0.10)
	Others*	49 (0.10)
Outcomes	Recovered/recovering	23,312 (47.48)
	Fatal	214 (0.44)
	Unknown	25,570 (52.08)
Single drug suspected		49,091 (99.99)

Abbreviation: DTaP-IPV-Hib, diphtheria, tetanus toxoids, pertussis, polio, and *Hemophilus influenza* type b; IQR, interquartile range; MMR, measles, mumps, and rubella; TTO, time to onset; WHO, World Health Organization.

\*Others: brucellosis, dengue vaccines, Ebola, leptospirosis, plague, rabies, tuberculosis, typhus, and yellow fever vaccines.

**Table 2.** Analysis of subgroups in vaccine-related pericarditis and myocarditis adverse events disproportionality.

	Total	Vaccine-associated pericarditis and myocarditis			IC (IC <sub>0.25</sub> ) based on age, years				
		Observed	ROR (95% CI)	IC (IC <sub>0.25</sub> )	0-11 years	12-17 years	18-44 years	45-64 years	≥65 years
<b>Total</b>					<b>1.34 (1.19)</b>	<b>2.91 (2.86)</b>	<b>2.87 (2.85)</b>	<b>3.39 (3.35)</b>	<b>4.89 (4.84)</b>
<b>Sex difference</b>									
Male	4,555,195	30,013	<b>34.22 (33.52 - 34.93)</b>	<b>3.46 (3.44)</b>	<b>1.52 (1.33)</b>	<b>3.10 (3.04)</b>	<b>2.94 (2.91)</b>	<b>3.51 (3.45)</b>	<b>3.73 (3.64)</b>
Female	6,842,897	18,779	<b>27.47 (26.81 - 28.14)</b>	<b>3.35 (3.32)</b>	<b>1.11 (0.86)</b>	<b>2.63 (2.50)</b>	<b>2.79 (2.75)</b>	<b>3.35 (3.30)</b>	<b>3.69 (3.60)</b>
<b>Vaccine types</b>									
Anthrax vaccines	15,832	223	<b>25.54 (22.37 - 29.16)</b>	<b>4.58 (4.35)</b>	N/A	N/A	<b>4.46 (4.24)</b>	1.83 (-0.24)	N/A
DTaP-IPV-Hib vaccines	1,849,474	292	0.28 (0.25 - 0.31)	-1.83 (-2.02)	<b>0.73 (0.45)</b>	-0.77 (-1.44)	<b>0.69 (0.33)</b>	0.08 (-0.82)	1.06 (-1.01)
Meningococcal vaccines	254,185	96	0.67 (0.55 - 0.82)	-0.57 (-0.91)	0.23 (-0.71)	-0.67 (-1.26)	<b>1.50 (1.01)</b>	0.00 (-3.78)	N/A
Pneumococcal vaccines	608,118	137	0.40 (0.34 - 0.47)	-1.31 (-1.59)	<b>1.30 (0.82)</b>	-0.37 (-2.96)	<b>0.85 (0.18)</b>	0.21 (-0.72)	<b>1.49 (0.72)</b>
Typhoid vaccines	35,132	121	<b>6.17 (5.16 - 7.38)</b>	<b>2.59 (2.29)</b>	N/A	N/A	<b>3.47 (3.15)</b>	1.34 (-0.43)	N/A
Encephalitis vaccines	37,396	42	<b>2.00 (1.48 - 2.71)</b>	<b>0.99 (0.47)</b>	1.61 (-0.46)	0.61 (-1.46)	<b>1.66 (0.94)</b>	<b>1.86 (0.56)</b>	1.20 (-2.58)
Influenza vaccines	478,337	500	<b>1.87 (1.71 - 2.04)</b>	<b>0.90 (0.75)</b>	0.45 (-0.57)	0.37 (-0.25)	<b>1.01 (0.78)</b>	<b>1.47 (1.16)</b>	<b>3.61 (3.30)</b>
Hepatitis A vaccines	153,470	65	0.76 (0.59 - 0.96)	-0.40 (-0.81)	-0.08 (-2.15)	-0.49 (-1.79)	<b>1.27 (0.75)</b>	0.32 (-1.45)	N/A
Hepatitis B vaccines	214,065	118	0.98 (0.82 - 1.18)	-0.02 (-0.33)	<b>2.99 (2.48)</b>	-0.66 (-1.80)	0.16 (-0.36)	<b>1.31 (0.41)</b>	1.22 (-2.56)
MMR vaccines	453,603	69	0.27 (0.21 - 0.34)	-1.87 (-2.27)	0.45 (-0.16)	-1.71 (-3.48)	0.35 (-0.27)	0.54 (-1.53)	N/A
Rotavirus diarrhea vaccines	218,477	20	0.16 (0.11 - 0.25)	-2.58 (-3.34)	0.78 (-0.09)	N/A	N/A	N/A	N/A

Zoster vaccines	340,721	81	0.42 (0.34 - 0.53)	-1.23 (- 1.60)	0.47 (- 0.51)	-0.69 (- 1.99)	0.09 (- 0.89)	0.14 (- 0.59)	<b>1.33 (0.43)</b>
Papillomavirus vaccines	204,234	87	0.76 (0.62 - 0.94)	-0.39 (- 0.75)	-0.70 (- 4.48)	-0.86 (- 1.30)	-0.29 (- 0.98)	N/A	N/A
Smallpox vaccines	14,696	579	<b>73.68 (67.79 - 80.10)</b>	<b>6.05 (5.91)</b>	N/A	1.32 (- 2.46)	<b>6.02 (5.87)</b>	<b>3.75 (2.88)</b>	N/A
COVID-19 mRNA vaccines	4,871,512	44,391	<b>37.77 (37.00 - 38.56)</b>	<b>3.07 (3.05)</b>	<b>4.52 (4.24)</b>	<b>4.16 (4.11)</b>	<b>3.41 (3.39)</b>	<b>3.90 (3.86)</b>	<b>5.34 (5.28)</b>
Ad5-vectored COVID-19 vaccines	1,450,737	2,178	<b>1.40 (1.34 - 1.46)</b>	<b>0.46 (0.39)</b>	<b>2.48 (0.72)</b>	-0.59 (- 2.66)	0.09 (- 0.03)	<b>2.20 (2.09)</b>	<b>4.37 (4.20)</b>
Inactivated whole-virus COVID-19 vaccines	197,881	48	0.22 (0.17 - 0.29)	-2.15 (- 2.63)	1.98 (- 0.61)	N/A	-2.05 (- 2.73)	-1.14 (- 2.28)	<b>1.80 (0.66)</b>
Others*	114,571	49	0.76 (0.58 - 1.01)	-0.39 (- 0.86)	-1.62 (- 5.40)	N/A	-0.09 (- 0.77)	0.74 (- 0.57)	<b>2.46 (0.69)</b>

Abbreviation: CI, confidence interval; DTaP-IPV-Hib, diphtheria, tetanus toxoids, pertussis, polio, and *Hemophilus influenza* type

b; IC, information component; MMR, measles, mumps, and rubella; ROR, reported odds ratio.

Bold style indicates when the value of  $IC_{0.25}$  is greater than 0.0 or the lower end of the ROR 95% CI is greater than 1.0. This means it is statistically significant.

Numbers in bold indicate a statistical significance.

\*Others included brucellosis, dengue vaccines, Ebola, leptospirosis, plague, rabies, tuberculosis, typhus, and yellow fever vaccines.

# Caregiver-Child Interaction Detection Model Based on Computer Vision to Measure Social Interaction Skills of Children with Developmental Disabilities

Byeonghun Kim<sup>1</sup>, Insu Jeon<sup>2</sup>, Chomyong Kim<sup>3</sup>, Jung-Yeon Kim<sup>3</sup>,  
Jiyoung Woo<sup>3,4</sup>, and Byeongjoon Noh<sup>4,\*</sup>

<sup>1</sup> Department of Future Convergence Technology, Soonchunhyang University, Asan, Republic of Korea

<sup>2</sup> Department of Medical Science, Soonchunhyang University, Asan, Republic of Korea

<sup>3</sup> Department of ICT Convergence, Soonchunhyang University, Asan, Republic of Korea

<sup>4</sup> Department of AI and Big Data, Soonchunhyang University, Asan, Republic of Korea

\*Contact: powernoh@sch.ac.kr, phone +82 10 9995 1231

**Abstract**— Measuring caregiver-child interactions is critical for early detection of developmental problems in children. Diagnostic assessments in a medical setting are time-consuming due to a variety of factors, and existing developmental screening tests rely on subjective parental ratings. Therefore, a scale that can objectively measure the quality of interaction is needed. Recent advances in computer vision have led to research on clinical decision systems for diagnosing developmental disorders in children. Previous works have limited application to real-world settings because they have limited experimental environments and focus on the behavior of caregivers and children separately. Therefore, it is difficult to comprehensively capture the interaction between them. To address this problem, we propose a vision-based model to evaluate the interaction between caregivers and children with disabilities in real-world situations. The proposed method measures the distance between the caregiver and the child, the number of times they look at the same location by estimating their gaze, and the number of times they make eye contact. The system estimates the gaze-based interaction between caregiver and child in real time, contributing to an objective assessment of the level of interaction. The model has been validated through extensive experiments, demonstrating that it can be used to detect gaze-based interactions.

## I. INTRODUCTION

Children with developmental disabilities have problems with social relationships, communication and cognitive developmental delays, and their social growth is significantly slower than their peers, causing them great difficulty in real life [1]. Early identification of these disorders is critical to prevent secondary disabilities and to allow for appropriate treatment. Currently, there are two main approaches to diagnosing developmental disabilities: biological and behavioral. Biological approaches have traditionally been used to diagnose and assess the severity of developmental disabilities using diagnostic methods such as blood tests, genetic tests, and brain scans. However, these methods cannot be used if the caregiver does not notice a problem with the child's behavior, making it impossible to be proactive. As a result, most medical settings

use a behavioral approach to diagnosing developmental disabilities [2].

To diagnose a developmental disorder in a child, a behavioral approach typically involves taking a medical history, interviewing caregivers, and observing the child's behavior. Children often exhibit biobehaviors such as repetitive behaviors (homologous behaviors), difficulties in social communication, and limited expressiveness [3]. The Korean Version for Learning Disability Evaluation Scale (K-LDES) [4], the Childhood Autism Rating Scale-2 (CARS-2) [5], and the Korean Developmental Screening Test for Infants & Children (K-DST) [6] are commonly used to record these observations. These scales typically include assessments of gross motor skills, fine motor skills, cognitive skills, language development, socialization, and self-help skills. The observation scores are summed, and if they exceed a predetermined threshold, the diagnosis is confirmed. After diagnosis, a functional assessment is conducted using a tool such as VBMAPP [7] to create a personalized developmental program aimed at improving skills needed for social integration for children with developmental disabilities. Functional assessments determine the skills a child with a developmental disability needs by evaluating the child's skills in a variety of areas, including developmental level, disability, transition, task analysis, intervention support skills, level placement, and IEP goals.

However, traditional methods of behavioral diagnosis and assessment have several limitations. First, the process of observing and interpreting a child's behavior requires significant time and effort [8]. Diagnostic assessments in unfamiliar medical settings can be time-consuming because of ambiguous findings due to a child's mild symptoms or nervousness. Second, they rely on the clinician's judgment, which can affect reliability [9]. Differences in training and clinical experience among clinicians, as well as different cultural backgrounds, may affect their interpretation of observed behaviors. Finally, ratings of caregiver-child interactions are subjective in nature. Appropriate caregiver-

child interaction is a key factor that has a significant impact on children's cognitive development, socialization, etc., and can also reduce caregiver stress [10], [11]. Therefore, there is a need to study objective measures of caregiver-child interaction in an unsupervised, relaxed environment.

Meanwhile, the rapid spread of advanced technologies in the field of computer vision based on deep learning has led to research to apply them to the diagnosis and evaluation of developmental disorder behaviour [12]-[15]. Qandeel et al. proposed using video-based behavior recognition to identify specific behaviors to determine whether a child has a developmental delay [14]. This method can objectively analyze and quantify a child's behavior, although it has the disadvantage of not taking into account the interaction with the caregiver. Using facial recognition technology to extract children's facial expressions and classify their emotions by Jordan et al. [15]. They identified specific emotions from children's facial expressions and used them to evaluate children's social and emotional development. However, it is difficult to apply to videos from real-world environments due to poor quality, and it has the disadvantage of not taking into account proper interaction with caregivers.

In this study, we propose a model for detecting caregiver-child interactions based on gaze estimation and distance. Our proposed method is adaptive to a variety of unsupervised real-world environments and has several advantages over conventional gaze tracking approaches. The extracted results can be used as a variable to predict whether a child has a developmental disorder, which can help clinicians diagnose the pathology and establish a functional assessment of the child.

The contributions of our study are as follows:

- **Caregiver-child interaction detection model:** Our model implements a novel approach to detect caregiver-child interaction based on gaze and distance.
- **Adaptability to real-world environments:** It works effectively in environments where traditional diagnostic assessments are difficult, such as homes and kindergartens, providing more flexibility for behavioral diagnosis and assessment.
- **Objective measurement of caregiver-child interactions:** Integrating gaze and distance information provides a model that can objectively measure caregiver-child interactions.

The remainder of this paper consists of three chapters as follows: Chapter II describes the methodology for implementing the proposed model, and Chapter III presents and discusses the used datasets and eye tracking results. Finally, Chapter IV concludes the paper, suggesting directions for future research.

## II. METHOD

This chapter explains in detail the overall structure of the proposed framework of models for detecting caregiver-child interactions, as well as the individual models used for this purpose.

### A. Overall Architecture

Fig. 1 shows the overall architecture of the proposed model, which consists of the following models: Adult-child detection, Head detection, Head Pose Estimation.

The proposed framework operates in the below sequence. It takes a video as input, divides it into single frames, and processes two tasks in parallel: distance calculation and gaze estimation. The distance calculation task works based on an adult-child detection model. It calculates the distance between Bounding Boxes based on the detection results. Then, the gaze estimation task performs head pose estimation based on the detection results of the head detection model. We combine these two results to find points that exceed an arbitrarily set threshold and judge them as interactions. In the following chapters, we provide detailed descriptions of the models we used to achieve these tasks.

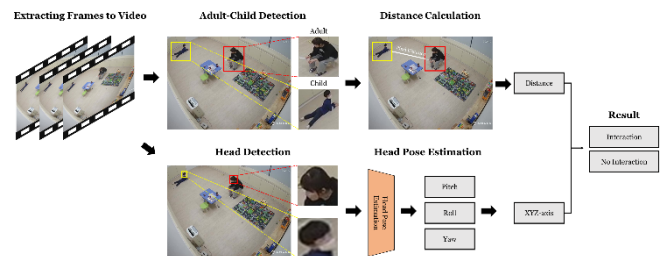


Fig. 1 The overall framework of interaction detection model

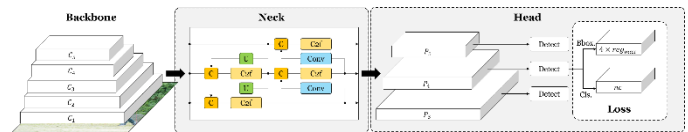


Fig. 2 The overall structure of YOLOv8, connected by the backbone, neck, and head.

### B. Object Detector

In this section, we briefly explain You Look Only Once (YOLO) version 8 model [16], widely used and state-of-the-art object detection due to its high performance on object detection tasks, as the object detector in the proposed framework. YOLOv8 is renowned for its ability to perform real-time object detection with a single network pass, making it highly efficient and effective for a wide range of applications [16].

The overall architecture of YOLOv8 is shown in Fig. 2, composing mainly backbone, neck, and head. One of the most significant changes in YOLOv8 is the shift to anchor-free design. Traditional anchor-based models rely on predetermined anchor boxes to estimate the position and size of objects. In contrast, anchor-free models like YOLOv8 directly estimate the position and size of objects without the constraints of a fixed anchor box. This design contributes to more accurate object detection results [17]. YOLOv8 also changes the kernel size of the first convolutional layer of the backbone model from  $6 \times 6$  to  $3 \times 3$ . As the kernel size was reduced by half, the number of parameters was decreased, resulting in faster training. Other architecture changes involved replacing the C3 module with a C2f module and replacing the first convolutional layer of the bottleneck from  $1 \times 1$  to  $3 \times 3$ . All of these architecture adjustments were implemented to improve the accuracy and speed of the model [17], [18].



In this study, the object detector is responsible for detecting adults, children, and heads. YOLOv8 was pre-trained using a benchmark dataset, Microsoft Common Objects [19], but since this dataset has unlabeled adults and children, and heads, which are the main target objects of this framework, we additionally trained with data that contains information about these labels. A detailed description of this can be found in section III-A.

### C. Head Pose Estimation for Gaze estimation

In this study, we use a method for estimating gaze using head pose to estimate gaze in a variety of environments. Most of the existing gaze estimation studies use facial landmarks to estimate gaze, which has high prediction performance for a narrow range of frontal views, but performs poorly when the target is looking backward and when the target is looking down or up [20]. However, we need to perform well in an unconstrained environment, so we adopt a gaze estimation method that does not rely on facial landmarks, but only on head pose.

To accomplish this, we employed the six degrees of freedom head pose estimation(6DoF-HPE) model [21]. 6DoF-HPE utilizes RGB images to simultaneously estimate rotation (pitch, roll, yaw) and translational components (x,y,z). The 6DoF-HPE framework consists of a head detection stage and a rotation estimation stage. In the head detection stage, we use SSD Detector(Single-Shot MultiBox) [22] to detect heads in various orientations and directions, and in the rotation estimation stage, we use a CNN-based network consisting of RepVGG-B1g4 backbone [23] and SENet [24]. We only use the rotation estimation model because we perform the head detection step before.

Fig. 3 shows the detailed network structure of the rotation estimation model. Feature maps are extracted using the properties of convolutional networks that consider specific parts of the whole image, and reconstructed into feature vectors by the SENet module. These vectors are concatenated with a fully connected layer to output a six-dimensional rotational representation. We use this rotation representation to calculate the human field of view (FOV) to estimate gaze. Typically, a human's perceived field of view is between -30 and 30 degrees from the front Fig. 4. Since the z-axis of the extracted translational vector represents the front of the head, we transform the axial coordinates of the z-axis to estimate the gaze.

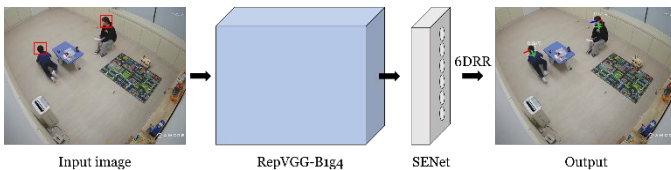


Fig. 3 Overview of the network of rotational estimation

In this section, we describe the datasets to train and validating the used models in the proposed framework. The experiments are divided into two parts: (1) evaluating the individual performance of each model performed in the proposed framework, and (2) checking the visualization results for gaze estimation.

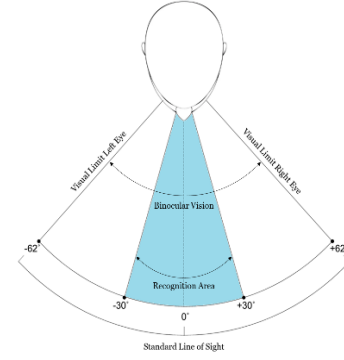


Fig. 4 Human field of view(FOV) and recognition area

## III. EXPERIMENTS AND RESULTS

### A. Data description

We collected videos of a caregiver and child playing freely in the environment. The average length of the collected videos is 10 minutes, and there are four different angles of the same scene. For training the model, we used some videos to label the parent and child classes and the head position. However, the 2D video images we took cannot be labeled for gaze vectors, so quantitative evaluation of gaze vectors is not possible. Therefore, we perform qualitative evaluation of gaze estimation results based on visualization.

### B. Object detection and distance estimation results

In our experiments, we quantitatively evaluate each model in the proposed framework by adopting accuracy, precision, and recall as evaluation metrics from the model accuracy perspective. In addition, object detectors typically use Average Precision (AP) score as metric to understand how well a model can identify and localize objects within image across different classes. It is calculated by finding the area under the precision-recall curve, representing the average precision across all possible recall levels. The mAP score is obtained by averaging the AP scores across different classes in the dataset. Each metric can be obtained through the following formula:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad (2)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (3)$$

$$AP = \int_0^1 \text{precision}(r) dr \quad (4)$$

$$mAP = \frac{1}{n} \sum_{i=1}^n AP_i \quad (5)$$

We use pre-trained models with YOLOv8n (nano version), the lightest and fastest model in the YOLOv8 series. This is especially optimized for environments where computing

resources are limited, such as edge devices or mobile applications, making it applicable to a wide range of environments. For additional training of this model, we set the hyperparameters to 100 epochs and 16 batch sizes. The results of this training process, including loss and accuracy as a function of epoch, are shown in the Fig. 5. In addition, the classification evaluation results for adult, children, and head are detailed in the TABLE I.

TABLE I  
OBJECT DETECTOR DETECTION PERFORMANCE EVALUATION RESULTS

Class	Accuracy	Precision	Recall	mAP@0.5	Time(s)
Adult	0.991	1.000	0.999	0.995	0.0017
Child	0.988	0.995	1.000	0.995	
Head	0.993	0.987	0.994	0.990	

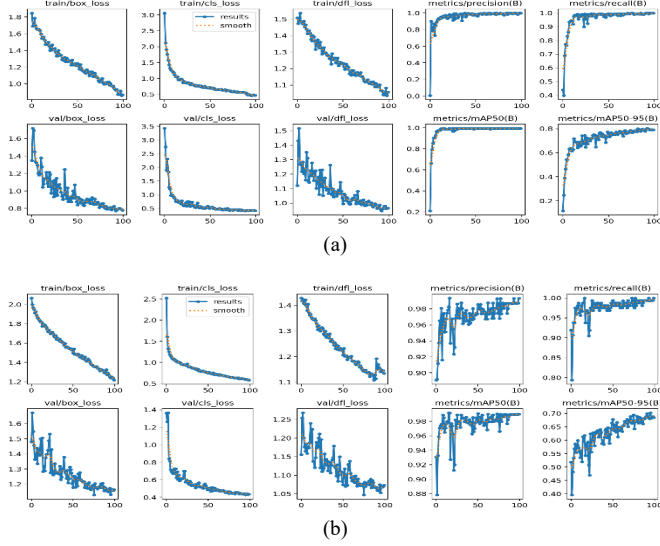


Fig. 5 Results of the training process for detectors pre-trained with YOLOv8n (a) Adult-child detector (b) Head detector

The experimental evaluation result mAP@0.5 is 0.995, which shows that the model's detection capability is very accurate. The model is also highly efficient with an inference time of 0.0017 seconds, making it applicable in environments with low computing power (e.g., homes and kindergartens), allowing for real-time inference in a variety of environments. The results of estimating the distance between a caregiver and a child based on these results can be seen in Fig. 6.

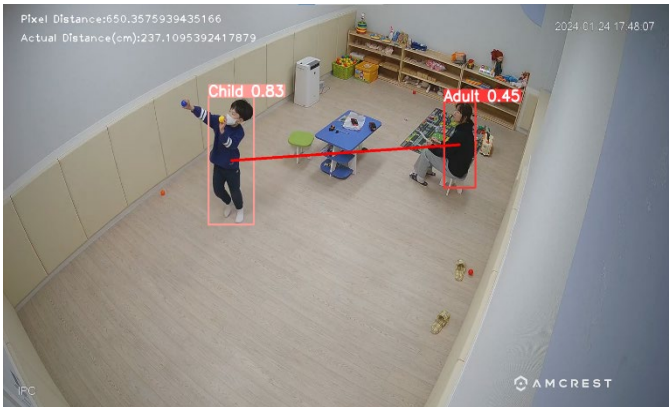


Fig. 6 Visualization of distance estimation results (distance estimation result text in the upper left)

### C. Gaze estimation results

In this study, we perform a qualitative evaluation of visualization-based gaze estimation results. The evaluation method is compared between the proposed model and a state-of-the-art L2CS [20] model that performs well in the field of head pose estimation.

Fig. 7 presents the results of applying the L2CS model and the proposed model to the same scene. The L2CS model can be seen to be unable to estimate gaze when the caregiver has their back to the camera. This error is a significant problem because it can lead the model to believe that the child and caregiver are not interacting with each other, even though they are looking at each other. On the other hand, our proposed model is able to estimate gaze more accurately, despite the caregivers having their backs to the camera.

Next, Fig. 8 demonstrates the results of applying the proposed model to videos taken from different angles and detecting gaze interactions. We interpret the results as robust performance despite significant changes in the image due to the camera angle. The model's robustness is important because it can be applied to various environments and still perform well.

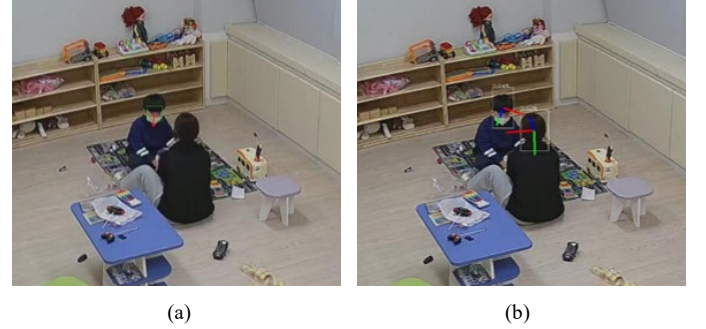


Fig. 7 Results of applying the L2CS model and the proposed model to the same scene (a) L2CS model (b) proposed model

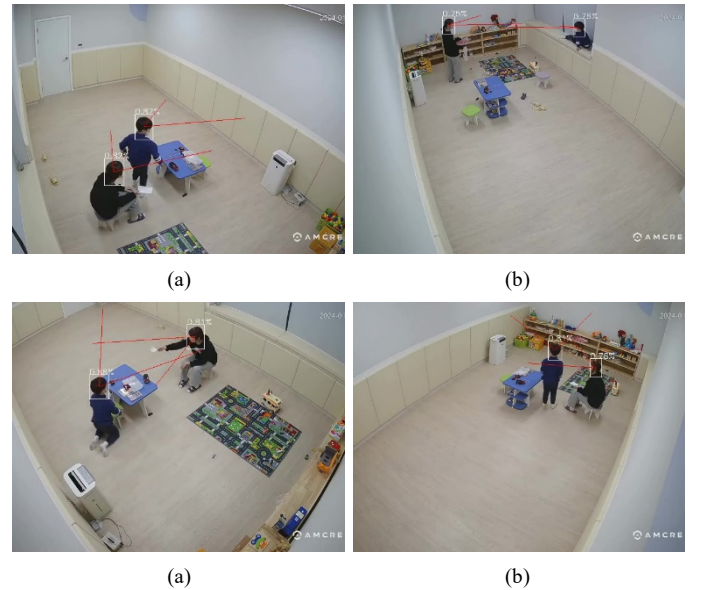


Fig. 8 Visualization of the results of detected gaze interactions when the proposed model is applied to videos taken from different angles

#### D. Discussions

This study proposes a gaze and distance based interaction detection model to detect caregiver-child interaction. The model consists of several models with specific functions of caregiver-child detection, distance estimation, head detection, head pose estimation, and gaze estimation. These models are used as variables for interaction detection and form the basis for measuring the degree of socialization.

In our experiments, we evaluated the performance based on accuracy and visualization results. The results showed that the individual models had high accuracy and real-time inference capabilities, and were robust in a variety of environments. In particular, the model that estimates gaze without face detection was found to be able to achieve high accuracy even in images with a lot of occlusion, such as CCTV.

These systems can be used in unsupervised real-world environments or environments with low computational power, allowing interaction to be measured even in kindergarten and home environments. They can also measure interaction objectively, which is an important benefit for clinicians.

#### IV. CONCLUSIONS

We have developed a computer vision-based system that automatically estimates the gaze of a child, a caregiver, and the distance between them to measure the developmental level of children with developmental disabilities. The proposed system further extends its applicability in the medical domain by establishing an automated protocol to detect interaction. However, the need to set a threshold for the criteria of interaction may limit its validity, and computer vision techniques still have limitations in that occlusion can significantly affect the estimation results.

In future work, we will recruit subjects with developmental disabilities, measure the degree of developmental disability, and develop a comprehensive model for various interactions such as human-object interaction and human-human interaction, estimate the actual distance, and improve the model to obtain more objective results. In future extensions of our work, we expect that more developmental assessments can be performed automatically, accurately, and objectively at home using lightweight devices prior to a thorough evaluation by a professional doctor, enabling early intervention for infants with developmental behavioral disorders.

#### ACKNOWLEDGMENT

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. RS-2023-00218176)

#### REFERENCES

- [1] S. Ghafghazi, A. Carnett, L. Neely, A. Das, and P. Rad, "Ai-augmented behavior analysis for children with developmental disabilities: Building toward precision treatment," *IEEE Systems, Man, and Cybernetics Magazine*, vol. 7, no. 4, pp. 4–12, 2021.
- [2] L. J. Barnhill, "The diagnosis and treatment of individuals with mental illness and developmental disabilities: An overview," *Psychiatric Quarterly*, vol. 79, pp. 157–170, 2008.
- [3] C. for Disease Control, Prevention et al., "Data and statistics on autism spectrum disorder—cdc," CDC. gov, 2019.
- [4] M.-S. SHIN, K.-E. HONG, Z.-S. KIM, and S.-C. CHO, "A standardization study of the korean version of learning disability evaluation scale," *Journal of Korean Neuropsychiatric Association*, pp. 1233–1245, 1998.
- [5] L. Jurek, M. Baltazar, S. Gulati, N. Novakovic, M. Nuñez, J. Oakley, and A. O'Hagan, "Response (minimum clinically relevant change) in asd symptoms after an intervention according to cars-2: consensus from an expert elicitation procedure," *European child & adolescent psychiatry*, vol. 31, no. 8, pp. 1–10, 2022.
- [6] H. J. Chung, D. Yang, G.-H. Kim, S. K. Kim, S. W. Kim, Y. K. Kim, Y. A. Kim, J. S. Kim, J. K. Kim, C. Kim et al., "Development of the korean developmental screening test for infants and children (k-dst)," *Clinical and Experimental Pediatrics*, vol. 63, no. 11, p. 438, 2020.
- [7] M. L. Sundberg, VB-MAPP Verbal Behavior Milestones Assessment and Placement Program: a language and social skills assessment program for children with autism or other developmental disabilities: guide. Mark Sundberg, 2008.
- [8] P. McCarty and R. E. Frye, "Early detection and diagnosis of autism spectrum disorder: why is it so difficult?" in *Seminars in Pediatric Neurology*, vol. 35. Elsevier, 2020, p. 100831.
- [9] F. P. GLASCOE and P. H. DWORKIN, "Obstacles to effective developmental surveillance: errors in clinical reasoning," *Journal of Developmental & Behavioral Pediatrics*, vol. 14, no. 5, pp. 344–349, 1993.
- [10] A. P. Kaiser, P. P. Hester, and A. S. McDuffie, "Supporting communication in young children with developmental disabilities," *Mental retardation and developmental disabilities research reviews*, vol. 7, no. 2, pp. 143–150, 2001.
- [11] R. P. Hastings, "Parental stress and behaviour problems of children with developmental disability," *Journal of intellectual and developmental disability*, vol. 27, no. 3, pp. 149–160, 2002.
- [12] A. Esteva, K. Chou, S. Yeung, N. Naik, A. Madani, A. Mottaghi, Y. Liu, E. Topol, J. Dean, and R. Socher, "Deep learning-enabled medical computer vision," *NPJ digital medicine*, vol. 4, no. 1, p. 5, 2021.
- [13] G. Brihadiswaran, D. Haputhanthri, S. Gunathilaka, D. Meedeniya, and S. Jayarathna, "Eeg-based processing and classification methodologies for autism spectrum disorder: A review," *Journal of Computer Science*, vol. 15, no. 8, 2019.
- [14] Q. Tariq, S. L. Fleming, J. N. Schwartz, K. Dunlap, C. Corbin, P. Washington, H. Kalantarian, N. Z. Khan, G. L. Darmstadt, and D. P. Wall, "Detecting developmental delay and autism through machine learning models using home videos of bangladeshi children: Development and validation study," *Journal of medical Internet research*, vol. 21, no. 4, p. e13822, 2019.
- [15] J. Hashemi, K. Campbell, K. Carpenter, A. Harris, Q. Qiu, M. Tepper, S. Espinosa, J. Schaich Borg, S. Marsan, R. Calderbank et al., "A scalable app for measuring autism risk behaviors in young children: a technical validity and feasibility study," in *Proceedings of the 5th EAI International Conference on Wireless Mobile Communication and Healthcare*, 2015, pp. 23–27.
- [16] G. Jocher, A. Chaurasia, and J. Qiu, "YOLO by Ultralytics," Jan. 2023. [Online]. Available: <https://github.com/ultralytics/ultralytics>
- [17] J. Solawetz, "What is yolov8? the ultimate guide," 2023.
- [18] R. Chhatrala, S. Patil, S. Lahudkar, and D. V. Jadhav, "Sparse multilinear laplacian discriminant analysis for gait recognition," *Pattern Analysis and Applications*, vol. 22, pp. 505–518, 2019.
- [19] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13*. Springer, 2014, pp. 740–755.
- [20] A. A. Abdelrahman, T. Hempel, A. Khalifa, and A. AlHamadi, "L2cnet: Fine-grained gaze estimation in unconstrained environments," 2022.
- [21] R. Algabri, H. Shin, and S. Lee, "Real-time 6dof fullrange markerless head pose estimation," *Expert Systems with Applications*, vol. 239, p. 122293, 2024.
- [22] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*. Springer, 2016, pp. 21–37.
- [23] X. Ding, X. Zhang, N. Ma, J. Han, G. Ding, and J. Sun, "Repvgg: Making vgg-style convnets great again," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 13 733–13 742.
- [24] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

# Prenatal and postnatal factors associated with sudden infant death syndrome: an umbrella review of meta-analyses

Hyeri Lee,<sup>1</sup> Selin Woo<sup>1\*</sup>

<sup>1</sup>Center for Digital Health, Medical Science Research Institute, Kyung Hee University Medical Center, Kyung Hee University College of Medicine, Seoul, South Korea

\*Correspondence: Selin Woo (dntpf1s@naver.com)

**Abstract**— A comprehensive quantitative evidence synthesis on the risk and protective factors for sudden infant death syndrome effects (SIDS) is lacking. We aimed to investigate the risk and protective factors related to SIDS. We conducted an umbrella review of meta-analyses of observational and interventional studies assessing the SIDS-related factors. PubMed/MEDLINE, Embase, EBSCO, and Google Scholar were searched from inception until January 18, 2023. Data extraction, quality assessment, and certainty of evidence were assessed by using AMSTAR2 following PRISMA guidelines. According to observational evidence, credibility was graded and classified by class and quality of evidence (CE; convincing, highly suggestive, suggestive, weak, or not significant). Our study protocol was registered with PROSPERO (CRD42023458696). The risk and protective factors related to SIDS are presented as equivalent odds ratios (eOR). We identified eight original meta-analyses, including 152 original articles, covering 12 unique risk and protective factors of SIDS across 21 countries and five continents. SIDS had eight risk factors, including prenatal drug exposure (eOR, 7.84 [95% CI, 4.81-12.79], CE=highly suggestive), prenatal opioid exposure (9.55 [95% CI, 4.87-18.72], CE=suggestive), prenatal methadone exposure (9.52 [95% CI, 3.34-27.10], CE=weak), prenatal cocaine exposure (4.38 [95% CI, 1.95-9.86], CE= weak), prenatal maternal smoking (2.25 [95% CI, 1.95-2.60], CE= highly suggestive), postnatal maternal smoking (1.97 [95% CI, 1.75-2.22], CE=weak), bed-sharing (2.89 [95% CI, 1.81-4.60], CE=weak), and infants found with heads covered by bedclothes after last sleep (11.01 [95% CI, 5.40-22.45], CE= suggestive). On the other hand, three protective factors, which are Breastfeeding (0.57 [95% CI, 0.39-0.83], CE=non-significant), supine sleeping position (0.48 [95% CI, 0.37-0.63], CE=suggestive), and pacifier use (0.44 [95% CI, 0.30-0.65], CE=weak), were also identified. Therefore, based on evidence, we suggested several risk and protective factors for SIDS. This study suggests a need for further studies on SIDS-related factors supported by weak credibility, no association, or without adequate research paper.

## I. INTRODUCTION

Sudden infant death syndrome (SIDS) is a sudden, unexpected death of a healthy infant under the age of one with no specific cause of death identified by autopsy or other means [1]. In the past, before 1990, deaths of unknown causes in infants under the age of 1 were defined as SIDS. However, as death investigations gradually became more common, many of the unspecified causes of infant death were able to be identified, so SIDS was re-defined as sudden unexplained infant death (SUID) that is still unspecified even after several death investigations including autopsy had done. Since more death investigations had been done, the sleeping position was suspected of having an association with SIDS, a sleep campaign recommending the supine position was done and the SIDS incidence was lowered [2]. In 1990, the incidence ranged from 1.5 to 3 per 1000 live births, but in 2000, the incidence was about 0.2 to 1 per 1000 live births in most countries [3]. However, according to the Centers for Disease Control and Prevention (CDC), SIDS in the United States in 2021 is still the third leading cause (7.3%) of infant deaths under the age of 1 [3]. Therefore, active research on SIDS is considered necessary [1, 4].

Despite several previous researches, the precise etiology of SIDS remains ambiguously defined, attributed largely to the intricate interplay of genetic, environmental, and developmental factors [5]. Both prenatal and postnatal periods are pivotal phases in the infant. Prenatal exposure to substances, including alcohol, cannabis, and opiates, is one of the potential risk factors associated with SIDS [6]. However, several protective factors have emerged, shedding hope on preventive strategies for SIDS. Supine position and breastfeeding have been suggested as protective factors against SIDS.

While numerous meta-analyses have provided insights into specific risk and protective factors, a unified synthesis of these findings is essential to understand the SIDS-associated factors comprehensively [6-12]. To provide an overview of the breadth, quality, and certainty of the previously reported associations between SIDS and risk and protective factors, we performed an umbrella review of the evidence across published meta-analyses. Therefore, we systematically identified relevant meta-analyses, summarized the risk and protective factors of SIDS, and

analyzed the certainty of evidence to provide a comprehensive overview of the SIDS-associated factors.

## II. METHODS

### *Literature search strategy and selection criteria*

We conducted an umbrella review to summarize and evaluate the risk and protective factors of SIDS. This review followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 guidelines, and its protocol was registered with PROSPERO (Registration No. CRD42023458696) [13, 14]. Two authors, H.L. and T.H.K., systematically searched online databases (PubMed/MEDLINE, Embase, EBSCO, and Google Scholar) for meta-analyses of observational studies or intervention trials examining the association between SIDS and its risk factors until January 18, 2024. Our search strategy was as follows: ('sudden infant death syndrome' OR 'SIDS' OR 'cot death') AND ('meta-analyses' OR 'systematic review') and their variants. We also searched the references of the eligible articles manually and reviewed the titles, abstracts, and full texts of the studies found through the search (T.H.K. and H.L.). Only observational studies or intervention trials meta-analyses were included because there was no randomized controlled trial (RCT). The following studies were excluded: duplicated studies, studies about outcomes of SIDS, studies not investigating the direct association between SIDS and its risk or protective factor, and a study reporting overlapped results. The SIDS-related factors eligible for our umbrella review are prenatal exposure to any drug and each of opioid, methadone, cocaine, prenatal exposure to PM10 prenatal and postnatal maternal smoking, bed sharing, supine sleep position, breastfeeding, head-covering, and use of a pacifier. Meta-analyses reporting odds ratio (OR) or relative risk ratio (RR) of the association of SIDS and its risk or protective factors were included in this review. We re-calculated the pool effect size, and RR was converted to OR in re-analysis [15].

### *Data extraction and quality assessment*

We assessed the methodological quality of the included studies and rated them based on the A Measurement Tool Assessment Systematic Reviews 2 (AMSTAR2) checklist. In cases of disagreement, another researcher (J.K.) gets involved in the discussion and reaches a consensus.

### *Data extraction*

Two independent researchers (T.H.K. and H.L.) screened the titles and abstracts and selected articles for full-text review. We extracted the following data from the selected articles: publication year, number of primary studies included, outcomes, country of study, number of cases and participants, study design, effect estimation model (random or fixed effects), heterogeneity, and maximally adjusted effect size with 95% confidence interval (CI). Each meta-analysis was re-analyzed by the Der Simonian and Laird random fixed effects model [4, 16]. We did not re-analyze any of the dose-response meta-analyses if the

data for dose-response analysis is insufficient. We performed several more analyses to evaluate specific aspects.  $I^2$  statistics was performed to evaluate heterogeneity, and  $I^2$  value exceeding 50% indicates significant heterogeneity. P-curve analysis was used to detect p-hacking [16]. The 95% prediction interval (PI) was examined to assess the uncertainty of the observed estimates and predict the value of new future observations based on Bayesian statistics. The Knapp-Sidik-Jonkman random effects model was used to reduce inappropriate type 1 errors [4, 16]. Egger's test estimates publication bias when the p-value is less than 0.1. We approximated equivalent odds ratios (eORs) for various metrics, including the relative risk (RR), in accordance with the latest guidelines [14]. All analyses were conducted based on the 'meta' package of R software (version 4.2.2; R Foundation, Vienna, Austria), and two-sided p value under 0.05 was considered as significant [17].

### *Assessment of quality of study and evidence*

In this review, we assessed the class and quality of evidence (CE) for each outcome, using criteria from previous umbrella reviews[18]. Observational study associations were categorized into five levels based on the strength of evidence for potential environmental risk or protective factors (class I, convincing; class II, highly suggestive; class III, suggestive; class IV, weak; NA, not significant). The credibility of evidence from observational studies was rated considering various factors, including the number of events related to the outcome of interest, the p-value of the association, the presence of small study effects, excess of significance bias, prediction intervals, statistical significance in the largest study, and heterogeneity.

According to the criteria of observational study, the credibility of evidence was graded. Class I: Involved exceeding 1000 cases (or over 20,000 participants); Significant summary associations as per random-effects calculations, with  $p < 10^{-6}$ ; Absence of indications pointing towards small-study effects; No observed evidence suggesting an excess of significance bias; Not including the null value in prediction intervals; The largest study is nominally significant,  $p < 0.05$ ; A low degree of heterogeneity, specifically,  $I^2$  value below 50%, Class II: Involved exceeding 1000 cases (or over 20,000 participants); Significant summary associations as per random-effects calculations, with  $p < 10^{-6}$ ; The largest study is nominally significant,  $p < 0.05$ , Class III: Involved exceeding 1000 cases (or over 20,000 participants); Significant summary associations as per random-effects calculations, with  $p < 10^{-3}$ , Class IV: Any other associations that present a p-value lower than 0.05, No significant evidence was defined when  $p > 0.05$ .

## III. RESULT



Among 270 articles identified through the literature review, 242 were excluded as duplicates, leaving 15 for the full-text screening (Figure 1). After excluding 8 articles based on the full text, we identified 8 meta-analyses for evaluating 12 unique risk and protective factors of SIDS. The eligible meta-analyses were published between 2005 to 2022, with 152 articles across 21 countries (Australia, Brazil, Belgium, Canada, Denmark, England, France, Germany, Hong Kong, Hungary, Ireland, Lithuania, Netherlands, New Zealand, Norway, Scotland, South Korea, and Sweden, United Kingdom, and United States) and five continents (Asia, Europe, North America, Oceania, and South America) were included (Table 1) [6-12]. Additionally, we present the characteristics of systematic reviews measuring sudden infant death syndrome.

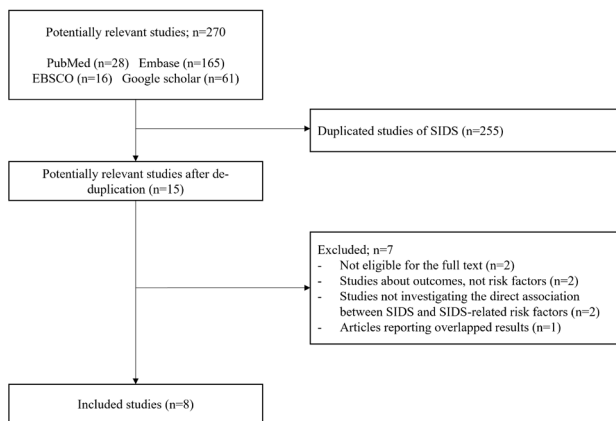


Fig. 1 Study flow chart

The quality of the original meta-analysis based on AMSTAR 2 was high in two meta-analyses, moderate in one, and low in five. Eight meta-analyses covered over 10 million participants, with 12 unique SIDS-associated prenatal and postnatal factors, including prenatal drug exposure, prenatal opioid exposure, prenatal methadone exposure, prenatal cocaine exposure, prenatal cocaine exposure, prenatal PM10 exposure, prenatal and postnatal maternal smoking, infants found with heads covered by bedclothes after last sleep, bed sharing, supine position, breastfeeding, and pacifier use. Overall, our re-analyses showed that 12 unique significant associations were identified reporting on the risk and protective factors of SIDS.

One meta-analytical association (8.3%) met the highly suggestive evidence three meta-analytical associations (25%) met the suggestive evidence, and five (41.7%) met weak evidence, two meta-analytical association (16.7%) met non-significant (8.3%) based on the CE. Except for five factors (prenatal opioid exposure, sleep, bed sharing, supine position, breastfeeding, and pacifier use), the shape of the p-curve was highly right-skewed for the binomial metrics ( $p < 0.25$ ), indicating no evidence of p-hacking. When we re-analyzed the 12 associations using random effects analyses, we found that 25.0% (3/12) meta-

analyses exhibited significant heterogeneity ( $I^2 > 75$ ). Using Egger's regression test, we observed statistical evidence of publication bias in 8.3% (1/12) of the studies. The forest plot, funnel plot, and p-curve for each association are shown in Supplementary Materials.

We identified 12 unique factors and grouped them into two categories, including prenatal and postnatal factors. Eleven significant associations between SIDS and factors were reported, and prenatal exposure to PM10 was reported that has no association with SIDS (Table 2). The results are summarized by presenting evidence maps of an umbrella review of each SIDS-associated prenatal and postnatal factor (Table 3).

### Prenatal factors

Among the five SIDS-associated prenatal factors, 80.0% (4/5) were supported by moderate certainty. SIDS was significantly associated with five risk factors: prenatal drug exposure (eOR, 7.84 [95% CI, 4.81-12.79]), opioid exposure (eOR, 9.55 [95% CI, 4.87-18.72]), methadone exposure (eOR, 9.52 [95% CI, 3.34-27.10]), cocaine exposure (eOR, 4.38 [95% CI, 1.95-9.86]), and maternal smoking (eOR, 2.25 [95% CI, 1.95-2.60]).

### Postnatal factors

Among SIDS-associated postnatal factors, one meta-analytical association had high certainty of evidence, two had moderate certainty, and three had low certainty. SIDS was significantly associated with three risk factors, including postnatal maternal smoking (eOR, 1.97 [95% CI, 1.75-2.22]), infants found with heads covered by bedclothes after last sleep (eOR, 11.01 [95% CI, 5.40-22.45]), and bed sharing (eOR, 2.89 [95% CI, 1.81-4.60]). On the other hand, we observed three protective factors, including supine position during sleep (eOR, 0.48 [95% CI, 0.37-0.63]), breastfeeding (eOR, 0.57 [95% CI, 0.39-0.83]) and use of pacifier (eOR, 0.44 [95% CI, 0.30-0.65]).

## IV. DISCUSSION

### Findings and Explanation

To our knowledge, our study was the first umbrella review about the relationship between the risk and protective factors and SIDS. Our findings can suggest recommendations for reducing SIDS based on evidence from several meta-analyses. From eight meta-analyses covering over 10 million participants, we identified 11 factors related to SIDS. We found eight factors that elevate the risk of SIDS, including prenatal drug exposure, prenatal opioid exposure, prenatal methadone exposure, prenatal cocaine exposure, prenatal and postnatal maternal smoking, bed-sharing, and head-covering. These relationships were supported by highly suggestive, suggestive, and weak evidence, except for prenatal PM10



exposure, and breastfeeding with non-significant evidence. Three factors were found as protective factors of SIDS, which are supine sleeping position, breastfeeding, and pacifier use. Their class and quality of evidence are evaluated as non-significant, weak evidence, suggestive evidence, and highly suggestive evidence, respectively.

To suggest appropriate guidance to reduce the risk of SIDS based on evidence, we summarized several guidelines for SIDS published in various institutions and countries and compared our findings to the guidelines (Supplementary Table 5). Almost all guidelines introduce prenatal maternal smoking, postnatal maternal smoking, head-covering as a risk factor for SIDS, and supine sleeping position and breastfeeding as a protective factor. In particular, many guidelines emphasize the supine sleeping position as a strong protective factor against SIDS, but our study supports this relationship with low certainty of the evidence, so we propose the need for further studies supported by higher certainty. Prenatal drug exposure, prenatal opioid exposure, prenatal methadone exposure, prenatal cocaine exposure, and pacifiers are only mentioned in some of those guidelines, so we expect to see them more in other guidelines. Overheating, sleep surface, room sharing without bed sharing are SIDS-related factors that many guidelines provide. However, additional research on these factors is expected to be needed since there is no appropriate evidence-based article supporting it.

### ***Plausible underlying mechanisms***

Many studies have been conducted to find out the pathophysiology of SIDS from the past, but no appropriate conclusion has been found. Recently, there was a consensus that SIDS is a multifactorial cause of death, so the Triple risk model has been proposed to explain SIDS. The triple risk model argues that SIDS occurs when an infant in a critical developmental period who has intrinsic vulnerability undergoes an exogenous trigger event during a critical developmental period [19].

In our study, we found that prenatal factors are related to exposure of substances, including opioids, methadone, and cocaine. To take opioids as an example, they pass the placenta, causing functional and structural changes to the developing nervous system [20]. Opioid-exposed infants have abnormal sleep patterns, apnea time prolongation, and decreased arousal of hypercapnia, which provide intrinsic vulnerability of triple risk model [21]. Maternal smoking is another factor that could provide infants' intrinsic vulnerability. Several hypotheses were suggested, including brainstem alteration or impaired lung maturation [22, 23]. Nicotine absorbed from maternal smoking can impair nicotinic acetylcholine receptors that cause abnormal cardiorespiratory responses to hypoxia [24-26].

In contrast, breastfeeding can protect infants from SIDS by reducing intrinsic vulnerability. Compared to

powdered milk, breastfeeding can provide a variety of immune substance that can prevent several infections causing intermittent apnea [27].

Factors including a head covering, bed sharing, supine sleeping position, and use of pacifier would be related to exogenous triggers of SIDS [26]. In a 2014 review of the National Institute for Health and Care Excellence, it concluded that bed sharing itself is not a risk factor [28]. Instead, when hazardous circumstances such as parental smoking, recent parental alcohol consumption, and sleeping on the sofa follow bed sharing, it could elevate the risk of SIDS, which means infants can be exposed to other risks of SIDS due to bed sharing [29]. Head-covering and supine sleep position are exogenous triggers that might be related to respiration of infant. If infants sleep in the supine position, their airway opens naturally, therefore, they need not express concerns regarding airway obstruction or aspiration [21]. In contrast, a prone position can induce hypercapnia and hypoxia due to the risk of rebreathing of expired gases [30]. Therefore, maintaining a supine position is important, especially for 4-6 months old infants before they develop the skills to choose and practice their position [31]. Pacifier use during sleep may improve autonomic control of breathing, airway patency, or both [32].

### ***Policy implication***

First of all, research considering various confounding factors should be conducted. For example, bed sharing was found as a risk factor for SIDS based on meta-analysis, but this might be due to differences in lifestyle according to culture or time [27]. Other lifestyle habits that exist only in cultures where bed sharing is common may act as a risk factor for SIDS, so it is necessary to identify what various confounding variables will be and study them in many ways [33]. Second, the guidelines to prevent SIDS should be updated frequently and should be provided to parents, especially young or ill-educated mothers. We have pointed out several factors that can be introduced more in guidelines and have proposed a need for further studies about several SIDS-related factors above based on evidence. Although the exact causal mechanism is still unknown, it is clear that the above factors are significantly related to SIDS, providing enough guidelines about SIDS is important to reduce the prevalence of SIDS [25, 34].

### ***Strengths and limitations***

To the best of our knowledge, this is the first umbrella review of meta-analyses about the factors related to SIDS. In addition to clear risk factors, factors that can lower risk are also specified, which can help in preventive measures. We have provided certainty of evidence about the relationship between SIDS and its related factors and made a proper suggestion for improving SIDS prevention guidelines. However, there are several limitations. First, this umbrella review, based on the existing meta-analyses,

acknowledges that various confounding factors for SIDS are not uniformly and fully controlled across original meta-analyses and original studies [35]. This gap highlights the need for future prospective studies to clearly understand the risk factors for SIDS. However, the ethical challenges in conducting such studies on SIDS are notable, emphasizing the value of our study in offering a comprehensive understanding despite these constraints. Second, the quality of the original texts contained in the meta-analyses was not directly evaluated [36]. Therefore, some problems may not have been identified with precision reliability. Third, various methodological approaches can be used to evaluate the sample size, heterogeneity, and statistical significance of each meta-analysis. If the mathematical method used in this umbrella review changes, certification of efficacy may also change [35]. Fourth, our umbrella meta-analysis suggested an overview and evidence for SIDS-associated factors, but further studies are needed to understand the potential causal mechanisms beyond each association. Fifth, there were no recent studies based on high evidence that we could include in our study, so we had no choice but to include relatively old studies.

## V. CONCLUSION

This umbrella review found factors related to SIDS supported by suggestive certainty, weak credibility, and no association based on several methodological approaches. Our findings suggested that eight factors increased the risk of SIDS, and three factors may be protective to SIDS. We have also compared our findings and several guidelines provided in various institutions and countries. This study suggests a need for further studies on SIDS-related factors supported by weak credibility, no association, or without adequate research paper

## ACKNOWLEDGMENT

This research was supported by grants from the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health and Welfare, Republic of Korea (grant number: HE23C002800).

**Table 1. Description of total meta-analysis to investigate the SIDS-associated factors among infants**

Outcome	First author	Published year	Included countries	AMSTAR2
<b>1. Prenatal factors</b>				
Prenatal drug exposure	Makarios L	2022	Australia, Germany, Norway, and USA	High
Prenatal opioid exposure	Makarios L	2022	Australia, Germany, Norway, and USA	High
Prenatal methadone exposure	Makarios L	2022	Australia, Germany, Norway, and USA	High
Prenatal cocaine exposure	Makarios L	2022	Australia, Germany, Norway, and USA	High
Prenatal maternal smoking	Zhang K	2013	Australia, Brazil, Denmark, England, Europe, France, Germany, Hungary, Lithuania, Netherlands, New Zealand, Nordic countries, Norway, Sweden, and USA	Low
Prenatal PM10 exposure	Kihal-Talantikite W	2020	Belgium, South Korea, UK, and USA	Low
<b>2. Postnatal factors</b>				
Postnatal maternal smoking	Zhang K	2013	Australia, Brazil, Denmark, England, Europe, France, Germany, Hungary, Lithuania, Netherlands, New Zealand, Nordic countries, Norway, Sweden, and USA	Low
Infants found with heads covered by bedclothes after last sleep	Blair PS	2008	Denmark, England, Germany, Hong Kong, Netherlands, Norway, Sweden, UK, and USA	Low
bed sharing	Vennemann MM	2012	Germany, Ireland, New Zealand, Norway, Scotland, UK, and USA	Moderate
Supine position	Priyadarshi M	2022	Australia, Brazil, Germany, Hong Kong, Ireland, Lithuania, Netherlands, New Zealand, Norway, Scotland, UK, and USA	High
Breastfeeding	Hauck FR	2011	Canada, Denmark, Germany, New Zealand, Norway, Scotland, Sweden, Tasmania, UK, and USA	Low
Pacifier use	Hauck FR	2005	Europe, Ireland, Netherlands, New Zealand, Scotland, UK, and USA	Low

SIDS, sudden infant death syndrome

**Table 2. Reanalysis of estimated effect using Der Simonian and Laird (DL) method and Hartung-Knapp-Sidik-Jonkman (HS) method, heterogeneity  $I^2$ , egger's p-value, 95% prediction interval, and CE**

Outcome	Included studies	Metrics	Total sample	Reported summary estimated effect (95% CI); random effect model	Re-analysed summary estimated effect (95% CI) DL method†			Re-analysed summary estimated effect (95% CI) HS method; random-effect model	Heterogeneity $I^2$ (%)	Tau <sup>2</sup>	Egger's p-value	95% prediction interval	CE*
					Fixed-effect model	Random-effect model	Largest study						
<b>1. Prenatal factors</b>													
Prenatal drug exposure	16	RR	675,310	<b>7.84 (5.21 to 11.81)</b>	<b>5.69 (4.98 to 6.50)</b>	<b>7.84 (5.25 to 11.72)</b>	<b>4.18 (3.41 to 5.14)</b>	<b>7.84 (4.81 to 12.79)</b>	79.87	0.38	0.17	(1.96, 31.39)	II
Prenatal opioid exposure	13	RR	675,310	<b>9.76 (5.28 to 18.05)</b>	<b>8.46 (6.85 to 10.45)</b>	<b>9.55 (5.50 to 16.56)</b>	<b>6.18 (4.60 to 8.30)</b>	<b>9.55 (4.87 to 18.72)</b>	74.84	0.59	0.72	(1.57, 57.92)	III
Prenatal methadone exposure	4	RR	675,310	<b>9.52 (4.60 to 19.70)</b>	<b>8.35 (6.25 to 11.17)</b>	<b>9.52 (4.66 to 19.45)</b>	<b>6.93 (4.92 to 9.76)</b>	<b>9.52 (3.34 to 27.10)</b>	68.27	0.32	0.72	(0.53, 171.22)	IV
Prenatal cocaine exposure	5	RR	327,046	<b>4.40 (2.52 to 7.67)</b>	<b>3.96 (3.09 to 5.07)</b>	<b>4.38 (2.55 to 7.53)</b>	<b>3.31 (2.43 to 4.52)</b>	<b>4.38 (1.95 to 9.86)</b>	63.09	0.20	0.52	(0.82, 23.47)	IV
Prenatal maternal smoking	23	OR	5,207,954	<b>2.25 (2.03 to 2.50)</b>	<b>2.46 (2.40 to 2.52)</b>	<b>2.25 (2.03 to 2.50)</b>	<b>2.50 (2.43 to 2.57)</b>	<b>2.25 (1.95 to 2.60)</b>	76.54	0.03	0.27	(1.57, 3.24)	II
Prenatal PM10 exposure	5	OR	192,332	<b>1.04 (1.01 to 1.08)</b>	<b>1.05 (1.01 to 1.08)</b>	<b>1.05 (1.00 to 1.10)</b>	1.03 (0.99 to 1.08)	1.05 (0.98 to 1.12)	37.32	0.00	0.94	(0.91, 1.20)	NS
<b>2. Postnatal factors</b>													
Postnatal maternal smoking	18	OR	789,912	<b>1.97 (1.77 to 2.19)</b>	<b>1.98 (1.86 to 2.10)</b>	<b>1.97 (1.77 to 2.20)</b>	<b>1.63 (1.43 to 1.86)</b>	<b>1.97 (1.75 to 2.22)</b>	56.60	0.02	0.55	(1.40, 2.78)	IV
Infants found with heads covered by bedclothes after last sleep	10	OR	7539	<b>9.60 (7.90 to 11.70)</b>	<b>9.55 (7.85 to 11.63)</b>	<b>11.01 (6.36 to 19.05)</b>	<b>12.00 (8.70 to 16.50)</b>	<b>11.01 (5.40 to 22.45)</b>	80.02	0.42	0.53	(2.15, 56.35)	IV
Bed sharing	11	OR	8959	<b>2.89 (1.99 to 4.18)</b>	<b>2.52 (2.02 to 3.13)</b>	<b>2.89 (1.99 to 4.18)</b>	<b>2.02 (1.35 to 3.03)</b>	<b>2.89 (1.81 to 4.60)</b>	57.35	0.20	0.07	(0.96, 8.65)	III
Supine position	26	OR	59,332	<b>0.51 (0.42 to 0.61)</b>	<b>0.48 (0.44 to 0.52)</b>	<b>0.48 (0.39 to 0.59)</b>	<b>0.49 (0.43 to 0.56)</b>	<b>0.48 (0.37 to 0.63)</b>	69.65	0.14	0.98	(0.22, 1.08)	NS
Breastfeeding	7	OR	5549	<b>0.55 (0.44 to 0.69)</b>	<b>0.55 (0.44 to 0.69)</b>	<b>0.57 (0.43 to 0.77)</b>	<b>0.50 (0.33 to 0.77)</b>	<b>0.57 (0.39 to 0.83)</b>	40.22	0.06	0.23	(0.27, 1.21)	IV
Pacifier use	8	OR	9459	<b>0.47 (0.40 to 0.55)</b>	<b>0.47 (0.40 to 0.55)</b>	<b>0.45 (0.34 to 0.58)</b>	<b>0.62 (0.46 to 0.83)</b>	<b>0.44 (0.30 to 0.65)</b>	62.79	0.09	0.38	(0.20, 0.99)	IV

CE, class and quality of evidence; CI, confidence interval; DL, Der Simonian and Laird; CE; HS, Hartung-Knapp-Sidik-Jonkman; OR, odds ratio; RR, risk relative risk. The numbers in bold indicate a significant difference ( $P < 0.05$ ).

\* Class and quality of evidence:

- Class I (convincing evidence): >1000 cases (or >20 000 participants for continuous outcomes); statistical significance at  $p < 10^{-6}$  (random effects); no evidence of small study effects and excess significance bias; 95% prediction interval excluded null value; no large heterogeneity ( $I^2 < 50\%$ ).
- Class II (highly suggestive evidence): >1000 cases (or >20 000 participants for continuous outcomes); statistical significance at  $p < 10^{-6}$  (random effects); largest study with 95% confidence interval excluding null value.
- Class III (suggestive evidence): >1000 cases (or >20 000 participants for continuous outcomes); statistical significance at  $p < 0.001$ .
- Class IV (weak evidence): remaining significant associations with  $p < 0.05$ .
- NS (non-significant):  $p > 0.05$ .

**Table 3. Evidence maps of umbrella review by SIDS-associated factors among infants**

	eOR (95% CI)	Class and quality of evidence	Direction
<b>1. Prenatal factors</b>			
Prenatal drug exposure	7.84 (4.81 to 12.79)	Highly suggestive	Associated
Prenatal opioid exposure	9.55 (4.87 to 18.72)	Suggestive	Associated
Prenatal methadone exposure	9.52 (3.34 to 27.10)	Weak	Associated
Prenatal cocaine exposure	4.38 (1.95 to 9.86)	Weak	Associated
Prenatal maternal smoking	2.25 (1.95 to 2.60)	Highly suggestive	Associated
Prenatal PM10 exposure	1.05 (0.98 to 1.12)	Non-significant	No association
<b>2. Postnatal factors</b>			
Postnatal maternal smoking	1.97 (1.75 to 2.22)	Weak	Associated
Infants found with heads covered by bedclothes after last sleep	11.01 (5.40 to 22.45)	Suggestive	Associated
Bed sharing	2.89 (1.81 to 4.60)	Weak	Associated
Supine position	0.48 (0.37 to 0.63)	Suggestive	Associated
Breastfeeding	0.57 (0.39 to 0.83)	Non-significant	Associated
Pacifier use	0.44 (0.30 to 0.65)	Weak	Associated

CI, confidence interval; eOR, equivalent odds ratio; SIDS, sudden infant death syndrome

Color represented the levels of OR and RR in data with statistically significance ( $p < 0.05$ )

# REFERENCES

1. Matthews, T., *Sudden unexpected infant death: infanticide or SIDS?* Lancet, 2005. **365**(9453): p. 3-4.
2. Willinger, M., L.S. James, and C. Catz, *Defining the sudden infant death syndrome (SIDS): deliberations of an expert panel convened by the National Institute of Child Health and Human Development.* Pediatr Pathol, 1991. **11**(5): p. 677-84.
3. Centers for Disease Control and Prevention. *Mortality in the United States, 2021.* 2022 [cited 2024 Jan, 19]; Available from: [https://www.cdc.gov/nchs/products/databriefs/db456.htm#section\\_4](https://www.cdc.gov/nchs/products/databriefs/db456.htm#section_4).
4. Park, S., et al., *The global burden of sudden infant death syndrome from 1990 to 2019: a systematic analysis from the Global Burden of Disease study 2019.* Qjm, 2022. **115**(11): p. 735-744.
5. Glinge, C., et al., *Risk of Sudden Infant Death Syndrome Among Siblings of Children Who Died of Sudden Infant Death Syndrome in Denmark.* JAMA Netw Open, 2023. **6**(1): p. e2252724.
6. Makarious, L., A. Teng, and J.L. Oei, *SIDS is associated with prenatal drug use: a meta-analysis and systematic review of 4 238 685 infants.* Arch Dis Child Fetal Neonatal Ed, 2022. **107**(6): p. 617-623.
7. Zhang, K. and X. Wang, *Maternal smoking and increased risk of sudden infant death syndrome: a meta-analysis.* Leg Med (Tokyo), 2013. **15**(3): p. 115-21.
8. Vennemann, M.M., et al., *Do immunisations reduce the risk for SIDS? A meta-analysis.* Vaccine, 2007. **25**(26): p. 4875-9.
9. Blair, P.S., et al., *Head covering - a major modifiable risk factor for sudden infant death syndrome: a systematic review.* Arch Dis Child, 2008. **93**(9): p. 778-83.
10. Vennemann, M.M., et al., *Bed sharing and the risk of sudden infant death syndrome: can we resolve the debate?* J Pediatr, 2012. **160**(1): p. 44-8.e2.
11. Priyadarshi, M., B. Balachander, and M.J. Sankar, *Effect of sleep position in term healthy newborns on sudden infant death syndrome and other infant outcomes: A systematic review.* J Glob Health, 2022. **12**: p. 12001.
12. Hauck, F.R., et al., *Breastfeeding and reduced risk of sudden infant death syndrome: a meta-analysis.* Pediatrics, 2011. **128**(1): p. 103-10.
13. Lee, S.W. and M.J. Koo, *PRISMA 2020 statement and guidelines for systematic review and meta-analysis articles, and their underlying mathematics: Life Cycle Committee Recommendations.* Life Cycle, 2022. **2**: p. e9.
14. Fusar-Poli, P. and J. Radua, *Ten simple rules for conducting umbrella reviews.* Evid Based Ment Health, 2018. **21**(3): p. 95-100.
15. Murad, M.H., et al., *When continuous outcomes are measured using different scales: guide for meta-analysis and interpretation.* BMJ, 2019. **364**: p. k4817.
16. Lee, J.S., et al., *Long-term health outcomes of early menarche in women: an umbrella review.* QJM, 2022. **115**(12): p. 837-847.
17. Lee, S.W., *Methods for testing statistical differences between groups in medical research: statistical standard and guideline of Life Cycle Committee.* Life Cycle, 2022. **2**: p. e1.
18. Solmi, M., et al., *Balancing risks and benefits of cannabis use: umbrella review of meta-analyses of randomised controlled trials and observational studies.* BMJ, 2023. **382**: p. e072348.
19. Spinelli, J., et al., *Evolution and significance of the triple risk model in sudden infant death syndrome.* J Paediatr Child Health, 2017. **53**(2): p. 112-115.
20. Abu, Y. and S. Roy, *Prenatal opioid exposure and vulnerability to future substance use disorders in offspring.* Exp Neurol, 2021. **339**: p. 113621.
21. Moon, R.Y., R.F. Carlin, and I. Hand, *Sleep-Related Infant Deaths: Updated 2022 Recommendations for Reducing Infant Deaths in the Sleep Environment.* Pediatrics, 2022. **150**(1).
22. Salihu, H.M. and R.E. Wilson, *Epidemiology of prenatal smoking and perinatal outcomes.* Early Hum Dev, 2007. **83**(11): p. 713-20.
23. Shin, Y.H., et al., *Autoimmune inflammatory rheumatic diseases and COVID-19 outcomes in South Korea: a nationwide cohort study.* Lancet Rheumatol, 2021. **3**(10): p. e698-e706.
24. McGrath-Morrow, S.A., et al., *The Effects of Nicotine on Development.* Pediatrics, 2020. **145**(3).
25. de Visme, S., et al., *Inconsistency Between Pictures on Baby Diaper Packaging in Europe and Safe Infant Sleep Recommendations.* J Pediatr, 2023. **264**: p. 113763.
26. Harrington, C.T., N.A. Hafid, and K.A. Waters, *Butyrylcholinesterase is a potential biomarker for Sudden Infant Death Syndrome.* EBioMedicine, 2022. **80**: p. 104041.
27. Perrone, S., et al., *Sudden Infant Death Syndrome: Beyond Risk Factors.* Life (Basel), 2021. **11**(3).
28. Tappin, D., et al., *Bed-sharing is a risk for sudden unexpected death in infancy.* Archives of Disease in Childhood, 2023. **108**(2): p. 79-80.
29. Marinelli, K.A., et al., *An Integrated Analysis of Maternal-Infant Sleep, Breastfeeding, and Sudden Infant Death Syndrome Research Supporting a Balanced Discourse.* J Hum Lact, 2019. **35**(3): p. 510-520.



30. Scragg, R.K., et al., *Infant room-sharing and prone sleep position in sudden infant death syndrome. New Zealand Cot Death Study Group.* Lancet, 1996. **347**(8993): p. 7-12.
31. Nelson, E.A., et al., *Rolling over in infants: age, ethnicity, and cultural differences.* Dev Med Child Neurol, 2004. **46**(10): p. 706-9.
32. Smith, R.W. and M. Colpitts, *Pacifiers and the reduced risk of sudden infant death syndrome.* Paediatr Child Health, 2020. **25**(4): p. 205-206.
33. Pease, A., et al., *Changes in background characteristics and risk factors among SIDS infants in England: cohort comparisons from 1993 to 2020.* BMJ Open, 2023. **13**(10): p. e076751.
34. Moon, R.Y., R.F. Carlin, and I. Hand, *Evidence Base for 2022 Updated Recommendations for a Safe Infant Sleeping Environment to Reduce the Risk of Sleep-Related Infant Deaths.* Pediatrics, 2022. **150**(1).
35. Kim, J.H., et al., *Environmental risk factors, protective factors, and peripheral biomarkers for ADHD: an umbrella review.* Lancet Psychiatry, 2020. **7**(11): p. 955-970.
36. IntHout, J., J.P. Ioannidis, and G.F. Borm, *The Hartung-Knapp-Sidik-Jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard DerSimonian-Laird method.* BMC Med Res Methodol, 2014. **14**: p. 25.

# Marketing Insights for Korean Medicine Clinics Through Consumer Review Analysis: A Focus on Latent Dirichlet Allocation Methodology

Chomyong Kim<sup>1</sup>, Yunyoung Nam<sup>2</sup>

<sup>1</sup>ICT Convergence Research Centre, Soonchunhyang University, Asan, South Korea

<sup>2</sup>Department of Computer Science and Engineering, Soonchunhyang University, Asan, South Korea

\*Contact: ynam@sch.ac.kr

**Abstract**— Medical practices must comprehend and act upon consumer feedback in order to succeed in the ever-changing world of healthcare services. This is especially true for Korean traditional medicine clinics, where patient experiences have a significant impact on how well patients are served overall as well as how well they receive care. The rise in online customer reviews in recent years has given rise to a useful collection of unstructured data reflecting the feelings and viewpoints of people who have visited these clinics for treatments.

With a focus on Latent Dirichlet Allocation (LDA), this research uses advanced text mining techniques to contribute to the rapidly growing field of healthcare marketing. With the help of LDA, a potent probabilistic model, we can uncover hidden subjects inside a corpus of textual data, exploring the various themes and issues that customers have raised about Korean traditional medicine clinics. By utilizing LDA, we want to identify the fundamental trends in customer feedback and offer practical recommendations for improving marketing tactics and maximizing patient experiences in the field of oriental medicine.

## I. INTRODUCTION

In the dynamic landscape of healthcare services, understanding and responding to customer feedback is paramount for the success of medical practices. This holds true for Korean traditional medicine clinics, where the nuances of patient experiences play a crucial role in shaping both service delivery and overall patient satisfaction. In recent years, the surge in online consumer reviews has provided a valuable repository of unstructured data that reflects the sentiments and opinions of individuals who have sought services in these clinics.

The influential role of online reviews in shaping consumers' purchasing decisions has been acknowledged over an extended period. Consequently, the significance of online reviews in the realm of product sales is noteworthy. Empirical studies have indicated that the volume, rating, and sentiment analysis of online reviews exert a discernible impact on sales outcomes. Furthermore, it has been discerned that both the quantity and rating of reviews positively correlate with increased sales. In light of these research findings, strategic marketing approaches can be formulated to enhance product sales through the judicious utilization of online reviews [1]. Elevating customer satisfaction is integral to augmenting sales, constituting a key objective within marketing strategies [2]. In the pursuit of

enhancing customer satisfaction, marketing strategies progress through various stages, which encompass the meticulous formulation of plans related to crucial elements such as the marketing mix, branding initiatives, and strategies associated with relationship marketing. Furthermore, these strategies extend to encompass specific considerations regarding product offerings, pricing structures, promotional activities, and distribution channels, all of which collectively aim to advance and optimize the overall satisfaction of customers [3-4]. Furthermore, within the realm of service design aiming to enhance customer satisfaction, the research and application of service quality are of paramount importance, playing a pivotal role in establishing brand image and generating market impact. Through these efforts, businesses significantly contribute to their development and competitive edge. Noteworthy is the substantial impact that service quality has on consumer satisfaction, experience, and brand loyalty. Shi's (2020) research underscores the leading role of the SERVQUAL model, extending its influence not only to sectors like retail, tourism, and other services but also notably within the healthcare services domain [5].

For example, in the United States, qualifications such as Acupuncturist and Herbalogist can be obtained by passing the NCCAOM (National Certification Commission for Acupuncture and Oriental Medicine) exams. In contrast, in Korea, the practice of Oriental Medicine involves a total of six years of education, comprising a two-year pre-medical program and a four-year traditional Oriental Medicine program. Subsequently, individuals must pass the national licensing examination to prescribe treatments like acupuncture, herbal remedies, and nutritional therapies. In essence, the medical services provided by Oriental Medicine clinics in Korea constitute specialized medical practices. According to the 2022 Korean Medicine Utilization Survey conducted by the Ministry of Health and Welfare, the primary purpose of utilizing Oriental Medicine in Korea is disease treatment (94.2%), with health promotion and beauty enhancement accounting for 14.9%. Among the treated conditions, musculoskeletal disorders are the most prevalent at 74.8%, followed by injuries, intoxication, and external factors at 35.5%, health supplement treatment at 12.8%, digestive system disorders at 8.1%, traffic accident sequelae at 7.3%, and respiratory system disorders at 6.8%. The therapeutic methods commonly used include acupuncture

(94.3%), cupping therapy (56.5%), moxibustion (54.6%), traditional Korean physical therapy excluding notification (44.5%), herbal medicine (28.5%), herbal acupuncture (28.4%), external herbal applications (26.7%), and nutritional therapy (9.4%). Particularly, there has been an increase in health promotion and beauty enhancement, rising from 13.5% in 2020 to 14.9%. Additionally, based on a report in 2022, obtaining information about Oriental Medicine clinics is primarily through word-of-mouth from family and friends (39.3%), followed by media broadcasts (27.6%), and internet websites (8.2%) [6]. This highlights the importance of review analysis in intuitively understanding consumer satisfaction based on service quality.

SERVQUAL is a widely used model in service quality management. Developed by Parasuraman, Zeithaml, and Berry, it assesses service quality by comparing customer expectations with their perceptions of the actual service received. The model identifies five key dimensions: reliability, assurance, tangibles, empathy, and responsiveness, providing a framework to understand and improve the quality of services across various industries [7].

Therefore, this study aims to utilize the Latent Dirichlet Allocation (LDA) technique to analyze consumer reviews with the goal of gaining insights that can fulfill the five dimensions of SERVQUAL, and subsequently making marketing recommendations.

## II. RELATED WORK

According to Choi et al. (2020) study utilized R to collect data on best-selling Bluetooth speakers from Amazon between August 13, 2019, and September 7, 2019. Subsequent sentiment analysis revealed that the quantity and rating of reviews positively influenced sales. Further analysis showed that titles with positive words correlated with higher sales, while titles with negative words were associated with lower sales. Negative words in the review content were linked to lower sales, while positive words did not significantly affect sales. Additionally, the interaction of brand reputation, review quantity, and sentiment index demonstrated that, in cases of low brand reputation, the impact of review quantity on sales was more significant. Both positive and negative scores in review titles and content influenced sales accordingly [8]. The study performed text preprocessing by splitting the text into two-word segments and conducted sentiment analysis using a sentiment dictionary to classify them as positive or negative. Although it revealed that the presence of negative words in the review content had a negative impact on sales, the study has a limitation in that it could not ascertain the reasons behind leaving negative reviews.

Choi et al. (2018) collected 28,924 reviews of a specific hospital from Yelp spanning from October 2005 to August 2016. The researchers classified these reviews into 264,565 sentences and employed Keyword Extraction Analysis (KEA) to extract representative keywords. Utilizing SERVQUAL's five dimensions, they categorized the keywords, considering words such as medical, surgery, service, doctor, time, care, hospital, office, nurse, excluding "patient," as relevant to SERVQUAL. The authors extracted sentences containing the identified representative words, and two experts selected words based on agreement, resulting in 19 dimensions of tangibility, 11 of

reliability, 13 of responsiveness, 13 of assurance, and 11 of empathy. After collecting sentences containing these words separately, researchers, with a one-month interval, verified twice whether the sentences were correctly classified into SERVQUAL units. The accuracy averaged 86.11%, with specificity for tangibility at 73.89%, reliability at 98.10%, responsiveness at 88.00%, assurance at 96.19%, and empathy at 82.67%. To assess the emotional scores of the sentences, they used the AFINN English dictionary, which categorizes words into very negative (-5, -4), negative (-3, -2, -1), positive (1, 2, 3), and very positive (4, 5). They calculated the frequency of words in sentences, multiplied by the average sentiment score of each group, and combined the sentences to calculate scores at the review level. Finally, a simple regression analysis verified that the scores of SERVQUAL's five dimensions influenced the review scores extracted by the researchers [9]. However, in this study, the sentences directly collected by researchers were classified into SERVQUAL's five dimensions. Due to this approach, it is challenging to ascertain the frequency and importance of words that can be explained for each dimension. This limitation may pose difficulties in formulating marketing strategies. Therefore, to propose marketing strategies for the satisfaction of traditional Korean medicine consumers, it is necessary to categorize reviews into positive and negative sentiments. Subsequently, one should examine which elements align with the five dimensions of SERVQUAL. In this regard, the utilization of Latent Dirichlet Allocation (LDA), a clustering technique in text mining, can be beneficial.

## III. METHODOLOGY

### A. Data Collection

This study collects information from medical service consumer review platform 'Modoodoc' by searching for Korean traditional medicine clinics. The target includes hospitals with a minimum of 2 reviews, and data such as hospital name, address, hospital rating, reviews, and review ratings are collected. A web crawler is developed using the Python programming language to collect comments.

### B. Preprocessing

The 1st quartile value of hospital ratings is calculated, and ratings above this value are classified as positive reviews, while those below are classified as negative reviews. To assess the importance of words within a document collection (corpus), Term Frequency-Inverse Document Frequency (TF-IDF), a numerical statistic, is utilized to extract representative nouns and adjectives.

TF-IDF is a frequently used technique in conjunction with word frequency analysis for extracting key keywords. Translated as 'Term Frequency-Inverse Document Frequency,' it assigns importance to each word in a Document-Term Matrix based on the frequency of the word and the inverse document frequency, which is a specific formula applied to the frequency of the word across documents. TF-IDF is employed in tasks such as calculating document similarity, determining importance in search systems, and extracting the importance of specific words in a document for use as key keywords. The formula for calculating TF-IDF is as follows:

$$tf(d, t) \times \log\left(\frac{n}{1 + df(t)}\right)$$

### C. Modelling: Latent Dirichlet Allocation (LDA)

For a last, the LDA technique provided by the Python library Scikit-learn is utilized. In this case, the number of topics (LDA results) is set to 6.

Latent Dirichlet Allocation (LDA), initially introduced by Blei et al. (2003), is a probabilistic model widely used in natural language processing and machine learning. It serves as a generative statistical model to discover topics based on the distribution of words in unstructured data sets [10-11]. In the context of text mining and topic modelling, LDA assumes that a document is a mixture of topics, and each word in the document is derived from a specific topic. Furthermore, LDA assumes that words in a document are associated with certain topics, and the distribution of topics in a document follows a Dirichlet distribution. This modelling approach enables the discovery of underlying topics in a collection of documents, contributing to extracting meaningful insights and patterns from large text datasets.

### D. Visualization

After visualizing the results using the pyLDAviz API, the topic distributions are examined to avoid overlap. Subsequently, the representative vocabulary for each topic is interpreted in alignment with the SERVQUAL 5 dimensions. The SERVQUAL 5 dimensions are as follows:

- **Tangible:** Physical facilities, equipment, and the appearance of staff, as well as communication materials.
- **Reliability:** The ability to believe in and accurately perform promised services.
- **Responsiveness:** Willingness to help customers willingly and provide prompt service.
- **Assurance:** The knowledge and competence of staff, politeness, reliability, and the ability to install confidence and safety.
- **Empathy:** Understanding and consideration of customers' personal requirements, accessibility, and smooth communication.

## IV. EXPERIMENTS AND RESULTS

To collect reviews written after purchasing medical services at actual Korean Medicine clinics, I developed a web crawler using the Python programming language. I gathered reviews from 3,120 clinics with 9,369 reviews by searching for 'Korean Medicine' on the 'Modoodoc' platform.

Region	Clinics	Reviews	min	Quantile (1 <sup>st</sup> )	mean	median
Seoul	681	3705	1.3	8.5	8.97	9.4
Gyeonggi	529	2203	1	8.7	9.11	9.5
Busan	166	676	1	8.67	9.03	9.3
Incheon	115	461	4.3	8.8	9.19	9.5
Daegu	99	418	3	8.72	9.13	9.5
Daejeon	86	346	4.3	8.72	9.14	9.5
Gwangju	70	214	3.3	8.8	9.13	9.5

Gyeongsangnam-do	65	243	5.3	8.8	9.2	9.5
Chungcheongnam-do	49	168	5.5	8.5	9.12	9.3
Jeollabuk-do	49	197	1	8.8	9.15	9.3
Gyeongsangbuk-do	47	186	1.8	9.15	9.34	9.8
Chungcheongbuk-do	40	119	4.8	8.9	9.23	9.5
Gangwon-do	39	158	4.3	8.82	9.27	9.5
Ulsan	34	106	6	9	9.36	9.5
Jeollanam-do	23	69	3	8.6	9.06	9.3
Sejong	14	48	7.3	8.8	9.3	9.5
Jeju	13	52	1	8.8	9.15	9.5

Table 1. Descriptive statistics of review dataset

Subsequently, to classify into positive and negative reviews, the first quartile values and average first quartile values of review scores for each region were calculated. Reviews with scores of 8.77 or higher were classified as positive reviews, while those below this threshold were classified as negative reviews. As a result, 6,905 positive reviews and 2,464 negative reviews were identified (Table 1).

After excluding unnecessary emoticons and onomatopoeic expressions from the reviews, the base forms of nouns and adjectives were extracted. The scikit-learn API's TF-IDF API was then utilized to extract 2,000 key words. Words that occurred less than 10 times were removed. Additionally, common Korean stop words from the default NLTK Korean stop words dictionary and irrelevant words such as '한줄평' (short comment), '리뷰' (review), '한의원' (Korean medicine clinic) frequently repeated on Modoodoc were eliminated.

The parameters for LDA were set to extract 6 topics for positive reviews and 3 topics for negative reviews. Other parameters such as batch size (128) and the number of learning iterations (10) were standardized. As a result, representative keywords for each of the 20 positive and 20 negative reviews, along with topic-specific representative keywords, were extracted.

The representative keywords for positive reviews were extracted in the following order: "clean, kind, results, detailed, explanation, staff, physical constitution, diet, rhinitis, herbal medicine, pain, ankle, acne, digestion, distance, weekend, hospitalization, friend, family, okay, atmosphere, pulse diagnosis, waist, recommendation, prescription, introduction, acquaintance, traffic accident, rumor, nutrition." For negative reviews, the representative keywords were extracted as follows: "detailed, clean, kind, results, staff, herbal medicine, explanation, nutrition, diet, constitution, counseling, nearby, famous, acquaintance, reservation, prescription, pain, sick, frequency, satisfaction, location, waist, improvement, intake, massage, interior, old, clean, massage, wrist."

When the extracted adjective base form keywords and some nouns are classified into the 5 dimensions of SERVQUAL (Table 2).

SERVQUAL	KEYWORDS
Tangible	Clean, Old, Facility, Distant, Weekend, Nearby, Atmosphere
Reliability	Detailed, Effect, Result
Responsiveness	Recommend, Prescription, Counseling
Assurance	Detailed, Thorough, Famous, Painful

Empathy	Kind, Explanation
---------	-------------------

Table 2. Keywords, based on SERVQUAL 5 dimensions.

This suggests that consumers tend to leave positive reviews when they are satisfied with each dimension, and in case of dissatisfaction, they are more likely to leave negative reviews. Additionally, as mentioned earlier, visits to oriental medicine clinics can be attributed to musculoskeletal disorders (such as back pain, ankle pain, general pain, traffic accidents, hospitalization, and nutrition) and visits for health promotion and beauty purposes ((physical constitutional-based) diet, rhinitis, acne). Visits prompted by recommendations, referrals, or suggestions from friends or family are also observed. This similarity in findings between the survey results and LDA's representative keywords indicates a consistent pattern.

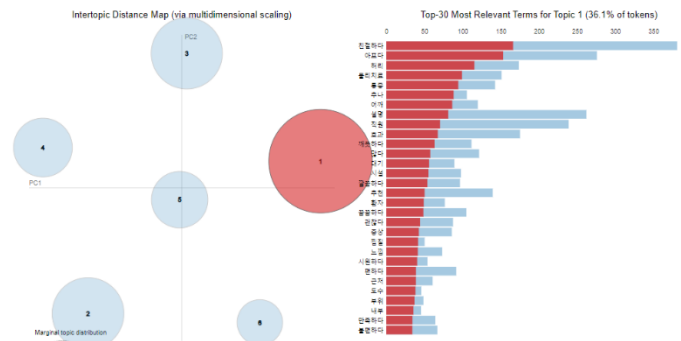
#### A. Positive Review analysis

The keywords for the 6 topics in positive reviews are as follows.

Topic1	Topic2	Topic3
Kind	Herbal Medicine	Kind
Painful	Diet	Painful
Waist (Spine)	(physical) constitution	(Actual)Pain
Physical Therapy	Prescription	Ankle
Pain	Consultation	Waist (Spine)
Chuna Manual Therapy	Kind	Effect
Shoulder	Effect	Weekend
Explanation	feel the pulse	Recommendation
Staff	Explanation	Physical Therapy
Effect	Condition	Wrist
Clean	Reservation	Atmosphere
Many	Take a dose	Treatment
Waiting	Many	Herbal Medicine
Facility	Famous	Staff
Cleanliness	Recommendation	constant
Recommendation	Help	comfortable
Patient	Life	Detailed
Detailed	Question	Overtreatment
Fine	Detailed	Explanation
Symptom	Speech of doctor	Headache
Topic4	Topic5	Topic6
Kindness	Rhinitis	Kind
Result	Acne	Digestion
Cleanliness	Friend	Hospitalization
Detailed	Far	Fine
Explanation	Suggestion	Pain
Staff	Introduction	Staff
Kind	Physiotherapist	Facility
Family	Positive rumours	Traffic Accident
Pain	Acquaintance	Location
Mother	Heavy	Many
Cold	Face	Building
Elders	Solution	Clean
Be moved	Age	Bad
Work	Registration	Recommendation
Effect	Work	Old
Menstrual Pain	KyungHee	Village
Western Medicine	Atmosphere	injection
Waist (Spine)	Pain	Effect
Recommendation	Far	Faulty
Clean	Head	Cleanliness

Table 3. Representing 20 keywords each positive topic

Fig 1. Distribution and Most Relevant Terms for positive Topic 1 (Topic number on graph is ignored)



Inferred and interpreted from Topic 1, patients visited a Korean Medicine clinic for Chuna Manual Therapy, including physical therapy, due to musculoskeletal disorders such as back and shoulder pain. While waiting at the clinic, they perceived the facilities as neat and clean. Furthermore, the thorough explanation of symptoms and effective treatment outcomes contributed to a positive perception, and the staff members were considered friendly.

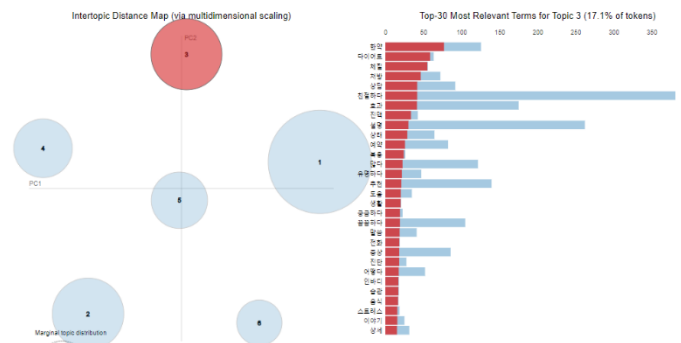


Fig 2. Distribution and Most Relevant Terms for positive Topic 2 (Topic number on graph is ignored)

Topic 2 suggests that individuals visited a renowned Korean Medicine clinic or one recommended to receive herbal prescriptions for dieting. They made appointments and, during the visit, obtained answers to questions about dieting. Additionally, they underwent constitutional examinations through pulse diagnosis, received prescriptions tailored to their constitution, and received counseling on aspects beneficial to their lifestyle.

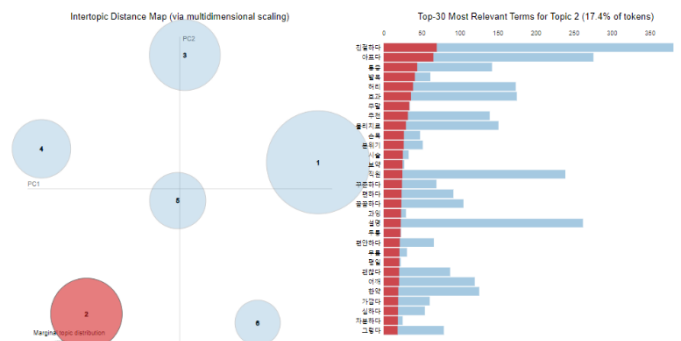


Fig 3. Distribution and Most Relevant Terms for positive Topic 3 (Topic number on graph is ignored)

Topic 3 appears similar to Topic 1, but it suggests that the Korean Medicine clinic is open on weekends and does not engage in excessive treatment.

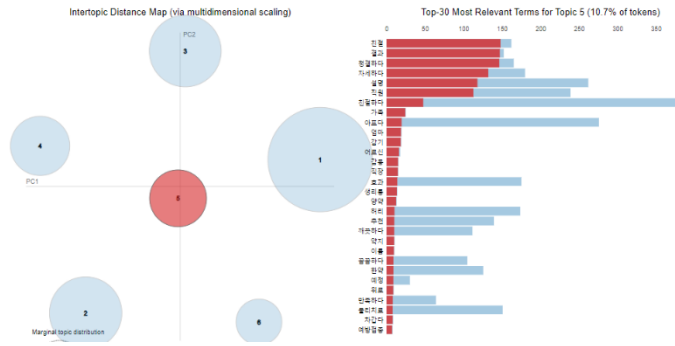


Fig 4. Distribution and Most Relevant Terms for positive Topic 4 (Topic number on graph is ignored)

Topic 4 represents cases where individuals visited the Korean Medicine clinic for symptoms related to 'menstrual pain' rather than musculoskeletal issues. It seems to be for addressing symptoms that could not be treated with Western Medicine, and the visit might have been influenced by family recommendations.

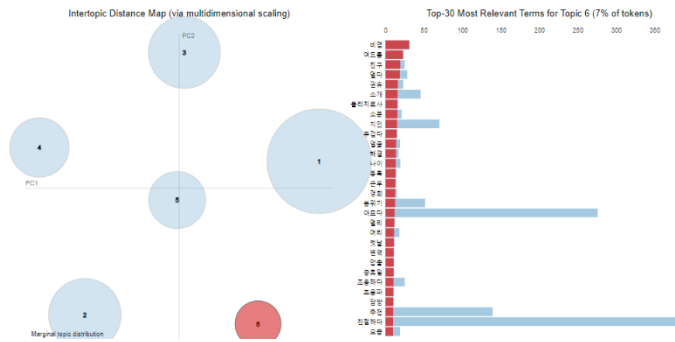


Fig 5. Distribution and Most Relevant Terms for positive Topic 5 (Topic number on graph is ignored)

Topic 5 also appears to involve visits for health promotion and beauty purposes (such as acne and rhinitis treatment), likely influenced by recommendations or introductions.

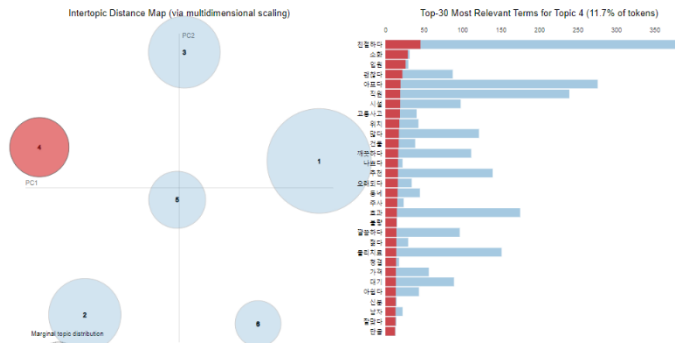


Fig 6. Distribution and Most Relevant Terms for positive Topic 6 (Topic number on graph is ignored)

Topic 6 seems to involve visits for gastrointestinal issues or traffic accidents. The fact that the building is not old, close, and clean suggests satisfaction with SERVQUAL's tangibles dimension.

## B. Negative Review analysis

The keywords for the 6 topics in positive reviews are as follows.

Topic1	Topic2	Topic3
Explanation	Painful	Herbal medicine
Staff	Kind	Effects
Detailed	Waist (Spine)	Consultation
Cleanliness	Pain	Many
Kindness	Physical Therapy	Kind
Kind	Chuma manual Therapy	Reservation
Result	Shoulder	Diet
Painful	Effects	Recommendation
Clean	Symptom	(physical) constitution
Satisfaction	Nearby	Prescription
Neat	Comfortable	Famous
Waist (Spine)	Manual Therapy	Acquaintance
Facility	Clean	(Neutral Verb)
Location	Improvement	Patient
Physical Therapy	Regional	Take a dose
Inside Clinic	Fine	Waiting
Many	Facility	Progress
Shoulder	Wrist	Detailed
Old	Severe	Condition
Pain	Explanation	Feel the pulse

Table 4. Representing 20 keywords each negative topic



Fig 7. Distribution and Most Relevant Terms for negative Topic 1 (Topic number on graph is ignored)

Topic 1 of negative reviews seems to involve visits for physical therapy, but the explanation was not detailed or friendly. Additionally, issues with the location, outdated interior facilities, and problems with the attitude of the staff (lack of friendliness) can be inferred.

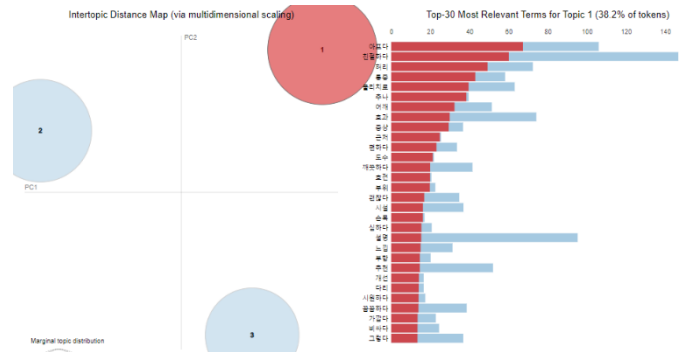


Fig 8. Distribution and Most Relevant Terms for negative Topic 2 (Topic number on graph is ignored)





## REFERENCES

- [1] Kühne, F., Maas, J., Wiesenthal, S., et al. (2019). Empirical research in clinical supervision: A systematic review and suggestions for future studies. *BMC Psychology*, 7(1), 54. <https://doi.org/10.1186/s40359-019-0327-7>
- [2] Otto, A. S., Szymanski, D. M., & Varadarajan, R. (2020). Customer satisfaction and firm performance: Insights from over a quarter century of empirical research. *Journal of the Academy of Marketing Science*, 48(4), 543–564. <https://doi.org/10.1007/s11747-019-00657-7>
- [3] Fong, S.-F., Loh, R.-Y., & Choi, S.-L. (2022). Marketing Strategies and Customer Satisfaction: A Study on the Higher Education Institutions in Johor. *Business and Economic Research*, 12(2), 61-83.
- [4] Fatah, S., & Ali, B. (2018). The impact of marketing strategy on customer satisfaction for e-learning: A marketing strategies model approach. *International Journal of Computer Science and Information Security*, 16, 95-102.
- [5] Shi, Z., & Shang, H. (2020). A review on quality of service and SERVQUAL model. In F.H. Nah & K. Siau (Eds.), *HCI in Business, Government and Organizations. HCII 2020* (Vol. 12204, pp. 15). Springer. [https://doi.org/10.1007/978-3-030-50341-3\\_15](https://doi.org/10.1007/978-3-030-50341-3_15)
- [6] Korean Ministry of Health and Welfare. (2023). Utilization Survey of Traditional Korean Medicine in 2022.
- [7] Parasuraman, A., Zeithaml, V. A., & Berry, L. L. (1988). SERVQUAL: A multiple-item scale for measuring consumer perceptions of service quality. *Journal of Retailing*, 64(1), 12–40.
- [8] Choi, J. Y., Kim, H. A., & Kim, Y. B. (2020). The impact of online review volume, rating, and sentiment score on sales: Focusing on the moderating effect of brand reputation. *Journal of Channel Retailing*, 25(3), 1-21. <https://doi.org/10.17657/jcr.2020.07.31.1>
- [9] Choi, J.-E., Kim, S., & Kim, H.-W. (2018). A study on sentiment score of healthcare service quality on the hospital rating. *Information Systems Review*, 20(2), 111-137.
- [10] Blei, D. M., Ng, A. Y. & Jordan, M. I. (2003). Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3, 993--1022. doi: <http://dx.doi.org/10.1162/jmlr.2003.3.4-5.993>
- [11] Jelodar, H., Wang, Y., Yuan, C. et al., (2019). Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey. *Multimed Tools Appl* 78, 15169 – 15211. doi.org/10.1007/s11042-018-6894-4

# National trends in the prevalence of self-perceived overweight among adolescents, 2005-2022: a nationwide representative study in South Korea

Kyeongmin Lee,<sup>1</sup> Selin Woo<sup>1\*</sup>

<sup>1</sup>Center for Digital Health, Medical Science Research Institute, Kyung Hee University Medical Center, Kyung Hee University College of Medicine, Seoul, South Korea

\*Correspondence: Selin Woo (dntpfls@naver.com)

**Abstract**— Despite several studies on self-evaluation of health and body shape, existing research on the risk factors of self-perceived overweight is insufficient, especially during the COVID-19 pandemic. Thus, this study focuses on elucidating the impact of risk factors on self-perceived overweight and how the prevalence of self-perceived overweight changed before and during the COVID-19 pandemic era. The data used in the study was obtained from middle and high school students who participated in the Korean Youth Risk Behavior Web-based Survey (KYRBS; total n=1,189,586). We grouped the survey results by years and estimated the slope in prevalence of self-perceived overweight before and during the pandemic, as well as the prevalence tendencies of self-perceived overweight according to various risk factors. The prevalence of self-perceived overweight is much higher than BMI-based overweight among 1,189,586 middle and high school participants (grade 7th-12th) from 2005 to 2022. From 2005 to 2019 (pre-pandemic) the prevalence of self-perceived overweight increased, but from 2020 to 2022 (pandemic) decreased. During the pandemic, individuals with higher levels of stress or lower economic status of households exhibited a more significant decrease in the rate of self-perceived overweight. The prevalence of self-perceived overweight tends to be higher among individuals with lower school performance, lower economic status, poorer subjective health and higher stress level. This nationwide study conducted over 18 years indicates that the prevalence of self-perceived overweight decreased during the COVID-19 pandemic. These findings suggest the necessity of facilities and policies for treatment of eating disorders and mental illnesses especially for adolescents with risk factors.

## I. INTRODUCTION

Amid the ongoing global health crisis, the psychological implications of the COVID-19 pandemic on body image perception needs to be investigated.<sup>1</sup> The World Health Organization (WHO) defines 'overweight' and 'obesity' as having a body mass index (BMI) of 25 or higher, and 30 or higher, respectively, highlighting excessive fat accumulation as a health hazard.<sup>2</sup> Despite these clear medical guidelines, individuals' perception of weight can diverge significantly from these standards.<sup>3</sup>

Self-perceived overweight, which refers to the distressing experience wherein an individual perceives themselves to be overweight irrespective of their actual BMI status, often leads

to rigorous dieting, body dissatisfaction, and potential eating disorders.<sup>4,5</sup> Given the high prevalence of eating disorder, especially among adolescents,<sup>6</sup> adolescents' self-perceived overweight must be researched.

Furthermore, as body dissatisfaction during adolescents can continue into adulthood,<sup>7</sup> identifying the risk factors for self-perceived overweight during adolescence is crucial. Nevertheless, prior research exploring body dissatisfaction has been limited by small sample sizes that challenge the veracity of findings (n=498), and also do not encompass the COVID-19 pandemic era, hence lacking data to assess the impact of this pandemic on self-perceived overweight.<sup>5</sup>

As a result, it's clear that there is a pressing need to analyze the prevalence of self-perceived overweight among adolescents, including the COVID-19 pandemic era, and identifying risk factors. This study examined the long-term and large-scale trends in the prevalence of overweight perception among adolescents, including the period of COVID-19, and identified risk factors. The finding of this study may be beneficial in developing specific public health strategies to encourage a positive body image and prevent negative outcomes. The ultimate goal of our study is to provide adolescents the resources and necessary understanding to cultivate a healthy self-perception, thereby enhancing overall well-being and reducing the risk of harmful behaviors.

## II. METHODS

### A. Patient selection and data collection

In this study, we used the nationwide Korea Youth Risk Behavior Web-based Survey (KYRBS) for a total 18 years from 2005 to 2022 to investigate the prevalence of self-perceived overweight among adolescents.<sup>8</sup> The KYRBS is an annual survey conducted by Korea Disease Control and Prevention Agency (KDCA) and the Ministry of Education to examine the health behavior statistics of the youth in Korea.<sup>9</sup> KYRBS surveys includes categories such as smoking, drinking, and obesity, and students voluntarily participate in anonymous online survey. Our study focused on students in grades 7th to 12th from 800 schools, aged 13 to 18, with an average response rate over 95.0%. The study protocol was

approved by the Institutional Review Board of the KDCA and Kyung Hee University (KNUH 2022-06-042), and all participants provided written informed consent. This study was conducted in accordance with the principles of the Declaration of Helsinki.

### B. Ascertainment of considering self-perceived overweight

The purpose of the study is to examine the proportion and trend of people consider themselves overweight among adolescents in South Korea, spanning from 2005 to 2022, including the era of the COVID-19 pandemic. To examine the number of students who reported self-perception of overweight, the students were asked “How would you describe your body type?” with 5 multiple choice options: very thin, bit thin, average, bit fat, very fat. The two options, bit fat and very fat, were combined to form the category ‘self-perceived overweight’.<sup>10</sup>

### C. Covariates

Eleven covariates were used in the analysis: grade (7th to 9th grade: middle school; 10th to 12th grade: high school), sex, region of residence (urban and rural), body mass index (BMI) group (underweight, normal, overweight, and obese), school performance (high, middle-high, middle, middle-low, and low), stress level (high, middle-high, middle, middle-low, and low),<sup>11</sup> subjective health status (very healthy, healthy, normal, and unhealthy), smoking status within one month of the survey,<sup>12</sup> alcohol consumption within one month of the survey,<sup>13</sup> and economic status of households (high, middle-high, middle, middle-low, and low).<sup>14</sup> BMI was calculated as weight in kilograms divided by height in meters squared based on student self-reported weight and height. Following the 2017 Korean National Growth Charts, BMI was divided into four groups: underweight (<5 percentile), normal (5 to 84 percentile), overweight (85 to 94 percentile), and obese ( $\geq 95$  percentile).<sup>14</sup> School performance, stress level, and economic status of households were divided into five groups according to the self-reports of participants: low (<20 percentile), middle low (20 to 39 percentile), middle (40 to 59 percentile), middle high (60 to 79 percentile), and high ( $\geq 80$  percentile). The definitions of all covariates were derived from established, peer-reviewed literature.<sup>15,16</sup>

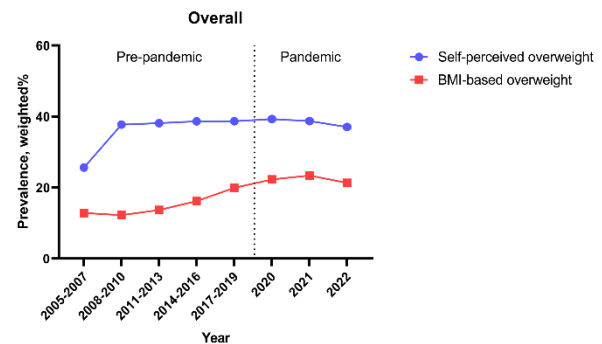
### D. Statistical analyses

In our study, we presented unweighted crude analysis results as frequencies and proportions to represent the overall characteristics of the study population. In contrast, the weighted Composite Sample Analysis with weights provided by KDCA was expressed using weighted percentages and 95% confidence intervals (CI) for the results. The prevalence of self-perceived overweight was calculated by grouping KYRBS data classified by year from 2005 to 2022 into three-year intervals. And the prevalence of self-perceived overweight and all regression model analyses were calculated, considering various variables such as grade, sex, region of residence, BMI, school performance, stress level, subjective health status, smoking status, alcohol consumption, and economic status of households. To calculate the 95% CI for weighted odds ratio (wOR) and the 95% CI for  $\beta$  coefficient, both binomial

logistic regression and linear logistic regression models were employed.<sup>17</sup> In this study, 95% CI values for  $\beta$ -coefficients were computed for analysis of trend in both pre-pandemic and during the pandemic, and  $\beta$  difference ( $\beta$ diff) was calculated to assess the impact of the COVID-19 pandemic on self-perceived overweight.<sup>18</sup> Additionally, we investigated which variables influence the vulnerability of the prevalence of self-perceived overweight. We conducted statistical analyses using SAS software (version 9.4; SAS Institute Inc., Cary, NC, USA) employing a two-sided test, and statistical significance was defined as a p-value less than 0.05.

## III. RESULTS

Table 1 shows the general characteristics of the participants. From 2005 to 2022, a total of 1,189,586 adolescents were enrolled in the KYRBS, as follows: grade (7th to 9th grade, 50.35% [95% CI, 49.97 to 50.73] and 10th to 12th grade, 49.65% [95% CI, 49.27 to 50.03]) and sex (male, 52.25%, [95% CI, 51.64 to 52.86] and female, 47.75% [47.14 to 48.36]).



Abbreviations: BMI, body mass index.

Fig. 1 Nationwide trend in self-perceived overweight and BMI-based overweight prevalence over 18 years (2005-2022) among Korean adults (n=1,189,586).

Figure 1 indicates that the prevalence of self-perceived overweight is much higher than BMI-based overweight (defined as individuals in overweight or obese group). Table 2 present self-perceived overweight rate and overweight prevalence from 2005 to 2022 including the COVID-19 pre-pandemic and pandemic eras. From 2005 to 2019 (pre-pandemic) overall self-perceived overweight rate increased ( $\beta$ , 2.80 [95% CI, 2.70 to 2.90]). In 2020-2022 (pandemic), overall self-perceived overweight rate significantly decreased ( $\beta$ , -0.53 [95% CI, -0.74 to -0.33]). During the COVID-19 pandemic, individuals with higher levels of stress showed a greater reduction in the rate of self-perceived overweight as follows: high group ( $\beta$ , -1.20 [95% CI, -1.72 to -0.68]), middle-high group ( $\beta$ , -1.02 [95% CI, -1.36 to -0.68]), middle group ( $\beta$ , -0.47 [95% CI, -0.76 to -0.19]), middle-low group ( $\beta$ , -0.19 [95% CI, -0.62 to 0.24]), low group ( $\beta$ , 0.10 [95% CI, -0.78 to 0.99]). Furthermore, during the COVID-19 pandemic, individuals with lower economic status of households presented a more significant decrease in the rate of self-perceived overweight as follows: high group ( $\beta$ , -0.54 [95% CI, -1.07 to -0.01]), middle-high group ( $\beta$ , -0.58 [95% CI, -0.90 to

-0.26]), middle group ( $\beta$ , -0.23 [95% CI, -0.50 to -0.04]), middle-low group ( $\beta$ , -0.72 [95% CI, -1.28 to -0.15]), low group ( $\beta$ , -1.26 [95% CI, -2.45 to -0.07]).

Table 3 illustrates wORs of self-perceived overweight according to risk factors. When the school performance was low, self-perceived overweight rate increased as follows: middle-high group (wOR, 1.16 [95% CI, 1.14 to 1.18]; reference high group), middle group (wOR, 1.20 [95% CI, 1.18 to 1.21]), middle low group (wOR, 1.47 [95% CI, 1.45 to 1.49]), and low group (wOR, 1.54 [95% CI, 1.51 to 1.57]). Additionally, as stress levels increase, the proportion of self-perceived overweight individuals rises as follows: middle-low group (wOR, 1.10 [95% CI, 1.06 to 1.13]; reference low group), middle group (wOR, 1.29 [95% CI, 1.25 to 1.32]), middle-high group (wOR, 1.58 [95% CI, 1.53 to 1.62]), and high group (wOR, 1.87 [95% CI, 1.81 to 1.92]). As subjective health status worsens, the prevalence of self-perceived overweight increases as follows: healthy group (wOR, 1.32 [95% CI, 1.30 to 1.33]; reference very healthy group), normal group (wOR, 1.63 [95% CI, 1.61 to 1.65]), and unhealthy group (wOR, 1.88 [95% CI, 1.85 to 1.92]). The proportion of self-perceived overweight individuals increases with a decrease in the economic status of households as follows: middle-high group (wOR, 1.07 [95% CI, 1.05 to 1.09]; reference high group), middle group (wOR, 1.16 [95% CI, 1.14 to 1.18]), middle-low group (wOR, 1.45 [95% CI, 1.43 to 0.48]), and low group (wOR, 1.54 [95% CI, 1.50 to 1.58]).

#### IV. CONCLUSIONS

Our study, which analyzed the prevalence of self-perceived overweight among 1,189,586 adolescents who participated in KYRBS from 2005 to 2022, revealed that the prevalence of self-perceived overweight is significantly higher than that of BMI-based overweight. Additionally, during the COVID-19 pandemic the prevalence of self-perceived overweight decreased. We also found that groups with lower school performance, subjective health status, and household economic status, as well as those with higher stress levels, were more likely to have a higher self-perceived overweight ratio. The prevalence of self-perceived overweight decreased more significantly among those with higher stress level or lower economic status of households during COVID-19 pandemic. These findings show which factors affect self-perceived overweight and further provide information on who is vulnerable to self-perceived overweight. Our findings suggest policies to prevent diseases arising from self-perceived overweight, particularly for adolescents with poor risk factors. This study indicates the need for further research on factors influencing self-perceived overweight and the impact of COVID-19 on self-perceived overweight.

#### V. ACKNOWLEDGMENTS

This research was supported by grants from the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health and Welfare, Republic of Korea (grant number:HE23C002800).

**Table 1.** Baseline characteristics of participants in KYRBS, 2005-2022 (total n=1,189,586).

	Total	Pre-pandemic					During the pandemic		
	2005-2022	2005-2007	2008-2010	2011-2013	2014-2016	2017-2019	2020	2021	2022
Overall, n	1,189,586	202,748	221,266	221,102	204,521	178,664	54,809	54,712	51,764
Grade, weighted % (95% CI)									
7 <sup>th</sup> -9 <sup>th</sup> grade (middle school)	50.35 (49.97 to 50.73)	56.22 (55.32 to 57.12)	50.85 (49.93 to 51.77)	49.14 (48.35 to 49.93)	46.98 (46.18 to 47.77)	46.53 (45.69 to 47.37)	49.67 (48.19 to 51.14)	51.04 (49.61 to 52.47)	51.68 (50.13 to 53.22)
10 <sup>th</sup> -12 <sup>th</sup> grade (high school)	49.65 (49.27 to 50.03)	43.78 (42.88 to 44.68)	49.15 (48.23 to 50.07)	50.86 (50.07 to 51.65)	53.02 (52.23 to 53.82)	53.47 (52.63 to 54.31)	50.33 (48.86 to 51.81)	48.96 (47.53 to 50.39)	48.32 (46.78 to 49.87)
Sex, weighted % (95% CI)									
Male	52.25 (51.64 to 52.86)	53.00 (51.44 to 54.55)	52.84 (51.25 to 54.42)	52.44 (51.02 to 53.87)	52.09 (50.64 to 53.54)	51.96 (50.55 to 53.38)	51.84 (49.57 to 54.10)	51.64 (49.46 to 53.82)	51.56 (49.40 to 53.72)
Female	47.75 (47.14 to 48.36)	47.00 (45.45 to 48.56)	47.16 (45.58 to 48.75)	47.56 (46.13 to 48.98)	47.91 (46.46 to 49.36)	48.04 (46.62 to 49.45)	48.16 (45.90 to 50.43)	48.36 (46.18 to 50.54)	48.44 (46.28 to 50.60)
Region of residence, weighted % (95% CI)									
Urban	93.85 (93.63 to 94.07)	92.54 (91.95 to 93.13)	94.29 (93.89 to 94.68)	93.55 (93.02 to 94.08)	93.77 (93.22 to 94.33)	94.05 (93.47 to 94.64)	94.11 (93.32 to 94.90)	94.45 (93.67 to 95.24)	94.42 (93.51 to 95.33)
Rural	6.15 (5.93 to 6.37)	7.46 (6.87 to 8.05)	5.72 (5.32 to 6.11)	6.45 (5.92 to 6.98)	6.23 (5.67 to 6.78)	5.95 (5.36 to 6.53)	5.89 (5.10 to 6.68)	5.55 (4.76 to 6.33)	5.58 (4.67 to 6.49)
BMI group, weighted % (95% CI) <sup>a</sup>									
Underweight	7.75 (7.69 to 7.82)	6.10 (5.95 to 6.25)	9.04 (8.89 to 9.19)	8.13 (7.99 to 8.26)	7.54 (7.41 to 7.68)	6.81 (6.68 to 6.94)	7.56 (7.31 to 7.81)	8.15 (7.89 to 8.40)	8.68 (8.43 to 8.93)
Normal	69.77 (69.62 to 69.93)	54.36 (53.67 to 55.05)	76.62 (76.38 to 76.87)	76.29 (76.07 to 76.52)	74.35 (74.11 to 74.60)	71.44 (71.17 to 71.71)	68.49 (67.99 to 68.99)	66.81 (66.29 to 67.33)	68.15 (67.63 to 68.68)
Overweight	7.94 (7.88 to 8.00)	5.05 (4.93 to 5.18)	6.83 (6.70 to 6.96)	7.39 (7.27 to 7.51)	8.21 (8.08 to 8.33)	9.11 (8.98 to 9.25)	9.95 (9.68 to 10.21)	9.70 (9.44 to 9.96)	8.99 (8.72 to 9.26)
Obese	8.22 (8.14 to 8.30)	3.87 (3.75 to 3.99)	5.14 (5.01 to 5.28)	6.00 (5.88 to 6.12)	7.62 (7.47 to 7.77)	10.38 (10.19 to 10.56)	11.86 (11.49 to 12.22)	13.20 (12.80 to 13.59)	11.83 (11.45 to 12.22)
Unknown	6.31 (6.17 to 6.45)	30.62 (29.79 to 31.45)	2.36 (2.29 to 2.44)	2.19 (2.12 to 2.26)	2.28 (2.21 to 2.35)	2.26 (2.18 to 2.33)	2.14 (2.01 to 2.28)	2.15 (2.01 to 2.29)	2.35 (2.20 to 2.49)
School performance, weighted % (95% CI) <sup>b</sup>									
High	12.37 (12.28 to 12.47)	13.53 (13.31 to 13.75)	11.34 (11.14 to 11.55)	10.81 (10.63 to 10.99)	12.44 (12.24 to 12.64)	13.21 (13.00 to 13.42)	12.19 (11.79 to 12.59)	12.64 (12.29 to 12.98)	13.47 (13.05 to 13.89)
Middle high	25.19 (25.09 to 25.30)	28.88 (28.62 to 29.13)	23.67 (23.45 to 23.88)	23.98 (23.78 to 24.18)	25.05 (24.84 to 25.25)	25.37 (25.14 to 25.59)	24.65 (24.23 to 25.08)	24.48 (24.05 to 24.92)	25.35 (24.92 to 25.78)
Middle	28.89 (28.79 to 29.00)	29.25 (28.98 to 29.51)	27.17 (26.93 to 27.41)	27.33 (27.12 to 27.53)	28.25 (28.03 to 28.47)	29.46 (29.23 to 29.70)	30.16 (29.74 to 30.57)	31.03 (30.61 to 31.44)	30.03 (29.60 to 30.46)
Middle low	23.16 (23.05 to 23.27)	20.44 (20.20 to 20.68)	25.68 (25.43 to 25.93)	25.36 (25.15 to 25.58)	23.65 (23.44 to 23.87)	22.15 (21.92 to 22.38)	23.01 (22.55 to 23.47)	22.01 (21.61 to 22.41)	21.76 (21.31 to 22.21)
Low	10.38 (10.31 to 10.46)	7.91 (7.74 to 8.08)	12.15 (11.97 to 12.32)	12.52 (12.35 to 12.68)	10.61 (10.45 to 10.77)	9.81 (9.65 to 9.97)	10.00 (9.68 to 10.31)	9.84 (9.53 to 10.15)	9.39 (9.08 to 9.70)
Stress level, weighted % (95% CI) <sup>b</sup>									
High	11.40 (11.31 to 11.48)	13.72 (13.51 to 13.93)	12.78 (12.58 to 12.98)	11.59 (11.41 to 11.76)	9.44 (9.29 to 9.60)	11.11 (10.91 to 11.30)	8.25 (7.96 to 8.54)	10.93 (10.60 to 11.25)	12.29 (11.95 to 12.63)
Middle high	29.14 (29.02 to 29.25)	32.48 (32.20 to 32.76)	30.77 (30.49 to 31.04)	30.19 (29.94 to 30.44)	27.11 (26.85 to 27.36)	27.98 (27.70 to 28.27)	25.87 (25.40 to 26.33)	27.82 (27.37 to 28.27)	29.03 (28.60 to 29.46)
Middle	41.85 (41.74 to 41.97)	39.20 (38.92 to 39.48)	40.97 (40.69 to 41.24)	41.61 (41.37 to 41.85)	43.47 (43.24 to 43.71)	41.72 (41.46 to 41.97)	44.49 (44.05 to 44.94)	42.58 (42.12 to 43.03)	41.89 (41.41 to 42.36)
Middle low	14.75 (14.66 to 14.84)	12.60 (12.40 to 12.79)	13.35 (13.15 to 13.55)	14.07 (13.89 to 14.26)	16.46 (16.25 to 16.66)	15.51 (15.29 to 15.73)	17.81 (17.40 to 18.23)	15.48 (15.12 to 15.84)	13.94 (13.60 to 14.28)
Low	2.87 (2.82 to 2.91)	2.00 (1.91 to 2.09)	2.13 (2.04 to 2.22)	2.54 (2.46 to 2.62)	3.52 (3.42 to 3.62)	3.69 (3.58 to 3.80)	3.58 (3.41 to 3.75)	3.19 (3.02 to 3.36)	2.85 (2.69 to 3.01)
Subjective health status, weighted % (95% CI)									
Very healthy	21.75 (21.62 to 21.88)	16.82 (16.57 to 17.07)	17.77 (17.51 to 18.03)	19.99 (19.72 to 20.26)	25.57 (25.27 to 25.87)	27.46 (27.13 to 27.79)	27.17 (26.56 to 27.78)	22.09 (21.60 to 22.58)	20.17 (19.69 to 20.65)

	21.88)	17.06)	18.03)	20.26)	25.87)	27.79)	27.77)	22.58)	20.65)
Healthy	44.95 (44.84 to 45.07)	45.04 (44.75 to 45.33)	46.46 (46.20 to 46.73)	47.83 (47.60 to 48.06)	46.23 (45.99 to 46.48)	43.75 (43.49 to 44.01)	42.47 (42.02 to 42.92)	42.60 (42.14 to 43.06)	42.98 (42.50 to 43.47)
Normal	25.30 (25.19 to 25.41)	28.95 (28.69 to 29.21)	27.73 (27.46 to 28.01)	25.24 (25.00 to 25.48)	22.11 (21.88 to 22.34)	21.95 (21.71 to 22.18)	22.65 (22.22 to 23.08)	26.08 (25.61 to 26.54)	26.30 (25.84 to 26.76)
Unhealthy	7.99 (7.93 to 8.06)	9.19 (9.02 to 9.37)	8.03 (7.88 to 8.18)	6.94 (6.81 to 7.07)	6.08 (5.96 to 6.21)	6.85 (6.71 to 6.99)	7.71 (7.44 to 7.99)	9.23 (8.96 to 9.51)	10.54 (10.24 to 10.84)
Smoking status, weighted % (95% CI)									
Non-smoker	81.55 (81.38 to 81.72)	72.37 (71.93 to 72.81)	73.04 (72.59 to 73.50)	76.07 (75.63 to 76.51)	82.71 (82.28 to 83.13)	86.38 (86.03 to 86.73)	89.84 (89.36 to 90.31)	90.18 (89.73 to 90.63)	91.18 (90.74 to 91.61)
Smoker	18.45 (18.28 to 18.62)	27.63 (27.19 to 28.07)	26.96 (26.50 to 27.42)	23.93 (23.49 to 24.37)	17.29 (16.87 to 17.72)	13.62 (13.27 to 13.97)	10.17 (9.69 to 10.64)	9.82 (9.37 to 10.27)	8.82 (8.39 to 9.26)
Alcohol consumption, weighted % (95% CI)									
0 days/month	82.48 (82.34 to 82.63)	72.53 (72.12 to 72.93)	77.96 (77.60 to 78.32)	81.39 (81.05 to 81.72)	84.01 (83.70 to 84.32)	84.16 (83.87 to 84.46)	89.41 (89.01 to 89.82)	89.33 (88.91 to 89.76)	87.01 (86.52 to 87.50)
1–5 days/month	13.35 (13.24 to 13.46)	20.31 (20.00 to 20.62)	16.13 (15.87 to 16.39)	14.22 (13.98 to 14.47)	12.64 (12.40 to 12.87)	12.41 (12.17 to 12.64)	8.25 (7.93 to 8.57)	8.38 (8.04 to 8.71)	10.35 (9.94 to 10.75)
6–30 days/month	4.17 (4.11 to 4.22)	7.17 (6.98 to 7.35)	5.92 (5.74 to 6.09)	4.39 (4.26 to 4.52)	3.35 (3.23 to 3.47)	3.43 (3.31 to 3.55)	2.34 (2.17 to 2.50)	2.29 (2.12 to 2.46)	2.64 (2.46 to 2.82)
Economic status of households, weighted % (95% CI) <sup>b</sup>									
High	8.93 (8.83 to 9.03)	7.41 (7.22 to 7.60)	6.34 (6.16 to 6.53)	6.83 (6.66 to 7.01)	8.74 (8.54 to 8.94)	10.90 (10.68 to 11.13)	11.20 (10.74 to 11.67)	10.83 (10.44 to 11.21)	11.85 (11.39 to 12.30)
Middle high	27.63 (27.48 to 27.78)	31.04 (30.61 to 31.47)	22.69 (22.36 to 23.02)	24.33 (24.03 to 24.63)	26.86 (26.55 to 27.17)	29.36 (29.04 to 29.67)	28.65 (28.10 to 29.20)	29.32 (28.71 to 29.93)	31.42 (30.85 to 32.00)
Middle	46.75 (46.59 to 46.90)	43.69 (43.31 to 44.07)	47.05 (46.74 to 47.36)	47.11 (46.83 to 47.39)	47.60 (47.29 to 47.92)	46.54 (46.21 to 46.88)	47.57 (46.92 to 48.22)	49.00 (48.35 to 49.65)	46.05 (45.36 to 46.73)
Middle low	13.20 (13.10 to 13.30)	14.03 (13.77 to 14.30)	17.86 (17.60 to 18.13)	16.96 (16.70 to 17.21)	13.56 (13.34 to 13.77)	10.90 (10.70 to 11.10)	10.39 (10.06 to 10.71)	8.96 (8.64 to 9.28)	8.82 (8.50 to 9.14)
Low	3.49 (3.44 to 3.53)	3.83 (3.70 to 3.95)	6.05 (5.90 to 6.20)	4.77 (4.65 to 4.88)	3.24 (3.14 to 3.33)	2.30 (2.22 to 2.38)	2.19 (2.06 to 2.32)	1.89 (1.77 to 2.01)	1.86 (1.74 to 1.99)

Abbreviations: BMI, body mass index (calculated as weight in kilograms divided by height in meters squared); CI, confidence interval; KYRBS, Korea Youth Risk Behavior Web-Based Survey.

<sup>a</sup> BMI was divided into four groups according to the 2017 Korean National Growth Charts: underweight (0-4 percentile), normal (5-84 percentile), overweight (85-94 percentile), and obese (95-100 percentile).

<sup>b</sup> School performance, stress level, and economic status of households were divided into five groups: Low (0-19 percentile), Middle low (20-39 percentile), Middle (40-59 percentile), Middle high (60-79 percentile), and High (80-100 percentile).

**Table 2.** The nationwide trend of self-perceived overweight prevalence before and during the COVID-19 pandemic, weighted % (95% CI), in the KYRBS.

	Pre-pandemic					During the pandemic			Trends in the pre-pandemic era, $\beta$ (95% CI) <sup>a</sup>	Trends in the pandemic era, $\beta$ (95% CI) <sup>a</sup>	Trend differences, $\beta_{diff}$ (95% CI) <sup>a</sup>
	2005-2007	2008-2010	2011-2013	2014-2016	2017-2019	2020	2021	2022			
Overall	25.64 (25.36 to 25.92)	37.74 (37.44 to 38.05)	38.14 (37.85 to 38.44)	38.68 (38.36 to 38.99)	38.72 (38.41 to 39.02)	39.32 (38.79 to 39.85)	38.74 (38.18 to 39.29)	37.08 (36.53 to 37.64)	<b>2.80 (2.70 to 2.90)</b>	<b>-0.53 (-0.74 to -0.33)</b>	<b>-3.33 (-3.56 to -3.10)</b>
Grade											
7 <sup>th</sup> –9 <sup>th</sup> grade (middle school)	24.43 (24.09 to 24.77)	37.21 (36.84 to 37.58)	36.62 (36.27 to 36.96)	35.50 (35.12 to 35.88)	35.87 (35.49 to 36.26)	38.40 (37.70 to 39.11)	37.75 (36.99 to 38.51)	35.71 (34.96 to 36.46)	<b>2.36 (2.23 to 2.49)</b>	-0.12 (-0.39 to 0.16)	<b>-2.48 (-2.78 to -2.18)</b>
10 <sup>th</sup> –12 <sup>th</sup> grade (high school)	27.20 (26.74 to 27.66)	38.30 (37.80 to 38.79)	39.61 (39.14 to 40.09)	41.49 (41.03 to 41.96)	41.19 (40.75 to 41.64)	40.22 (39.44 to 41.01)	39.77 (38.96 to 40.57)	38.55 (37.75 to 39.36)	<b>3.07 (2.93 to 3.22)</b>	<b>-0.84 (-1.13 to -0.54)</b>	<b>-3.91 (-4.24 to -3.58)</b>
Sex											
Male	24.71 (24.38 to 25.04)	34.08 (33.74 to 34.43)	33.29 (32.99 to 33.60)	34.01 (33.66 to 34.35)	35.75 (35.40 to 36.11)	38.72 (38.03 to 39.40)	38.94 (38.24 to 39.65)	37.18 (36.49 to 37.88)	<b>2.23 (2.12 to 2.35)</b>	<b>0.49 (0.24 to 0.74)</b>	<b>-1.74 (-2.02 to -1.47)</b>
Female	26.69 (26.26 to 27.12)	41.84 (41.44 to 42.25)	43.49 (43.13 to 43.85)	43.75 (43.34 to 44.17)	41.93 (41.48 to 42.37)	39.97 (39.21 to 40.73)	38.52 (37.71 to 39.33)	36.98 (36.20 to 37.76)	<b>3.37 (3.23 to 3.52)</b>	<b>-1.63 (-1.92 to -1.34)</b>	<b>-5.00 (-5.32 to -4.68)</b>
Region of residence											
Urban	25.77 (25.47 to 26.07)	37.78 (37.46 to 38.10)	38.14 (37.83 to 38.45)	38.53 (38.20 to 38.86)	38.55 (38.24 to 38.87)	39.18 (38.62 to 39.73)	38.46 (37.88 to 39.03)	36.79 (36.22 to 37.36)	<b>2.70 (2.60 to 2.81)</b>	<b>-0.59 (-0.80 to -0.38)</b>	<b>-3.29 (-3.53 to -3.06)</b>
Rural	24.04 (23.38 to 24.70)	37.12 (36.38 to 37.85)	38.18 (37.27 to 39.09)	40.90 (39.79 to 42.01)	41.32 (40.12 to 42.52)	41.61 (40.13 to 43.09)	43.56 (42.00 to 45.11)	42.09 (40.19 to 43.98)	<b>4.16 (3.85 to 4.47)</b>	0.44 (-0.26 to 1.13)	<b>-3.72 (-4.48 to -2.96)</b>
BMI group <sup>b</sup>											
Underweight	1.16 (0.90 to 1.41)	1.69 (1.46 to 1.92)	1.52 (1.32 to 1.73)	1.88 (1.64 to 2.12)	1.60 (1.35 to 1.86)	1.41 (1.03 to 1.80)	0.93 (0.64 to 1.22)	1.11 (0.78 to 1.45)	<b>0.10 (0.02 to 0.18)</b>	<b>-0.19 (-0.33 to -0.06)</b>	<b>-0.10 (-0.13 to -0.08)</b>
Normal	16.35 (15.97 to 16.72)	31.50 (31.14 to 31.87)	30.68 (30.33 to 31.03)	29.13 (28.77 to 29.48)	26.02 (25.66 to 26.37)	25.31 (24.75 to 25.88)	23.66 (23.05 to 24.26)	23.26 (22.71 to 23.82)	<b>1.40 (1.27 to 1.52)</b>	<b>-0.99 (-1.20 to -0.78)</b>	<b>-0.34 (-0.56 to -0.13)</b>
Overweight	88.39 (87.65 to 89.12)	95.41 (94.96 to 95.86)	94.55 (94.14 to 94.95)	92.68 (92.23 to 93.13)	89.49 (88.96 to 90.02)	86.43 (85.36 to 87.50)	86.22 (85.20 to 87.23)	88.21 (87.15 to 89.27)	<b>-0.33 (-0.52 to -0.15)</b>	<b>-0.44 (-0.83 to -0.06)</b>	<b>-1.76 (-1.99 to -1.54)</b>
Obese	96.95 (96.11 to 97.79)	98.63 (98.37 to 98.88)	98.09 (97.82 to 98.37)	98.09 (97.86 to 98.33)	97.37 (97.10 to 97.64)	96.95 (96.51 to 97.39)	97.10 (96.67 to 97.53)	96.60 (96.06 to 97.14)	-0.06 (-0.21 to 0.09)	<b>-0.21 (-0.41 to -0.02)</b>	<b>-1.34 (-1.49 to -1.20)</b>
School performance <sup>c</sup>											
High	22.54 (21.86 to 23.21)	34.37 (33.60 to 35.15)	33.36 (32.69 to 34.04)	32.34 (31.69 to 33.00)	32.55 (31.87 to 33.22)	33.81 (32.56 to 35.07)	32.74 (31.44 to 34.04)	30.92 (29.67 to 32.18)	<b>1.97 (1.75 to 2.19)</b>	<b>-0.57 (-1.03 to -0.11)</b>	<b>-1.31 (-1.48 to -1.14)</b>
Middle high	24.44 (24.00 to 24.88)	36.74 (36.23 to 37.26)	36.88 (36.41 to 37.36)	37.28 (36.79 to 37.77)	36.68 (36.17 to 37.19)	36.29 (35.31 to 37.28)	36.22 (35.26 to 37.18)	34.60 (33.65 to 35.55)	<b>2.76 (2.60 to 2.92)</b>	<b>-0.63 (-0.97 to -0.28)</b>	<b>-1.62 (-2.10 to -1.15)</b>
Middle	25.29 (24.82 to 25.75)	36.47 (35.98 to 36.96)	36.92 (36.46 to 37.38)	37.28 (36.81 to 37.75)	37.61 (37.13 to 38.09)	38.70 (37.83 to 39.56)	37.71 (36.87 to 38.55)	35.78 (34.92 to 36.65)	<b>2.65 (2.50 to 2.80)</b>	<b>-0.63 (-0.94 to -0.31)</b>	<b>-1.09 (-1.27 to -0.91)</b>
Middle low	28.09 (27.50 to 28.68)	40.47 (39.93 to 41.01)	40.96 (40.45 to 41.47)	42.70 (42.14 to 43.26)	43.25 (42.67 to 43.83)	43.92 (42.93 to 44.91)	43.59 (42.51 to 44.67)	42.65 (41.58 to 43.72)	<b>3.19 (3.00 to 3.38)</b>	-0.20 (-0.59 to 0.19)	<b>-1.28 (-1.42 to -1.14)</b>



	28.68)	41.01)	41.47)	43.26)	43.84)	44.90)	44.68)	43.71)			
Low	30.33 (29.43 to 31.23)	39.92 (39.15 to 40.69)	41.63 (40.92 to 42.35)	44.15 (43.37 to 44.94)	45.39 (44.52 to 46.27)	44.79 (43.16 to 46.43)	45.11 (43.64 to 46.58)	43.88 (42.24 to 45.52)	<b>3.29 (3.01 to 3.56)</b>	-0.42 (- 1.01 to 0.17)	<b>-1.29 (-1.42 to -1.16)</b>
Stress level <sup>c</sup>											
High	32.29 (31.56 to 33.02)	45.03 (44.24 to 45.82)	45.96 (45.23 to 46.69)	47.26 (46.43 to 48.09)	47.15 (46.36 to 47.95)	47.79 (46.05 to 49.53)	45.68 (44.31 to 47.06)	43.73 (42.33 to 45.13)	<b>3.44 (3.20 to 3.69)</b>	-1.20 (- 1.72 to - 0.68)	<b>-2.04 (-2.23 to -1.86)</b>
Middle high	27.54 (27.10 to 27.99)	40.34 (39.84 to 40.84)	41.52 (41.05 to 41.99)	43.32 (42.80 to 43.84)	42.66 (42.16 to 43.16)	43.36 (42.43 to 44.29)	41.84 (40.87 to 42.81)	39.70 (38.78 to 40.61)	<b>3.54 (3.38 to 3.70)</b>	-1.02 (- 1.36 to - 0.68)	<b>-2.75 (-3.13 to -2.38)</b>
Middle	23.23 (22.82 to 23.64)	35.67 (35.27 to 36.07)	36.34 (35.97 to 36.72)	36.89 (36.50 to 37.27)	36.84 (36.42 to 37.25)	38.06 (37.33 to 38.78)	36.88 (36.14 to 37.62)	35.60 (34.82 to 36.39)	<b>2.88 (2.75 to 3.02)</b>	-0.47 (- 0.76 to - 0.19)	<b>-1.57 (-1.78 to -1.37)</b>
Middle low	21.74 (21.09 to 22.39)	32.22 (31.57 to 32.86)	31.58 (31.00 to 32.17)	32.72 (32.15 to 33.29)	32.68 (32.07 to 33.30)	34.42 (33.35 to 35.50)	34.74 (33.55 to 35.92)	31.74 (30.53 to 32.96)	<b>2.20 (2.00 to 2.40)</b>	-0.19 (- 0.62 to 0.24)	<b>-1.49 (-1.62 to -1.36)</b>
Low	20.92 (19.20 to 22.63)	31.03 (29.37 to 32.68)	28.14 (26.83 to 29.44)	29.91 (28.73 to 31.10)	30.07 (28.79 to 31.34)	30.73 (28.51 to 32.96)	32.12 (29.63 to 34.60)	29.70 (27.16 to 32.23)	<b>1.57 (1.09 to 2.04)</b>	0.10 (- 0.78 to 0.99)	<b>-1.24 (-1.35 to -1.13)</b>
Subjective health status											
Very healthy	22.60 (21.99 to 23.21)	32.15 (31.54 to 32.77)	29.84 (29.33 to 30.34)	30.57 (30.10 to 31.04)	30.00 (29.51 to 30.48)	32.33 (31.50 to 33.16)	29.51 (28.62 to 30.41)	27.82 (26.90 to 28.74)	<b>1.19 (1.01 to 1.36)</b>	-0.83 (- 1.16 to - 0.50)	<b>-1.46 (-1.63 to -1.29)</b>
Healthy	24.85 (24.45 to 25.24)	37.07 (36.67 to 37.48)	37.25 (36.88 to 37.61)	38.49 (38.10 to 38.88)	38.58 (38.17 to 38.99)	38.13 (37.38 to 38.88)	37.28 (36.52 to 38.04)	34.84 (34.06 to 35.62)	<b>2.97 (2.84 to 3.10)</b>	-1.19 (- 1.47 to - 0.91)	<b>-1.22 (-1.31 to -1.13)</b>
Normal	27.18 (26.70 to 27.65)	41.01 (40.49 to 41.52)	44.20 (43.72 to 44.68)	45.40 (44.86 to 45.93)	45.91 (45.34 to 46.48)	45.72 (44.71 to 46.73)	44.73 (43.80 to 45.66)	43.30 (42.31 to 44.29)	<b>4.51 (4.34 to 4.68)</b>	-0.88 (- 1.26 to - 0.51)	<b>-0.70 (-0.85 to -0.55)</b>
Unhealthy	30.26 (29.35 to 31.16)	42.70 (41.77 to 43.63)	46.19 (45.28 to 47.10)	49.76 (48.77 to 50.74)	51.55 (50.55 to 52.56)	51.69 (50.04 to 53.33)	50.63 (49.11 to 52.16)	48.45 (46.91 to 49.98)	<b>5.23 (4.93 to 5.53)</b>	-1.09 (- 1.70 to - 0.48)	<b>-1.73 (-1.84 to -1.63)</b>
Smoking status											
Non-smoker	25.64 (25.32 to 25.97)	38.50 (38.16 to 38.84)	38.81 (38.50 to 39.12)	39.01 (38.67 to 39.34)	38.69 (38.37 to 39.02)	39.28 (38.73 to 39.83)	38.74 (38.17 to 39.32)	36.99 (36.41 to 37.57)	<b>2.65 (2.54 to 2.76)</b>	-0.56 (- 0.77 to - 0.34)	<b>-3.21 (-3.32 to -3.10)</b>
Smoker	25.63 (25.12 to 26.14)	35.69 (35.15 to 36.23)	36.02 (35.48 to 36.55)	37.09 (36.48 to 37.70)	38.89 (38.19 to 39.59)	39.67 (38.18 to 41.16)	38.68 (37.16 to 40.20)	38.05 (36.52 to 39.59)	<b>3.11 (2.92 to 3.30)</b>	-0.29 (- 0.81 to 0.23)	<b>-3.11 (-3.66 to -2.56)</b>
Alcohol consumption											
0 days/month	25.81 (25.50 to 26.12)	38.11 (37.78 to 38.44)	38.28 (37.97 to 38.59)	38.39 (38.06 to 38.71)	38.16 (37.84 to 38.49)	38.97 (38.42 to 39.52)	38.48 (37.91 to 39.05)	36.83 (36.25 to 37.40)	<b>2.47 (2.37 to 2.58)</b>	-0.44 (- 0.65 to - 0.23)	<b>-1.43 (-1.74 to -1.12)</b>
1–5 days/month	25.35 (24.77 to 25.94)	37.09 (36.45 to 37.73)	38.19 (37.54 to 38.84)	40.54 (39.80 to 41.29)	42.15 (41.40 to 42.90)	43.01 (41.45 to 44.58)	41.19 (39.58 to 42.80)	39.57 (38.01 to 41.13)	<b>4.07 (3.85 to 4.28)</b>	-0.89 (- 1.43 to - 0.34)	<b>-1.22 (-1.31 to -1.13)</b>
6–30 days/month	24.77 (23.90 to 25.63)	34.68 (33.67 to 35.69)	35.40 (34.28 to 36.51)	38.90 (37.54 to 40.26)	39.95 (38.55 to 41.34)	39.64 (36.62 to 42.66)	39.89 (36.96 to 42.82)	35.79 (32.93 to 38.66)	<b>3.82 (3.46 to 4.17)</b>	-1.17 (- 2.17 to - 0.18)	<b>-1.44 (-1.60 to -1.28)</b>
Economic status of households <sup>c</sup>											
High	23.39 (22.47 to 24.30)	35.23 (34.27 to 36.20)	33.89 (33.03 to 34.76)	33.58 (32.82 to 34.34)	34.20 (33.48 to 34.93)	35.48 (34.19 to 36.78)	35.29 (33.90 to 36.68)	32.48 (30.99 to 33.97)	<b>1.98 (1.72 to 2.25)</b>	-0.54 (- 1.07 to - 0.01)	<b>-1.92 (-2.06 to -1.78)</b>
Middle high	24.08 (23.62)	36.30 (35.76)	36.36 (35.90)	36.46 (35.98)	36.62 (36.14)	37.55 (36.67)	37.27 (36.38)	34.76 (33.89)	<b>2.78 (2.62 to 2.93)</b>	-0.58 (- 0.90 to -	<b>-2.92 (-3.20 to -2.65)</b>

	to 24.53)	to 36.83)	to 36.82)	to 36.95)	to 37.10)	to 38.43)	to 38.15)	to 35.63)		<b>0.26)</b>	
Middle	25.32 (24.93 to 25.71)	36.79 (36.39 to 37.19)	37.24 (36.86 to 37.62)	38.43 (38.04 to 38.83)	38.73 (38.33 to 39.13)	39.29 (38.59 to 39.99)	38.82 (38.10 to 39.54)	38.09 (37.35 to 38.83)	<b>2.86 (2.73 to 2.99)</b>	-0.23 (- 0.50 to 0.04)	<b>-1.26 (-1.51 to -1.02)</b>
Middle low	29.47 (28.79 to 30.16)	41.49 (40.85 to 42.12)	43.10 (42.46 to 43.74)	45.06 (44.32 to 45.79)	46.55 (45.72 to 47.39)	46.11 (44.67 to 47.55)	45.41 (43.82 to 46.99)	44.36 (42.76 to 45.96)	<b>3.87 (3.63 to 4.11)</b>	<b>-0.72 (- 1.28 to - 0.15)</b>	<b>-1.29 (-1.43 to -1.15)</b>
Low	32.28 (30.94 to 33.63)	42.15 (41.04 to 43.26)	44.57 (43.43 to 45.70)	47.65 (46.29 to 49.02)	49.58 (47.88 to 51.28)	50.58 (47.45 to 53.71)	47.51 (44.30 to 50.72)	46.23 (42.83 to 49.63)	<b>4.12 (3.66 to 4.59)</b>	<b>-1.26 (- 2.45 to - 0.07)</b>	<b>-1.06 (-1.17 to -0.96)</b>

Abbreviations: BMI, body mass index (calculated as weight in kilograms divided by height in meters squared); CI, confidence interval; KYRBS, Korea Youth Risk Behavior Web-Based Survey; OR, odds ratio; wOR, weighted odds ratio.

Numbers in bold indicate a significant difference ( $P < 0.05$ ).

<sup>a</sup> All  $\beta$ s and  $\beta$ diffs were expressed by multiplying 100.

<sup>b</sup> BMI was divided into four groups according to the 2017 Korean National Growth Charts: underweight (0-4 percentile), normal (5-84 percentile), overweight (85-94 percentile), and obese (95-100 percentile).

<sup>c</sup> School performance, stress level, and economic status of households were divided into five groups: Low (0-19 percentile), Middle low (20-39 percentile), Middle (40-59 percentile), Middle high (60-79 percentile), and High (80-100 percentile).

**Table 3.** Ratio of ORs for association between self-perceived overweight prevalence of adolescents and each socioeconomic factor, 2005-2022.

		Overall (2005–2022)		Pre-pandemic (2005–2019)		During the pandemic (2020–2022)		Ratio of wORs (95% CI), pre-pandemic (reference) versus pandemic	
		wOR (95% CI)	P-value	wOR (95% CI)	P-value	wOR (95% CI)	P-value	wOR (95% CI)	P-value
Grade	7 <sup>th</sup> –9 <sup>th</sup> grade (middle school) (ref)	1.00 (ref)		1.00 (ref)		1.00 (ref)		1.00 (ref)	
	10 <sup>th</sup> –12 <sup>th</sup> grade (high school)	<b>1.18 (1.17 to 1.20)</b>	<b>&lt;0.001</b>	<b>1.20 (1.18 to 1.21)</b>	<b>&lt;0.001</b>	<b>1.10 (1.07 to 1.13)</b>	<b>&lt;0.001</b>	<b>0.92 (0.89 to 0.94)</b>	<b>&lt;0.001</b>
Sex	Male (ref)	1.00 (ref)		1.00 (ref)		1.00 (ref)		1.00 (ref)	
	Female	<b>1.32 (1.30 to 1.33)</b>	<b>&lt;0.001</b>	<b>1.37 (1.36 to 1.39)</b>	<b>&lt;0.001</b>	1.01 (0.98 to 1.03)	0.489	<b>0.74 (0.72 to 0.76)</b>	<b>&lt;0.001</b>
Region of residence	Urban (ref)	1.00 (ref)		1.00 (ref)		1.00 (ref)		1.00 (ref)	
	Rural	1.01 (0.99 to 1.04)	0.167	0.99 (0.97 to 1.02)	0.527	<b>1.20 (1.15 to 1.25)</b>	<b>&lt;0.001</b>	<b>1.21 (1.15 to 1.27)</b>	<b>&lt;0.001</b>
BMI group <sup>a</sup>	Normal (ref)	1.00 (ref)		1.00 (ref)		1.00 (ref)		1.00 (ref)	
	Underweight	<b>0.04 (0.04 to 0.04)</b>	<b>&lt;0.001</b>	<b>0.04 (0.04 to 0.05)</b>	<b>&lt;0.001</b>	<b>0.04 (0.03 to 0.04)</b>	<b>&lt;0.001</b>	1.00 (0.83 to 1.20)	1.000
	Overweight	<b>28.92 (28.14 to 29.73)</b>	<b>&lt;0.001</b>	<b>31.99 (30.98 to 33.03)</b>	<b>&lt;0.001</b>	<b>20.93 (19.81 to 22.11)</b>	<b>&lt;0.001</b>	<b>0.65 (0.61 to 0.70)</b>	<b>&lt;0.001</b>
	Obese	<b>111.41 (105.05 to 118.16)</b>	<b>&lt;0.001</b>	<b>120.24 (111.64 to 129.49)</b>	<b>&lt;0.001</b>	<b>98.30 (89.76 to 107.65)</b>	<b>&lt;0.001</b>	<b>0.82 (0.73 to 0.92)</b>	<b>0.001</b>
School performance <sup>b</sup>	High (ref)	1.00 (ref)		1.00 (ref)		1.00 (ref)		1.00 (ref)	
	Middle high	<b>1.16 (1.14 to 1.18)</b>	<b>&lt;0.001</b>	<b>1.16 (1.14 to 1.18)</b>	<b>&lt;0.001</b>	<b>1.16 (1.11 to 1.20)</b>	<b>&lt;0.001</b>	1.00 (0.96 to 1.04)	1.000
	Middle	<b>1.20 (1.18 to 1.21)</b>	<b>&lt;0.001</b>	<b>1.19 (1.17 to 1.21)</b>	<b>&lt;0.001</b>	<b>1.24 (1.20 to 1.29)</b>	<b>&lt;0.001</b>	<b>1.04 (1.01 to 1.08)</b>	<b>0.043</b>
	Middle low	<b>1.47 (1.45 to 1.49)</b>	<b>&lt;0.001</b>	<b>1.45 (1.43 to 1.48)</b>	<b>&lt;0.001</b>	<b>1.60 (1.53 to 1.66)</b>	<b>&lt;0.001</b>	<b>1.10 (1.06 to 1.15)</b>	<b>&lt;0.001</b>
	Low	<b>1.54 (1.51 to 1.57)</b>	<b>&lt;0.001</b>	<b>1.53 (1.50 to 1.56)</b>	<b>&lt;0.001</b>	<b>1.67 (1.59 to 1.75)</b>	<b>&lt;0.001</b>	<b>1.09 (1.04 to 1.15)</b>	<b>0.001</b>
Stress level <sup>b</sup>	Low (ref)	1.00 (ref)		1.00 (ref)		1.00 (ref)		1.00 (ref)	
	Middle low	<b>1.10 (1.07 to 1.14)</b>	<b>&lt;0.001</b>	<b>1.10 (1.06 to 1.13)</b>	<b>&lt;0.001</b>	<b>1.14 (1.06 to 1.22)</b>	<b>0.003</b>	1.04 (0.96 to 1.12)	0.365
	Middle	<b>1.29 (1.25 to 1.32)</b>	<b>&lt;0.001</b>	<b>1.29 (1.25 to 1.33)</b>	<b>&lt;0.001</b>	<b>1.31 (1.22 to 1.40)</b>	<b>&lt;0.001</b>	1.02 (0.94 to 1.10)	0.690
	Middle high	<b>1.58 (1.53 to 1.62)</b>	<b>&lt;0.001</b>	<b>1.58 (1.53 to 1.63)</b>	<b>&lt;0.001</b>	<b>1.59 (1.49 to 1.70)</b>	<b>&lt;0.001</b>	1.01 (0.94 to 1.08)	0.866
	High	<b>1.87 (1.81 to 1.92)</b>	<b>&lt;0.001</b>	<b>1.88 (1.82 to 1.94)</b>	<b>&lt;0.001</b>	<b>1.86 (1.73 to 2.00)</b>	<b>&lt;0.001</b>	0.99 (0.91 to 1.07)	0.791
Subjective health status	Very healthy (ref)	1.00 (ref)		1.00 (ref)		1.00 (ref)		1.00 (ref)	
	Healthy	<b>1.32 (1.30 to 1.33)</b>	<b>&lt;0.001</b>	<b>1.31 (1.30 to 1.33)</b>	<b>&lt;0.001</b>	<b>1.35 (1.31 to 1.39)</b>	<b>&lt;0.001</b>	1.03 (0.99 to 1.06)	0.063
	Normal	<b>1.63 (1.61 to 1.65)</b>	<b>&lt;0.001</b>	<b>1.60 (1.58 to 1.63)</b>	<b>&lt;0.001</b>	<b>1.86 (1.80 to 1.92)</b>	<b>&lt;0.001</b>	<b>1.16 (1.12 to 1.20)</b>	<b>&lt;0.001</b>
	Unhealthy	<b>1.88 (1.85 to 1.92)</b>	<b>&lt;0.001</b>	<b>1.81 (1.77 to 1.85)</b>	<b>&lt;0.001</b>	<b>2.33 (2.23 to 2.43)</b>	<b>&lt;0.001</b>	<b>1.29 (1.23 to 1.35)</b>	<b>&lt;0.001</b>
Economic status of households <sup>b</sup>	High (ref)	1.00 (ref)		1.00 (ref)		1.00 (ref)		1.00 (ref)	
	Middle high	<b>1.07 (1.05 to 1.09)</b>	<b>&lt;0.001</b>	<b>1.07 (1.05 to 1.09)</b>	<b>&lt;0.001</b>	<b>1.10 (1.05 to 1.14)</b>	<b>&lt;0.001</b>	1.03 (0.98 to 1.08)	0.230
	Middle	<b>1.16 (1.14 to 1.18)</b>	<b>&lt;0.001</b>	<b>1.16 (1.14 to 1.18)</b>	<b>&lt;0.001</b>	<b>1.21 (1.16 to 1.25)</b>	<b>&lt;0.001</b>	<b>1.04 (1.01 to 1.09)</b>	<b>0.044</b>
	Middle low	<b>1.45 (1.43 to 1.48)</b>	<b>&lt;0.001</b>	<b>1.46 (1.43 to 1.49)</b>	<b>&lt;0.001</b>	<b>1.58 (1.50 to 1.66)</b>	<b>&lt;0.001</b>	<b>1.08 (1.02 to 1.14)</b>	<b>0.005</b>
	Low	<b>1.54 (1.50 to 1.58)</b>	<b>&lt;0.001</b>	<b>1.54 (1.50 to 1.59)</b>	<b>&lt;0.001</b>	<b>1.77 (1.63 to 1.92)</b>	<b>&lt;0.001</b>	<b>1.15 (1.05 to 1.25)</b>	<b>0.002</b>

Abbreviations: BMI, body mass index (calculated as weight in kilograms divided by height in meters squared); CI, confidence interval; OR, odds ratio; wOR, weighted odds ratio.

Numbers in bold indicate a significant difference ( $P < 0.05$ ).

<sup>a</sup> BMI was divided into four groups according to the 2017 Korean National Growth Charts: underweight (0-4 percentile), normal (5-84 percentile), overweight (85-94 percentile), and obese (95-100 percentile).

<sup>b</sup> School performance, stress level, and economic status of households were divided into five groups: Low (0-19 percentile), Middle low (20-39 percentile), Middle (40-59 percentile), Middle high (60-79 percentile), and High (80-100 percentile).

## REFERENCES

1. Rhee SY. Obesity: lessons learned and the way forward. *Life Cycle* 2023; 3: e6.
2. Eum S, Rhee SY. Age, ethnic, and sex disparity in body mass index and waist circumference: a bi-national large-scale study in South Korea and the United States. *Life Cycle* 2023; 3: e4.
3. Longo MR. Distortion of mental body representations. *Trends Cogn Sci* 2022; 26(3): 241-54.
4. Dahlenburg SC, Gleaves DH, Hutchinson AD, Coro DG. Body image disturbance and sexual orientation: An updated systematic review and meta-analysis. *Body Image* 2020; 35: 126-41.
5. Amaral ACS, Ferreira MEC. Body dissatisfaction and associated factors among Brazilian adolescents: A longitudinal study. *Body Image* 2017; 22: 32-8.
6. Yang F, Qi L, Liu S, et al. Body Dissatisfaction and Disordered Eating Behaviors: The Mediation Role of Smartphone Addiction and Depression. *Nutrients* 2022; 14(6).
7. Mishina K, Kronström K, Heinonen E, Sourander A. Body dissatisfaction and dieting among Finnish adolescents: a 20-year population-based time-trend study. *European Child & Adolescent Psychiatry* 2024.
8. Kim MJ, Lee KH, Lee JS, et al. Trends in body mass index changes among Korean adolescents between 2005-2020, including the COVID-19 pandemic period: a national representative survey of one million adolescents. *Eur Rev Med Pharmacol Sci* 2022; 26(11): 4082-91.
9. Kim Y, Choi S, Chun C, Park S, Khang Y-H, Oh K. Data Resource Profile: The Korea Youth Risk Behavior Web-based Survey (KYRBS). *International Journal of Epidemiology* 2016; 45(4): 1076-e.
10. Kim B, Kim HS, Park S, Kwon JA. BMI and perceived weight on suicide attempts in Korean adolescents: findings from the Korea Youth Risk Behavior Survey (KYRBS) 2020 to 2021. *BMC Public Health* 2023; 23(1): 1107.
11. Kim SY, Kim HR, Park B, Choi HG. Comparison of Stress and Suicide-Related Behaviors Among Korean Youths Before and During the COVID-19 Pandemic. *JAMA Netw Open* 2021; 4(12): e2136137.
12. Yang H, Kim MS, Rhee SY, et al. National prevalence and socioeconomic factors associated with the acceptance of COVID-19 vaccines in South Korea: a large-scale representative study in 2021. *Eur Rev Med Pharmacol Sci* 2023; 27(18): 8943-51.
13. Oh J, Kim M, Rhee SY, et al. National Trends in the Prevalence of Screen Time and Its Association With Biopsychosocial Risk Factors Among Korean Adolescents, 2008-2021. *J Adolesc Health* 2023.
14. Woo HG, Park S, Yon H, et al. National Trends in Sadness, Suicidality, and COVID-19 Pandemic-Related Risk Factors Among South Korean Adolescents From 2005 to 2021. *JAMA Network Open* 2023; 6(5): e2314838-e.
15. Shin H, Park S, Yon H, et al. Estimated prevalence and trends in smoking among adolescents in South Korea, 2005-2021: a nationwide serial study. *World J Pediatr* 2023; 19(4): 366-77.
16. Oh J, Lee M, Lee H, et al. Hand and Oral Hygiene Practices of South Korean Adolescents Before and During the COVID-19 Pandemic. *JAMA Netw Open* 2023; 6(12): e2349249.
17. Kang J, Park J, Lee H, et al. National trends in depression and suicide attempts and COVID-19 pandemic-related factors, 1998-2021: A nationwide study in South Korea. *Asian J Psychiatr* 2023; 88: 103727.
18. Park J, Lee M, Lee H, et al. National trends in rheumatoid arthritis and osteoarthritis prevalence in South Korea, 1998-2021. *Sci Rep* 2023; 13(1): 19528.

# Improvement of Mosquito Activity Prediction Performance Using Attention-Assisted Hybrid 1D CNN-LSTM Model

Minjoong Kim<sup>1</sup>, Dayeong So<sup>1</sup>, and Jihoon Moon<sup>1,2,\*</sup>

<sup>1</sup> *Department of ICT Convergence, Soonchunhyang University, Asan, South Korea*

<sup>2</sup> *Department of AI and Big Data, Soonchunhyang University, Asan, South Korea*

\*Contact: jmoon22@sch.ac.kr, phone +82-41-530-4956

**Abstract**—With the recent increase in malaria cases in South Korea, predicting mosquito activity has become essential to prevent disease transmission and implement effective pest control measures. Traditional mosquito activity prediction models have mainly used statistical methods, but these methods struggle to fully capture complex and time-varying patterns. Therefore, this study proposes a new approach using a deep learning-based model that has recently attracted attention. In particular, convolutional long short-term memory (ConvLSTM) is recognized as an effective model for time series prediction problems because it can simultaneously handle spatial and temporal patterns of time series data. However, the ConvLSTM model faces the challenge of treating all-time dependencies equally. The factors influencing mosquito activity may vary in importance over time. To address this issue, this study presents a model that combines the attention mechanism with ConvLSTM to dynamically adjust importance over time. This model was evaluated using mosquito prediction data in Seoul, and aims to develop an effective mosquito prediction model that is responsive to climate and environmental changes. This research is expected to contribute to a better understanding of mosquito occurrence in cities and reduce social and economic losses.

## I. INTRODUCTION

Mosquitoes are important vectors of several diseases, such as malaria and dengue fever, and have a significant impact on public health [1]. Therefore, predicting mosquito activity is critical for preventing disease transmission and implementing effective control measures. However, accurate prediction of mosquito activity is challenging due to its dependence on complex factors.

Existing mosquito prediction models have primarily relied on statistical approaches [5]. However, these methods struggle to effectively capture complex patterns and temporal changes. As a result, models based on machine learning and deep learning have attracted interest. In particular, the combination of one-dimensional convolutional neural network (1D-CNN) and long short-term memory (LSTM) has been recognized as effective for time series prediction problems due to its ability to process both spatial and temporal patterns of time series data [2].

However, a limitation of convolutional LSTM (ConvLSTM) is its uniform treatment of all temporal dependencies, which hinders its ability to account for the varying importance of different factors in predicting mosquito activity. To address this issue, this study introduces a model that uses an attention

mechanism with ConvLSTM to dynamically adjust the importance of factors over time.

This paper evaluates the proposed model using mosquito prediction data from Seoul. Through this evaluation, we aim to develop an effective mosquito prediction model that adapts to climate and environmental changes. This research is expected to improve our understanding of mosquito occurrence in urban areas and minimize social and economic losses.

The paper is organized as follows: Section II reviews related studies, Section III describes the structure and working principles of the model, Section IV discusses the experimental design and results, and Section V concludes the paper with a summary of findings and directions for future research.

## II. RELATED WORK

Methods of predicting mosquito activity included methods of directly predicting mosquito populations using factors such as environmental variables [5]–[9] and methods of indirectly identifying mosquito populations through vector-borne diseases such as dengue fever and malaria [3, 4]. In this paper, we aimed to predict the mosquito activity rate using a method that directly measured the mosquito population. In addition, recent studies have identified the factors used to predict mosquito activity and characterized their characteristics.

As vector-borne diseases emerged as a social problem, research was actively conducted to uncover various information, such as factors of mosquito occurrence and its degree, by analyzing mosquito-related data. Villena et al. [5] estimated the occurrence and abundance of mosquitoes using traditional statistical models such as classification and regression tree (CART), general linear model (GLM), and generalized linear mixed models (GLMM) based on environmental factors, such as temperature and precipitation, distance to anthrax characteristics, and trap type (i.e., Centers for Disease Control and Prevention light trap (CDC light trap) and CDC gravid trap). Specifically, they provided reasons for using different types of traps based on environmental and geographic factors. Lee and Park [6] predicted mosquito occurrence patterns by considering the characteristics of urban areas (i.e., landscape, land use, meteorological factors, and mosquito control activities), which had different mosquito occurrence characteristics from the natural environment. In particular, the random forest (RF) model, which showed the best performance in [7] in predicting mosquito occurrence using

support vector machine (SVM), CART, and RF, was analyzed focusing on the geographical characteristics of waterfront and variable areas. This demonstrated a high ability to model ecological problems involving non-linear relationships between data. In addition, SHAP, or Shapley additive explanations, was used to select important variables for mosquito distribution among climate data. Xia et al. [8] predicted mosquito occurrence using RF, decision tree (DT), multilayer perceptron (MLP), and SVM and provided a thorough and easy way to investigate the correlation between environmental factors (i.e., precipitation, specific humidity, enhanced vegetation index, and surface skin temperature) and mosquito population.

However, there were limitations due to the opportunistic nature of recording mosquito occurrence in areas of interest (AOI), the lack of complete mosquito abundance data when running machine learning models, and the rapid change of environmental factors over the years due to climate change. These limitations suggested that machine learning predictors were more appropriate for analyzing trends from a few years ago. Chen et al. [9] quantified the relationship between mosquito abundance and socioeconomic, landscape, and combined two factors by constructing models such as k-nearest neighbor (kNN), artificial neural network (ANN), and SVM, including socioeconomic factors (i.e., population size at the sampling site, average household income, employment rate, educational status, population density, and average house sale price) and geographic factors. In addition, the proposed model used socioeconomic and landscape factors as inputs, which were characterized by reducing monitoring and data collection costs because they were not as dynamic as other commonly used environmental factors such as temperature and humidity.

In the case of existing studies, to increase predictive power, mosquito capture methods had been supplemented in the data collection process, or variables related to mosquito occurrence had been added or replaced. Each study used machine learning and deep learning models with better performance, but there was a limitation that there were not many attempts to improve the predictive model itself.

### III. FORECASTING MODEL CONSTRUCTION

In this study, we used a Att-ConvLSTM model to predict the mosquito activity index by simultaneously addressing complex temporal patterns and important temporal steps with collected time series data. The proposed model consists of three main components:

- The LSTM is a type of recurrent neural network (RNN) that specializes in processing sequential, or continuous, data. While basic RNNs are adept at learning the temporal characteristics of sequences, they struggle with long-term dependencies. LSTM overcomes this by incorporating a “gate” mechanism in each recurrent unit that allows it to decide which information to retain or discard. This structure allows LSTM to effectively handle data with long-term dependencies [10].
- The 1D-CNN differs from its more common counterpart by applying convolutional operations to one-dimensional data, making it ideal for analyzing time-series data or natural language processing. The core of 1D-CNN includes the convolution layer and the pooling layer, where the former

applies multiple filters to the input data to extract features, and the latter reduces the feature dimensions to simplify the model and prevent overfitting [11].

- The attention mechanism, a technique used in deep learning models for sequence data, allows the model to focus on significant information by assigning an “attention score” to each input, indicating its importance. In this way, the model learns which inputs most influence the output [12].

For the layer configuration, the 1D-CNN extracts high-dimensional features from time-domain data through multiple convolutional layers, as shown in Fig. 1. The weights of the CNN are then updated by backpropagating the error from the loss function. To mitigate the effects of gradient loss across multiple network layers, each convolutional layer includes a connection layer with batch normalization, rectified linear unit (ReLU), and maximum pooling layers [13].

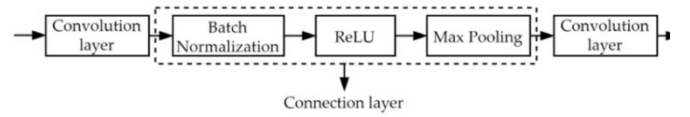


Fig. 1 Layer Configuration for 1D-CNN

The proposed model extracts features from the input data via the 1D-CNN layer, as shown in Fig. 2, and passes them to the LSTM layer for use as input. The LSTM layer learns the temporal relationship between multidimensional vectors and obtains the corresponding hidden state vector. The attention layer then determines the attention weight of each input, and the nonlinear regression layer derives the predicted mosquito activity index.

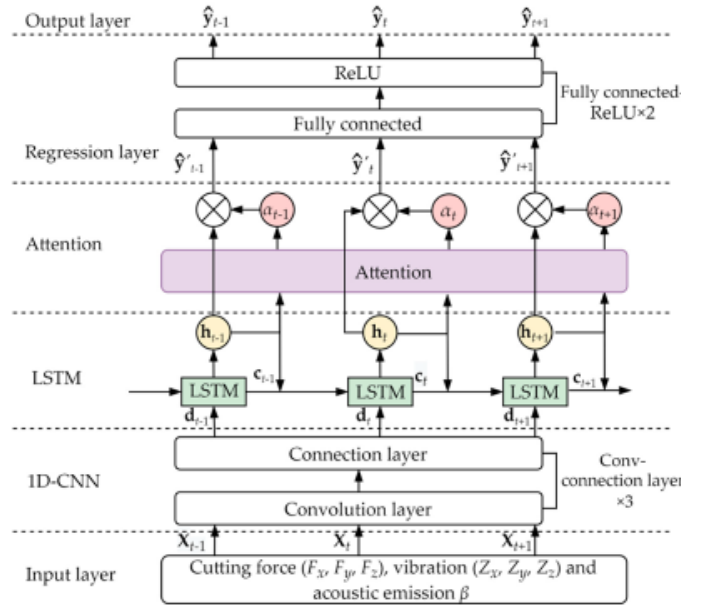


Fig. 2 Architecture of the Att-ConvLSTM Model

### IV. RESULTS AND DISCUSSION

In this paper, approximately 1704 days of data collected from May 1, 2016 to December 31, 2020 provided by the Seoul Metropolitan Mosquito Forecast were used. This dataset includes date, mosquito activity index (averages for Seoul city,

waterfronts, residential areas, and parks), precipitation, average temperature, minimum temperature, maximum temperature, collection amount, and mosquito species. In this study, experiments focused on the average mosquito activity index for Seoul under different labels based on the data collection area. Duplicate data with overlapping dates were removed by preprocessing, with 70% of the total data used for training and 30% for testing.

To evaluate the predictive performance of the model, this study used two indicators, namely the mean absolute error (MAE) and the root mean square error (RMSE), calculated using Equations (1)–(2):

$$MAE = 1/n \times \sum |F_t - A_t|, \quad (1)$$

$$RMSE = (\sqrt{\sum (F_t - A_t)^2 / n}), \quad (2)$$

here  $A_t$  and  $F_t$  represent the actual value and the predicted value at time  $t$ , and  $n$  represents the number of observations. These measurements quantify the difference between the predicted and actual mosquito incidence rates. The MAE measures the mean absolute difference between the predicted and actual values, while the RMSE measures the square root of the mean squared differences between the predicted and actual values.

The total number of learning iterations was set to 150, and the batch size was set to 72. Learning continued until the loss value between the training and validation datasets showed no significant variance beyond a certain threshold.

TABLE 1  
PERFORMANCE COMPARISON

Models	RMSE	MSE	R <sup>2</sup>
Support Vector Machine	0.074	54.271	0.934
Artificial Neural Network	0.066	37.250	0.949
Random Forest	0.064	27.291	0.950
LSTM	0.064	24.569	0.950
ConvLSTM	0.062	24.156	0.954
ConvGRU	0.062	23.905	0.955
Att-ConvLSTM	0.061	23.842	0.955

Table 1 presents the learning results for both the existing and the newly proposed models, showing that the feature extraction method that combines the attention mechanism after the integration of 1D CNN and LSTM layers achieved the highest performance. For the proposed models, it was observed that the average MAE was reduced by 13.3% compared to the Random Forest model, which had the best performance among the existing models.

## V. CONCLUSIONS

In this paper, we proposed a model designed to accurately predict mosquito activity by simultaneously considering complex temporal patterns and significant temporal steps. This model integrated ConvLSTM and attention mechanism techniques, which enabled learning of time series data by simultaneously considering these patterns and steps. The performance evaluation showed that the proposed method could improve the performance by about 13%. These results indicated that the proposed model was capable of adapting to rapidly changing environmental factors due to climate change and the diverse characteristics collected from around the world. It was

also expected that this approach would enable more accurate detection of abnormal patterns and sentiment analysis not only in healthcare, but also in fields such as medicine, speech recognition, and natural language processing.

Future studies could focus on several aspects. First, the learning speed and performance of the model could be improved by using different optimization algorithms. Second, the accuracy of mosquito occurrence prediction could be further improved by incorporating additional meteorological variables. Third, the proposed model could be applied to regions with different climatic conditions to assess its generalizability and scalability. Finally, future research could integrate the proposed model with mosquito abundance and vector-borne disease (VBD) incidence prediction technology to explore comprehensive mosquito prevention strategies.

## ACKNOWLEDGEMENTS

This research was supported by a Korea Institute for Advancement of Technology (KIAT) grant funded by the Korea Government (MOTIE) (P0012724, HRD Program for Industrial Innovation).

## REFERENCES

- [1] M. K. Kindhauser, T. Allen, V. Frank, R. S. Santhana, and C. Dye, "Zika: the origin and spread of a mosquito-borne virus," *Bull. World Health Organ.*, vol. 94, no. 9, pp. 675–686, Sep. 2016.
- [2] P. Li, J. Zhang, and P. Krebs, "Prediction of Flow Based on a CNN-LSTM Combined Deep Learning Approach," *Water*, vol. 14, no. 6, p. 993, Mar. 2022.
- [3] A. N. A. Kamarudin, Z. Zainol, and N. F. A. Kassim, "Forecasting the Dengue Outbreak using Machine Learning Algorithm: A Review," in *Proc. 2021 Int. Conf. Women in Data Science at Taif Univ.*, 2021.
- [4] M. C. Wimberly, J. K. Davis, M. B. Hildreth, and J. L. Clayton, "Integrated Forecasts Based on Public Health Surveillance and Meteorological Data Predict West Nile Virus in a High-Risk Region of North America," *Environ. Health Perspect.*, vol. 130, no. 8, 2022.
- [5] O. C. Villena et al., "Environmental and geographical factors influence the occurrence and abundance of the southern house mosquito, *Culex quinquefasciatus*, in Hawai'i," *Sci. Rep.*, vol. 14, no. 604, 2024.
- [6] D. Lee and Y. Park, "Interpretable machine learning approach to analyze the effects of landscape and meteorological factors on mosquito occurrences in Seoul, South Korea," *Environ. Sci. Pollut. Res.*, vol. 30, pp. 532–546, 2023.
- [7] Y. Kwon et al., "Modeling Occurrence of Urban Mosquitoes Based on Land Use Types and Meteorological Factors in Korea," *Environ. Res. Public Health*, vol. 12, no. 12, pp. 13131–13147, Dec. 2015.
- [8] I. Xia et al., "Using Machine Learning Models for Predicting *Culex* Mosquito Habitats and Breeding Patterns in Washington D.C.," *Research Square*, 2023.
- [9] S. Chen et al., "An operational machine learning approach to predict mosquito abundance based on socioeconomic and landscape patterns," *Landscape Ecol.*, vol. 34, pp. 1295–1311, 2019.
- [10] Y. Yu, X. Si, C. Hu, and J. Zhang, "A Review of Recurrent Neural Networks: LSTM Cells and Network Architectures," *Neural Comput.*, vol. 31, no. 7, pp. 1235–1270, Jul. 2019.
- [11] S. M. Shahid, S. Ko, and S. Kwon, "Performance Comparison of 1D and 2D Convolutional Neural Networks for Real-Time Classification of Time Series Sensor Data," in *Proc. Int. Conf. Information Networking*, 2022.
- [12] X. Ran, Z. Shan, Y. Fang, and C. Lin, "An LSTM-Based Method with Attention Mechanism for Travel Time Prediction," *Sensors*, vol. 19, no. 4, p. 861, Feb. 2019.
- [13] R. Li, X. Ye, F. Yang, and K. Du, "ConvLSTM-Att: An Attention-Based Composite Deep Neural Network for Tool Wear Prediction," *Machines*, vol. 11, no. 2, p. 297, 2023.

# Time Series-Based Multi-Fusion Deep Learning for Day-Ahead Photovoltaic Forecasting

Dayeong So<sup>1</sup>, Minjoong Kim<sup>1</sup>, Yongsung Kim<sup>2</sup>, Jihoon Moon<sup>1,3,\*</sup>

<sup>1</sup> Department of ICT Convergence, Soonchunhyang University, Asan, South Korea

<sup>2</sup> Department of Technology Education, Chungnam National University, Daejeon, South Korea

<sup>3</sup> Department of AI and Big Data, Soonchunhyang University, Asan, South Korea

\*Contact: jmoon22@sch.ac.kr, phone +82-41-530-4956

**Abstract**—Recently, multi-fusion deep learning (DL) approaches, which exploit the advantages of DL, have attracted attention in the field of artificial intelligence. Because these methods have the advantage of improving the prediction accuracy by considering the characteristics and advantages of the combined models, they have been applied to renewable energy research, which requires excellent prediction performance. In this study, we propose a GRU-TCN (short for gated recurrent unit and temporal convolutional network) model to perform multistep-ahead solar photovoltaic (PV) power generation forecasting for the next 24 hours by considering the multi-time series characteristics. To demonstrate that the GRU-TCN model can be used for stable power system operation, we perform a day-ahead solar PV power generation forecasting for Sanyo solar panels collected at Alice Spring, DKASC, Australia. The experimental results show that the GRU-TCN model outperforms the benchmark forecasting models in terms of mean squared error (MSE) and root MSE (RMSE).

## I. INTRODUCTION

Virtual power plants (VPPs) are gaining a lot of attention as a technology to better manage the participation of distributed energy resources in the electricity market and to improve the operation of the power grid. This initiative is in support of the implementation of the Special Law for the Promotion of Distributed Energy in South Korea. A VPP consists of distributed energy resources that work together as a single large power plant by combining multiple small distributed energy resources [1]. Advances in Internet of Things (IoT) and smart grid technologies are expected to enable bi-directional control of distributed energy resources (DERs) to solve technical problems and increase the efficiency of grid operation [2]. These technological advances offer the advantage of increasing the flexibility of the energy system through the integrated management of distributed resources. They can also help manage energy imbalances by enabling VPPs to predict future energy generation, store excess energy in energy storage systems (ESS), and adjust supply schedules to match demand from distributed resources [3].

Typically, electricity markets operate with day-ahead markets, where forecasting energy generation one day in advance is critical to ensuring stable energy supply and demand [4]. However, for renewable energy generation, due to its reliance on clean resources, predicting the exact amount of power generation is challenging. This is because power

generation can fluctuate due to environmental factors and internal conditions. Therefore, developing a reliable solar photovoltaic (PV) power generation forecasting model that can simulate virtual environments within a VPP has become an important research issue.

In the field of artificial intelligence (AI), a multi-fusion deep learning approach that exploits the advantages of deep learning has recently gained attention [5]. This method improves the prediction accuracy by incorporating the characteristics and strengths of different models. As a result, it is applied to solar PV power generation prediction research, which requires high accuracy. In this study, we propose a multi-fusion deep learning model with self-attention, which is applied to a gated recurrent unit (GRU)-temporal convolutional network (TCN) model. This model predicts the solar PV power generation forecasting for the next 24 hours by analyzing multi-time series data.

This paper is organized as follows. Section 2 introduces the related work, and Section 3 details the data preprocessing process. Sections 4 and 5 describe the multi-fusion deep learning model for solar PV power generation forecasting in detail and present experimental results to demonstrate its superiority over benchmark forecasting models, respectively. Finally, Section 6 outlines the conclusions of this paper and suggests future research directions.

## II. RELATED WORK

AI-based forecasting models have been developed to accurately predict solar PV power generation. AlShafeey and Csáki [6] recognized that solar PV power generation is influenced by various meteorological variables such as solar irradiance and temperature. They organized the input variables in three different ways: structure, hybrid, and time-series. The experimental results showed that the artificial neural network (ANN) model outperformed the multiple linear regression (MLR) model in terms of prediction accuracy, regardless of how the input variables were organized. Zhao et al. [7] introduced automated machine learning (AML) to streamline repetitive tasks such as data preprocessing and hyperparameter optimization for day-ahead solar PV power generation forecasting. In addition, they improved the forecasting accuracy by using a genetic algorithm (GA) to select operators well suited to weather factors, thereby optimizing the reconstruction of input variables and minimizing the error of the forecasting model.



Recently, DL, an important method of artificial intelligence, has gained attention for its effectiveness in predicting solar PV power generation. DL is characterized by its ability to learn from multiple variables and complex patterns through its layered structure and nodes, which makes it highly suitable for analyzing solar PV power generation data with diverse trends and patterns. In addition, recent studies have introduced various multi-fusion DL frameworks, which aim to leverage the strengths and characteristics of different DL models to build and train hybrid models. Pengtao et al. [8] developed a wavelet packet transform (WPD)-long short-term memory (LSTM) model, which integrates WPD with LSTM and linear weighting methods, with the goal of predicting solar PV power generation one hour in advance at 5-minute intervals. Their results show that the WPD-LSTM model outperforms traditional single DL models under different seasonal and weather conditions. Similarly, Agga et al. [9] introduced a hybrid convolutional neural network (CNN)-LSTM model tailored for next-day short-term solar PV power generation forecasting. This model was shown to outperform both single machine learning models (i.e., MLR, k-nearest neighbors, and decision tree) and single DL models (i.e., ANN, CNN, and LSTM) in experimental tests.

However, these models do not account for the time-series nature of the training data, which limits their ability to accurately learn and reflect the dynamic aspects of time-series data. To address these challenges, this study introduces a multi-fusion deep learning model that combines the strengths of GRU, TCN, and self-attention mechanisms after data preprocessing. This approach aims to improve the accuracy of multi-level solar PV power generation prediction through a comprehensive multi-fusion deep learning framework.

### III. DATA PREPROCESSING

The solar power data and meteorological observations for this study were obtained from the Desert Knowledge Australia Solar Center (DKASC) in Alice Springs, Australia [10]. The DKASC houses 38 solar panels, of which 17 are Sanyo panels (array rating: 6.3 kW, material: other, array structure: ground mount, installed: 2010, etc.) were used in the experiments. Data were collected hourly from April 1, 2016 to April 30, 2019, and details of the solar panel specifications and meteorological observations are shown in Table 1.

TABLE 1  
DATASET COLLECTED BY DKASC

No.	Columns
1	Timestamp
2	Wind Speed
3	Weather Temperature Celsius
4	Weather Relative Humidity
5	Global Horizontal Radiation
6	Diffuse Horizontal Radiation
7	Wind Direction
8	Weather Daily Rainfall
9	Radiation Global Tilted
10	Radiation Diffuse Tilted
11	Active Energy Delivered Received
12	Current Phase Average
13	Active Power

Because solar irradiance, the primary energy source for solar PV power, is time- and weather-dependent, it is critical to effectively use external environmental data, especially time and weather. However, traditional date and time formats are sequential, making it difficult to accurately reflect periodicity. For example, 23:00 pm and midnight are adjacent in time, but their sequence difference is marked as 23. In this study, to better capture this periodic nature, one-dimensional sequence data are transformed into two-dimensional sequence data by Equations (1)–(6).

$$\text{Month}_x = \sin(\text{Month} \times (2\pi/12)) \quad (1)$$

$$\text{Month}_y = \cos(\text{Month} \times (2\pi/12)) \quad (2)$$

$$\text{Date}_x = \sin(\text{Day} \times (2\pi/\text{DOTM})) \quad (3)$$

$$\text{Date}_y = \cos(\text{Day} \times (2\pi/\text{DOTM})) \quad (4)$$

$$\text{Hour}_x = \sin(\text{Hour} \times (2\pi/24)) \quad (5)$$

$$\text{Hour}_y = \cos(\text{Hour} \times (2\pi/24)) \quad (6)$$

In Equations (3) and (4), DOTM refers to the day of the month, which indicates the total number of days in a month—for example, February, March, and April have 28 or 29, 31, and 30, respectively.

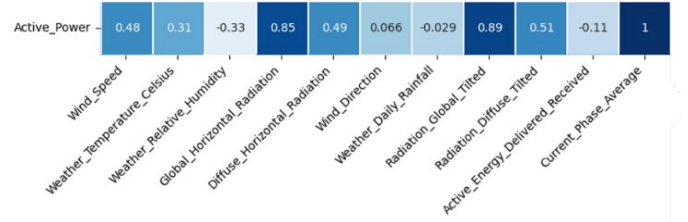


Fig. 1. Correlation Between Input Variables and Active Power (PV)

Fig. 1 is a heat map visualization of the relationship between active power and 11 input variables, excluding time of day, for the Sanyo panel. This visualization shows a strong positive correlation between active power and solar irradiance. In contrast, wind direction, daily precipitation, and active energy supply showed negligible correlation with active power, with values close to zero, and were therefore omitted from the forecasting model input variables.

The dataset contained missing meteorological observations that were attributed to either half or full days due to maintenance or equipment failures. Specifically, there were a total of 22,065 missing observations for the wind speed variable and 971 missing observations for the slope and scattered solar irradiance variables. Due to its significant proportion of missing data, the wind speed variable was excluded. In contrast, missing values for the slope and scattered solar irradiance variables were imputed using the linear interpolation method. In addition, the input variables were subjected to min-max normalization to account for differences in plant capacity, transforming the data to a 0 to 1 scale.

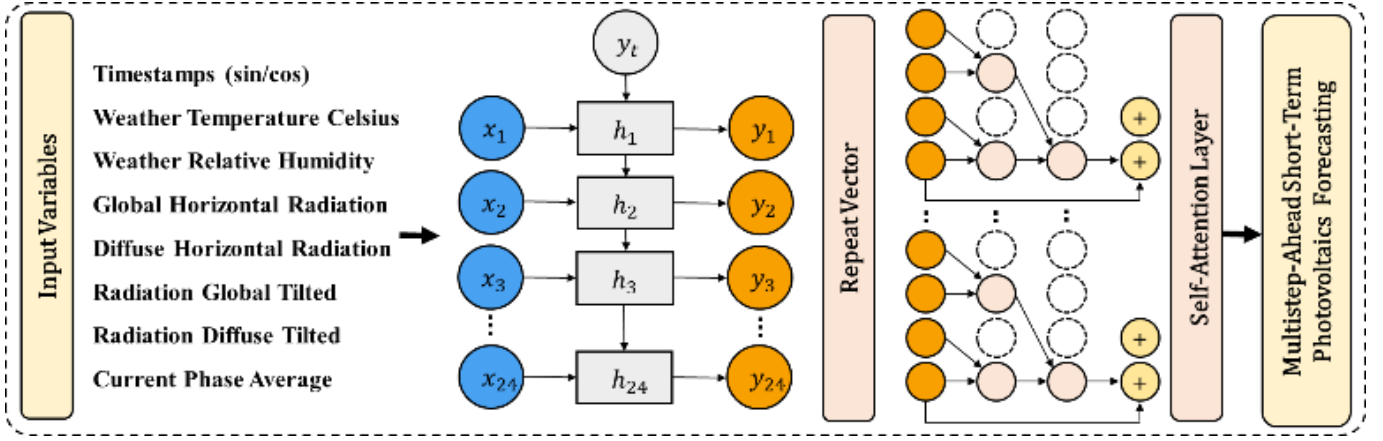


Fig. 2 Architecture of a Time Series-Based Multi-Fusion Deep Learning Model for Multistep-Ahead PV Power Generation Forecasting

#### IV. PROPOSED MODEL

The structure of the multi-fusion DL model proposed in this paper is shown in Figure 2. This model is a sequence-to-sequence fusion that combines a GRU as an encoder and a TCN as a decoder to predict solar PV power generation over the next 24 hours. The GRU, an improvement over the LSTM, includes an update gate that combines the functions of a forget gate and an input gate, allowing it to mix new information with retained memory fragments. This results in a simpler but similarly effective structure compared to the LSTM [12]. The TCN, derived from the CNN, excels at identifying large patterns in sequential data through extended convolution, with the advantage of parallel operation and reduced memory requirements for training [13].

In the model, a "repeat vector" function replicates the output of the GRU encoder—corresponding to the predicted 24-hour period—as input to the TCN decoder, which then produces a 24-hour output. A self-attention mechanism then refines this output to emphasize variables critical to the forecasting. A dense layer is added to ensure that the model output has a single dimension and spans 24 hours. The scaled exponential linear unit (SELU) function is chosen as the activation function to address issues of long term dependence and vanishing gradients [14]. Furthermore, considering the advantages of multistep-ahead over single-step-ahead forecasting for stable grid operation [15], this study adopts a many-to-many approach and trains the model to predict hourly solar PV power generation for the next day.

#### V. RESULTS AND DISCUSSION

The dataset was divided into training and test sets in a 2:1 ratio, with each sample consisting of 24 consecutive hours of input variables and corresponding labels. For both the training and test sets, the sample start time and sequence length were set to 24 hours. The performance of the proposed model was compared with two benchmark forecasting models, LSTM-TCN and bidirectional LSTM (Bi-LSTM)-TCN, whose hyperparameter settings are detailed in Table 2.

TABLE 2  
HYPERPARAMETER CONFIGURATION

No.	Hyperparameter	Setting
1	Epochs	25
2	Batch size	24
3	Optimizer	Adam
4	Metrics	MAE
5	Learning rate	0.001
6	Activation function	SELU
7	Loss	Huber loss
8	Random state	42
9	Early stopping	10

Model validation was performed using mean squared error (MSE) and root MSE (RMSE) metrics, as described by Equations (7) and (8):

$$\text{MSE} = (\sum (F_t - A_t)^2) / n, \quad (7)$$

$$\text{RMSE} = \sqrt{\text{MSE}}, \quad (8)$$

where  $A_t$  is the actual active power (PV) at time  $t$ , and  $F_t$  is the predicted active power (PV) value at time  $t$ .

In this study, a multi-fusion DL model was proposed as a novel approach for multistep-ahead prediction of solar PV power generation over the next 24 hours. For the experimental setup, data preprocessing was applied to account for the characteristics of multi-time-series data, and a multi-fusion DL model was constructed by integrating GRU, TCN, and self-attention mechanisms. Table 3 evaluates the prediction performance for the next 24 hours using the test set from Sanyo solar panels.

The experimental results showed that our model outperformed the benchmark forecasting models, achieving an average RMSE of 1.970 and an average MSE of 3.882. These results confirmed that the proposed multi-fusion DL model could effectively learn time-series patterns and temporal dependencies, focus on relevant information, and ensure high accuracy in predicting solar PV power generation.

TABLE 3  
PERFORMANCE COMPARISON OF MULTI-FUSION MODELS FOR SANYO

Steps	MSE			RMSE		
	LSTM-TCN	Bi-LSTM-TCN	GRU-TCN (Ours)	LSTM-TCN	Bi-LSTM-TCN	GRU-TCN (Ours)
1	1.985	1.979	2.000	3.940	3.915	4.000
2	2.002	2.013	1.979	4.007	4.051	3.915
3	1.997	2.002	1.917	3.988	4.009	3.677
4	2.016	2.002	1.941	4.063	4.010	3.768
5	2.031	1.998	1.943	4.124	3.993	3.776
6	2.038	1.985	1.927	4.152	3.939	3.714
7	2.007	1.967	1.920	4.029	3.867	3.688
8	1.989	1.964	1.919	3.955	3.856	3.683
9	1.970	1.974	1.923	3.883	3.895	3.697
10	1.949	1.977	1.930	3.800	3.910	3.726
11	1.930	1.974	1.923	3.724	3.897	3.698
12	1.931	1.970	1.922	3.728	3.882	3.695
13	1.926	1.968	1.965	3.711	3.873	3.862
14	1.952	1.974	2.007	3.811	3.898	4.027
15	1.968	1.981	2.014	3.872	3.925	4.056
16	2.002	1.967	2.031	4.010	3.870	4.124
17	1.990	1.970	2.029	3.961	3.879	4.119
18	2.013	1.975	2.026	4.051	3.900	4.105
19	2.036	1.979	2.035	4.147	3.917	4.143
20	2.029	1.984	2.001	4.115	3.937	4.004
21	2.004	1.991	1.999	4.016	3.964	3.996
22	1.996	1.995	1.967	3.984	3.981	3.867
23	1.965	1.988	1.975	3.860	3.954	3.901
24	1.970	1.985	1.979	3.880	3.939	3.916
Avg.	1.987	1.982	1.970	3.950	3.928	3.882

## VI. CONCLUSION

In this study, we proposed a multi-fusion deep learning model for predicting solar power generation over the next 24 hours. The experimental results showed that our model outperformed two benchmark forecasting models, LSTM-TCN and Bi-LSTM-TCN, achieving an average RMSE of 1.970 and an average MSE of 3.882. In future work, we aim to conduct further experiments with solar panel data from DKASC to evaluate the model's performance and applicability in different environments, thereby improving the generalizability of the proposed model.

## ACKNOWLEDGEMENTS

This research was supported by the MSIT (Ministry of Science and ICT), Korea, under the ICAN (ICT Challenge and Advanced Network of HRD) program (IITP-2024-2020-0-01832) supervised by the IITP (Institute of Information & Communications Technology Planning & Evaluation).

## REFERENCES

- [1] KEPCO Management Research Institute, "VPP operational status and activation plan," 2021. [Online]. Available: <http://www.keaj.kr/news/articleView.html?idxno=4057>
- [2] KISTI, "ASTI Market Insight Report," 2022. [Online]. Available: <https://repository.kisti.re.kr/bitstream/10580/17882/3/ASTI%20MARKET%20INSIGHT%20028%280712%29.pdf>
- [3] E. Perez et al., "Predictive power control for PV plants with energy storage," *IEEE Transactions on Sustainable Energy*, vol. 4, no. 2, pp. 482–490, 2012.
- [4] F. Wang et al., "A day-ahead PV power forecasting method based on LSTM-RNN model and time correlation modification under partial daily pattern prediction framework," *Energy Conversion and Management*, vol. 212, Article 112766, 2020.
- [5] K. Wang, X. Qi, and H. Liu, "A comparison of day-ahead photovoltaic power forecasting models based on deep learning neural network," *Applied Energy*, vol. 251, Article 113315, 2019.
- [6] M. AlShafeey and C. Csaki, "Evaluating neural network and linear regression photovoltaic power forecasting models based on different input methods," *Energy Reports*, vol. 7, pp. 7601–7614, 2021.
- [7] W. Zhao et al., "A point prediction method based automatic machine learning for day-ahead power output of multi-region photovoltaic plants," *Energy*, vol. 223, Article 120026, 2021.
- [8] P. Li et al., "A hybrid deep learning model for short-term PV power forecasting," *Applied Energy*, vol. 259, Article 114216, 2020.
- [9] A. Agga et al., "CNN-LSTM: An efficient hybrid deep learning architecture for predicting short-term photovoltaic power production," *Electric Power Systems Research*, vol. 208, Article 107908, 2022.
- [10] Desert Knowledge Australia Solar Centre. [Online]. Available: <https://dkasolarcentre.com.au/download?location=alice-springs>
- [11] S. Park et al., "Explainable Photovoltaic Power Forecasting Scheme Using BiLSTM," *KIPS Transactions on Software and Data Engineering*, vol. 11, no. 8, pp. 339–346, 2022.
- [12] A. Mellit, A. Massi Pavan, and V. Lughi, "Deep learning neural networks for short-term photovoltaic power forecasting," *Renewable Energy*, vol. 172, pp. 276–288, 2021.
- [13] D. So et al., "BiGTA-Net: A hybrid deep learning-based electrical energy forecasting model for building energy management systems," *Systems*, vol. 11, no. 9, Article 456, 2023.
- [14] G. Klambauer et al., "Self-normalizing neural networks," *Advances in Neural Information Processing Systems*, vol. 30, 2017.
- [15] J. Liang and W. Tang, "Ultra-short-term spatiotemporal forecasting of renewable resources: An attention temporal convolutional network-based approach," *IEEE Transactions on Smart Grid*, vol. 13, no. 5, pp. 3798–3812, 2022.

# Developing Voice Recognition Models Tailored to Children's Speech Patterns

Hyojin Shin<sup>1\*</sup>, and Jiyoung Woo<sup>2</sup>

<sup>1</sup>Dept. ICT convergence, University of SoonChunHyang, Asan, South Korea

<sup>2</sup>Dept. AI and Bigdata, University of SoonChunHyang, Asan, South Korea

\*Contact: hyojin8296@sch.ac.kr, phone +82-9589-2031

**Abstract**— Despite the increasing use of digital devices by children, research on children's speech recognition remains relatively sparse compared to that on adult speech recognition. This study aims to enhance children's speech recognition capabilities by fine-tuning the Whisper model to more accurately recognize and interpret the unique patterns in children's voices. By utilizing a dataset of children's speech, this study develops a recognition model specifically tailored to child users. Furthermore, this improvement in speech recognition accuracy aims to promote enhanced interactions between children and a range of platforms, including educational, entertainment, and gaming systems, while also providing support to children with developmental disorders.

## I. INTRODUCTION

With the advent of the digital era, the use of digital devices and voice technologies by children has become an essential part of everyday life. However, the majority of existing voice recognition technologies have been primarily developed with adult voices, leading to limitations in accurately capturing the subtle voice patterns and emotional nuances of children's speech. Notably, children's voices have unique characteristics that differ significantly from adults', and technologies that fail to adequately consider these differences can not only limit the user experience but also pose a greater barrier for children with developmental disabilities. This study aims to refine and enhance the performance of the Whisper voice recognition model for children by utilizing a dataset specifically tailored to their voices. Through this effort, we seek to improve the accuracy of voice recognition in challenging environments with significant background noise, such as those with music or other sounds. These technological advances will significantly contribute to the development of customized educational content and applications that specifically consider the needs of children. Furthermore, they will offer substantial support in the early diagnosis and support for children with developmental disabilities, treatment of language development disorders, and the creation of accessible captions. Through this research, we aim to create an equitable digital environment where all children can seamlessly adapt to and integrate with the digital world.

## II. RELATED WORK

Research on adult voice recognition has been extensively conducted across various studies; however, research on

children's voice recognition remains relatively scarce [1], [2]. In the study by Nagano et al. [3] audio data recorded during class discussions by middle school students aged 12 to 15 years old were augmented, as well as data from children aged 11 to 15 years old reading prepared texts. They then applied a Convolutional Neural Network (CNN)-based voice model, achieving Character Error Rate (CER) performances of 49.49% and 22.22%, respectively. In addition, Yu et al. [4] fine-tuned the CTC-CRF model a speech recognition system that combines Connectionist Temporal Classification (CTC) and Conditional Random Field (CRF)—on datasets for adult reading, child reading, and child conversation. This approach yielded CER performances of 9.8% and 9.7%, respectively.

Previous studies have consistently shown lower performance on conversational data as compared to reading scripted text and have primarily involved older children. Furthermore, research on speech recognition for Korean children is scant, especially when contrasted with the work done in English-speaking countries, China, and Japan. In this study, we aim to meticulously fine-tune the Whisper model using speech data exclusively from conversations of Korean children aged 3 to 10 years. Our goal is to develop and enhance a speech recognition model that is specialized for the nuances of Korean children's voices.

## III. METHODOLOGY

### A. Data

The dataset utilized in this study was sourced from AI Hub's "free conversation voice" public dataset [5], which includes a mixed collection of male, female, and child voices. Specifically, the dataset comprises voice data collected from children aged between 3 to 10 years, amounting to 3,000 hours of audio from more than 1,000 distinct speakers. The data encompasses various elements, such as voice recordings in WAV format, transcriptions in JSON format, details about the recorded participants, and the environments where the recordings took place. To facilitate efficient data processing and storage, a subset of 60,000 voice samples has been selected from the comprehensive dataset. This subset was chosen to provide a representative and manageable pool of data for in-depth analysis and for improving speech recognition models specifically designed for the distinctive

speech characteristics of children. Table 1 below details the number of datasets used.

TABLE I  
NUMBER OF DATASET

train	validation	Test
48,000	6,000	6,000

### B. Preprocessing

In this study, both audio and textual data were preprocessed to ensure they were well-suited for model training. Initially, audio data was loaded with a sampling rate of 16kHz and then converted into log-Mel spectrograms. This conversion is essential as it translates voice signal frequencies into a format that the model can easily process, thereby enhancing its ability to learn crucial auditory information effectively. Concurrently, the text data was cleansed to improve the model's learning efficiency by removing non-character symbols and superfluous spaces. We then prepared the text for model training through tokenization and assigned a unique identifier (ID) to each token to preprocess the text data.

### C. CER

The CER was employed to assess the performance of the speech recognition model. CER calculates recognition errors at the character level, making it particularly relevant for languages like Korean, where a single word can consist of multiple morphemes. Utilizing Word Error Rate (WER) in such contexts might lead to an overestimation of the error rate. Therefore, CER, which provides a more accurate reflection of Korean speech recognition performance, was chosen. The formula for calculating CER is presented as Equation 1 below.

$$CER = \frac{S + D + I}{N} \times 100\% \quad (1)$$

In calculating the CER, 'S' stands for 'substitution,' quantifying the characters misidentified by the system and replaced with incorrect ones. 'D' stands for 'deletion,' accounting for characters that the system failed to recognize and consequently omitted. 'I' represents 'insertion,' which marks the number of characters erroneously added by the system, and 'N' stands for 'number of characters,' representing the total count of characters in the reference sequence. The CER is presented as a percentage, simplifying the interpretation of the system's recognition error rate.

### D. Whisper

Whisper model [6] is developed by OpenAI, this large-scale speech recognition model, known as Whisper, demonstrates exceptional accuracy in voice recognition across complex environments, including those with unusual accents or significant background noise. It possesses the advanced capability to convert voice information into text, taking into account the coherence between sentences to

ensure a natural flow of dialogue. The models are categorized by size into tiny, base, small, medium, and large, each defined by the number of parameters they contain. The foundational architecture of the model is illustrated in Figure 1.

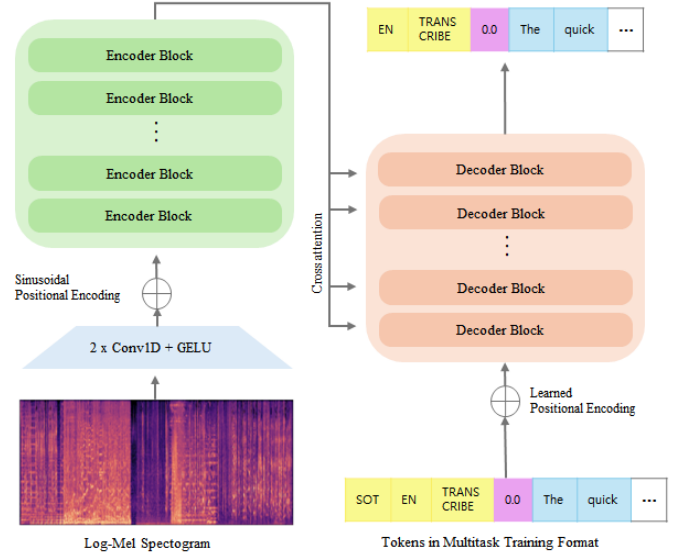


Fig. 1 Whisper architecture

In this study, we built and improved a children's speech recognition model by fine-tuning both the Whisper base, small and medium model and compared their performance. These models have higher accuracy than Whisper tiny model with fewer parameters but are more efficient because they do not require as many computational resources as Whisper large model with many parameters. The hyperparameter information for fine-tuning is shown in Table 2.

TABLE II  
HYPERPARAMETER INFORMATION

Parameters	Value
per device train batch size	32
gradient accumulation steps	1
learning rate	5e-5
warmup steps	50
max steps	6000
gradient checkpointing	True
fp16	True
evaluation strategy	steps
per device eval batch size	16
predict with generate	True
generation max length	50
save steps	1500
eval steps	1500
logging steps	20
metric for best model	cer
greater is better	False

The training and performance of the model depend on key parameters. Gradient accumulation steps allow for larger

batch handling by accumulating gradients across multiple steps. The learning rate adjusts weights and gradually increases through warm-up steps. Max steps limit training duration. Gradient checkpointing and fp16 improve efficiency by reducing memory usage and speeding up processing. Evaluation, saving, and logging are governed by specific strategies. The criteria for the best model are determined by metrics and evaluation settings. Output predictions are controlled by generation limits and techniques. The report to parameter specifies where training logs are sent, optimizing the workflow.

#### IV. RESULTS

In this section, we examine the results of analyzing children's speech recognition data using the Whisper model. The performance of the Whisper model, segmented by size, is presented in Table 3. The model's outputs for the actual dataset are displayed in Table 4 below.

TABLE III  
CER SCORE BY MODEL

Model	CER	Loss
Whisper tiny	63.2139	5.2438
Whisper base	55.0622	3.2975
Whisper small	35.0561	5.4551
Whisper medium	24.4708	5.1488
Whisper base finetuning	16.1683	0.3944
Whisper small finetuning	12.0924	0.3312
Whisper medium finetuning	<b>10.5064</b>	<b>0.288</b>

TABLE IV  
MODEL OUTPUT

Model	Output Text
Reference Text	자동차를 타고 동물원에 갔어요 <b>자동차를 타고 동물원에 갔어요</b> 가족들과 코끼리차 맨 뒤에 앉았어요 <b>코끼리 차에 맨 뒤에 앉았어요</b> 음 그럼 이번에는 짧게 해줄게 가족들과 코끼리 차 <b>가족들과 코끼리 차 맨 뒤에 앉았어요 맨 뒤에 앉았어요</b>
Whisper-medium-finetuning	자동차를 타고 동물원에 갔어요 <b>상자를 타고 동물원에 갔어요</b> 가족들과 코끼리차 메인 뒤에 앉았어요 <b>코끼리차에 메인 뒤에 앉았어요</b> 음 그럼 이번에는 짧게 해줄게 가족들과 코끼리 차 <b>가족들과 코끼리 차 맨 뒤에 앉았어요 맨 뒤에 앉았어요</b>

The Whisper medium finetuning model significantly outperformed other models. It achieved a lowest CER of 10.5064% and a minimum loss of 0.288 This represents an approximately 14% reduction in CER compared to the standard Whisper medium model, underscoring the performance enhancements gained through the fine-tuning process. Therefore, in specialized areas such as children's speech data, careful customization and optimization tailored to the specific characteristics of the data are more important. Additionally, we present the results of some audio data that was collected during a sentence-learning activity between a

teacher and a 4-year-old child at a daycare center, as output by the model. The sections in bold indicate the model's recognition of the child's voice. As a result of comparing the text labels with the manually entered audio data, it was confirmed that the model accurately captured the characteristics of the child's voice and produced generally accurate outputs.

#### V. CONCLUSIONS

This study focused on improving child voice recognition models to more accurately identify their voices as children increasingly use digital devices and voice technology. Predominantly, existing speech recognition models are optimized for adult speech, which poses limitations in accurately capturing and reflecting the unique characteristics of children's speech. To address these issues, we constructed a finetuning Whisper speech recognition model utilizing a dataset of children's voices. The results demonstrated that tailoring the model to specific data characteristics significantly enhances its performance. Enhancements in children's speech recognition models are pivotal in fostering a child-friendly digital environment, potentially aiding in the support of language development disorders through early diagnosis and treatment. Future research should focus on a more comprehensive exploration of children's speech characteristics and the further fine-tuning of other size models to improve performance.

#### ACKNOWLEDGMENT

This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the ICAN(ICT Challenge and Advanced Network of HRD) support program(IITP-2024-2020-0-01832) supervised by the IITP(Institute for Information & Communications Technology Planning & Evaluation)

#### REFERENCES

- [1] Yu, Fan, et al. "The SLT 2021 children speech recognition challenge: Open datasets, rules and baselines." IEEE Spoken Language Technology Workshop (SLT), pp. 1116-1123, 2021.
- [2] Oralbekova, Dina, et al. "Difficulties Developing a Children's Speech Recognition System for Language with Limited Training Data." International Conference on Computational Collective Intelligence. Cham: Springer Nature Switzerland, pp. 419-429, 2023.
- [3] Nagano et al. "Data augmentation based on vowel stretch for improving children's speech recognition." IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), pp. 502-508, 2019.
- [4] Yu et al., "The SLT 2021 Children Speech Recognition Challenge: Open Datasets, Rules and Baselines," IEEE Spoken Language Technology Workshop (SLT), pp. 1117-1123, 2021.
- [5] free conversation voice (mixed male, female, infant, etc.), online available: <https://www.aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSe=realm&dataSetSn=108>
- [6] Radford et al., "Robust speech recognition via large-scale weak supervision," International Conference on Machine Learning. PMLR, pp. 28492-28518, 2022.



# Global, regional, and national trends in substance use disorder mortality rates, across 60 countries, from 1990 to 2021, with projections up to 2040: comprehensive analysis from the WHO Mortality Database

Soeun Kim,<sup>1</sup>Dong Keon Yon<sup>1#</sup>

<sup>1</sup>Center for Digital Health, Medical Science Research Institute, Kyung Hee University College of Medicine, Seoul, South Korea

# Corresponding authors

\*Contact: soeun9342@naver.com, phone +82 10-7688-5487

## Abstract—

**Objective:** Estimate global trends in substance use disorder (SUD) mortality rates from 1990 to 2021 and project SUD deaths until 2040 across 60 countries using a Bayesian age-period-cohort (BAPC) analysis.

**Design:** Analyzed WHO Mortality Database. Calculated age-standardized country-specific SUD mortality rates for 60 countries from 1990 to 2021 using LOESS curves. Estimated future SUD mortality projections up to 2040 with the BAPC model.

**Results:** Of 60 countries studied, 37 were high-income (HICs), and 23 were low to middle-income (LMICs). Global SUD mortality rose from 2.70 deaths per 1,000,000 people (95% CI: 1.28-4.12) in 1990 to 3.69 (95% CI: 2.21-5.17) in 2021. LMICs showed a significant increase from 0.25 (95% CI: -1.12 to 1.63) in 1990 to 5.10 (95% CI: 3.86-6.35) in 2021. Mortality rose notably among ages 45 and above. Positive correlations were found between SUD mortality rates and Human Development Index, Socio-demographic Index, and Gender Gap Index. Predictive models suggest global SUD deaths could rise from 4.15 (95% CI: 4.02-4.31) per 1,000,000 people in 2021 to 7.67 (95% CI: 6.60-8.95) in 2030 and 18.74 (95% CI: 11.98-29.47) in 2040.

**Conclusions:** Global SUD mortality increased from 1990 to 2021, especially in LMICs. Proactive strategies are urgently needed to reduce SUD-related mortality rates.

## I. INTRODUCTION

Substance use disorders (SUD) pose a significant challenge to public health, necessitating immediate attention to their global trends and future projections to develop effective health policies and interventions. [1] The aftermath of the COVID-19 pandemic has seen a surge in the prevalence of SUD, particularly in North America, where an opioid crisis has significantly impacted the United States and Canada. [2] In 2019, the rates of opioid-related mortality were 15.8 and 6.4 per 100,000 individuals in the United States and Canada, respectively, highlighting the severity of this substance use crisis. [3] Additionally, the pandemic period accompanied a reduction in hospital admissions, coinciding with a surge in drug overdose fatalities. This reduction in hospital accessibility during the pandemic may have inadvertently contributed to the increase in mortality rates of SUD; such recent shifts are likely to influence international trends in SUD, suggesting the need for understanding global and longitudinal trends in SUD mortality.

Furthermore, the consumption of substances such as alcohol and psychoactive stimulants is associated with heightened risks of various health complications issues, including fractures, cognitive impairments, cardiovascular diseases, and delirium, each contributing to the overall

morbidity and mortality associated with SUD. [4, 5] Therefore, this study utilized the World Health Organization (WHO) Mortality Database to provide insights into the global trends in SUD mortality rates. It also aimed to estimate the future burden of SUD up to 2040 across 60 countries.

## II. METHOD

In brief, we investigated the international trends in SUD mortality rates by utilizing the WHO Mortality Database for 60 countries between 1990 and 2021. [6] Then, we fitted Bayesian age-period-cohort (BAPC) models to the previous trends to predict SUD mortality rates up to 2040. [7]

### A. DATA SOURCES

Primary data were sourced from the WHO Mortality Database, containing mortality statistics reported by member countries annually. [6] We extracted data on SUD-related deaths using specific ICD codes and population estimates from all available countries, sexes, and age groups. Data collection underwent rigorous verification by two independent researchers using Python software.

### B. STATISTICAL ANALYSIS

We calculated age-standardized SUD mortality rates to adjust for population differences across countries and timeframes. [6] LOESS curves were employed for trend analysis, supplemented by categorization based on income levels and the Human Development Index (HDI) [6]. Correlation analysis was conducted using socioeconomic indicators like HDI, Socio-demographic Index (SDI), Gender Gap Index (GGI), and Gini Coefficient. [8-11] Decomposition analysis examined differences in SUD deaths attributed to population growth, aging, and epidemiological changes. [12] BAPC models were fitted to estimate SUD mortality rates from 2022 to 2040 across various demographics and countries. [13]

### C. PATIENT AND PUBLIC INVOLVEMENT

Given the focus on global epidemiological trends, patient or public involvement was not applicable as the study relied on secondary data from the WHO Mortality Database. No patients were directly involved in setting the research question, data collection, analysis, interpretation, or manuscript writing.

## III. RESULT

Age-standardized SUD mortality rates from 1990 to 2021 were analyzed across 60 countries using the WHO Mortality Database. In 1990, the LOESS smoothed rate was 2.70 deaths per 1,000,000 people (95% CI, 1.28-4.12), increasing to 3.69 deaths per 1,000,000 people (95% CI, 2.21-5.17) in 2021 (Figure 1). Of these countries, 37 were HICs and 23 were LMICs (Figure 1). Notably, LMICs experienced a significant 20.4 times increase in SUD mortality rates, rising from 0.25 deaths per 1,000,000 people (95% CI, -1.12 to 1.63) in 1990 to 5.10 deaths per 1,000,000 people (95% CI, 3.86-6.35) in 2021 (Figure 1). Geographical variations were evident, with Africa and North America showing increasing trends, while Asia Pacific, Europe, and Latin America and the Caribbean appeared to plateau (Figure 2). Significant increases in SUD mortality rates were observed in older age groups, particularly those aged 45 and above (Figure 3). Positive correlations were found between SUD mortality rates and indicators such as the Human Development Index (HDI), Socio-demographic Index (SDI), and Gender Gap Index (GGI) (Figure 4). Analysis of factors contributing to the increase in SUD deaths from 1990 to 2021 revealed impacts from population aging, population growth, and epidemiological changes, particularly pronounced in LMICs (Figure 5). Predictive modeling indicated a significant projected increase in SUD deaths globally, with rates expected to rise to 7.67 deaths per 1,000,000 people by 2030 and 18.74 by 2040 (Figure 6).

#### IV. DISCUSSION

##### A. FINDING OF THIS STUDY

This study observed a global increase in age-standardized SUD mortality rates from 1990 to 2021, with rates rising from 2.70 to 3.69 deaths per 1,000,000 people. Notably, LMICs experienced a substantial 20.4-fold increase in SUD mortality. Geographical variations were evident, with increasing trends in Africa and North America. Older age groups, particularly those aged 45 and above, showed significant increases in SUD mortality rates. Positive correlations were found between SUD mortality rates and indices such as HDI, SDI, and GGI. Factors contributing to the rise in SUD deaths included population aging and epidemiological changes. BAPC models projected a continued significant increase in SUD mortality rates, with rates expected to rise by 2040.

##### B. COMPARISONS WITH PREVIOUS STUDIES

Compared to previous GBD-based studies, this research provides advancements by utilizing raw mortality data from the WHO Mortality Database, enhancing data reliability. [8]

Additionally, our study extends beyond 2019 and covers a broader scope of countries, projecting future trends up to 2040.

##### C. POSSIBLE EXPLANATIONS

The observed increase in SUD mortality rates aligns with rising prevalence of alcohol and drug use disorders, driven by population growth and aging. [14] LMICs face challenges in providing adequate treatment and mental health services, contributing to steeper increases in SUD mortality. [15] Regional variations in North America and Africa may be attributed to factors such as the opioid crisis and historical drug use patterns. [16] Aging populations present another contributing factor, with older individuals becoming more vulnerable to SUD-related mortality due to physiological changes and comorbidities. [17] Positive correlations with socioeconomic indices suggest a complex interplay between socioeconomic development and SUD mortality rates. [18]

##### D. LIMITATIONS AND STRENGTHS

Limitations include potential data incompleteness and variability, as well as underestimation of SUD mortality rates. Projections rely on assumptions about future trends and population estimates. [7] Further prospective studies controlling for confounding factors are needed to accurately estimate SUD mortality risks. [19]

#### V. CONCLUSIONS

In conclusion, there has been a marked and progressive increase in international SUD mortality since 1990, especially in LMICs and the older population. Using models that explicitly adjusted for the effects of age, period, and cohort on trends in SUD mortality, future SUD deaths are predicted to increase up to 2040 at the global levels. These findings suggest urgent and proactive strategies to reduce the mortality rates related to SUD are needed.

#### ACKNOWLEDGMENT

This study was conducted based on the World Health Organization (WHO) Mortality Database and population estimates provided by the United Nations (UN). We retrieved mortality data concerning age, sex, and causes of death from the WHO Mortality Database, which collects information reported by individual countries. Additionally, population estimates for each country were obtained from the datasets provided by the United Nations.



## REFERENCES

- [1] J. H. Cantor, C. M. Whaley, B. D. Stein, and D. Powell, "Analysis of Substance Use Disorder Treatment Admissions in the US by Sex and Race and Ethnicity Before and During the COVID-19 Pandemic," *JAMA Netw Open*, vol. 5, no. 9, p. e2232795, Sep 1 2022, doi: 10.1001/jamanetworkopen.2022.32795.
- [2] T. Gomes, M. Tadrus, M. M. Mamdani, J. M. Paterson, and D. N. Juurlink, "The Burden of Opioid-Related Mortality in the United States," *JAMA Netw Open*, vol. 1, no. 2, p. e180217, Jun 1 2018, doi: 10.1001/jamanetworkopen.2018.0217.
- [3] H.-A. The Lancet Regional, "Opioid crisis: addiction, overprescription, and insufficient primary prevention," *Lancet Reg Health Am*, vol. 23, p. 100557, Jul 2023, doi: 10.1016/j.lana.2023.100557.
- [4] T. Gress, M. Miller, C. Meadows, 3rd, and S. M. Neitch, "Benzodiazepine Overuse in Elders: Defining the Problem and Potential Solutions," *Cureus*, vol. 12, no. 10, p. e11042, Oct 19 2020, doi: 10.7759/cureus.11042.
- [5] M. Tadrus *et al.*, "Assessment of Stimulant Use and Cardiovascular Event Risks Among Older Adults," *JAMA Netw Open*, vol. 4, no. 10, p. e2130795, Oct 1 2021, doi: 10.1001/jamanetworkopen.2021.30795.
- [6] S. Ebmeier, D. Thayabaran, I. Braithwaite, C. Bénamara, M. Weatherall, and R. Beasley, "Trends in international asthma mortality: analysis of data from the WHO Mortality Database from 46 countries (1993-2012)," (in eng), *Lancet*, vol. 390, no. 10098, pp. 935-945, Sep 2 2017, doi: 10.1016/s0140-6736(17)31448-4.
- [7] E. Kiyoshige *et al.*, "Projections of future coronary heart disease and stroke mortality in Japan until 2040: a Bayesian age-period-cohort analysis," (in eng), *Lancet Reg Health West Pac*, vol. 31, p. 100637, Feb 2023, doi: 10.1016/j.lanwpc.2022.100637.
- [8] Y. H. Shin *et al.*, "Global, regional, and national burden of allergic disorders and their risk factors in 204 countries and territories, from 1990 to 2019: A systematic analysis for the Global Burden of Disease Study 2019," (in eng), *Allergy*, vol. 78, no. 8, pp. 2232-2254, Aug 2023, doi: 10.1111/all.15807.
- [9] J. W. Hahn *et al.*, "Global Incidence and Prevalence of Eosinophilic Esophagitis, 1976-2022: A Systematic Review and Meta-analysis," (in eng), *Clin Gastroenterol Hepatol*, vol. 21, no. 13, pp. 3270-3284.e77, Dec 2023, doi: 10.1016/j.cgh.2023.06.005.
- [10] L. The, "2020: a critical year for women, gender equity, and health," (in eng), *Lancet*, vol. 395, no. 10217, p. 1, Jan 4 2020, doi: 10.1016/s0140-6736(19)33170-8.
- [11] B. Clarsen *et al.*, "Changes in life expectancy and disease burden in Norway, 1990-2019: an analysis of the Global Burden of Disease Study 2019," (in eng), *Lancet Public Health*, vol. 7, no. 7, pp. e593-e605, Jul 2022, doi: 10.1016/s2468-2667(22)00092-5.
- [12] X. Cheng *et al.*, "Population ageing and mortality during 1990-2017: A global decomposition analysis," (in eng), *PLoS Med*, vol. 17, no. 6, p. e1003138, Jun 2020, doi: 10.1371/journal.pmed.1003138.
- [13] V. J. Schmid and L. Held, "Bayesian Age-Period-Cohort Modeling and Prediction - BAMP," *Journal of Statistical Software*, vol. 21, no. 8, pp. 1 - 15, 10/16 2007, doi: 10.18637/jss.v021.i08.
- [14] L. Degenhardt *et al.*, "The global burden of disease attributable to alcohol and drug use in 195 countries and territories, 1990-2016: a systematic analysis for the Global Burden of Disease Study 2016," *The Lancet Psychiatry*, vol. 5, no. 12, pp. 987-1012, 2018/12/01/ 2018, doi: [https://doi.org/10.1016/S2215-0366\(18\)30337-7](https://doi.org/10.1016/S2215-0366(18)30337-7).
- [15] V. Patel *et al.*, "Addressing the burden of mental, neurological, and substance use disorders: key messages from Disease Control Priorities, 3rd edition," (in eng), *Lancet*, vol. 387, no. 10028, pp. 1672-85, Apr 16 2016, doi: 10.1016/s0140-6736(15)00390-6.
- [16] O. J. Onaolapo, A. T. Olofinnade, F. O. Ojo, O. Adeleye, J. Falade, and A. Y. Onaolapo, "Substance use and substance use disorders in Africa: An epidemiological approach to the review of existing literature," (in eng), *World J Psychiatry*, vol. 12, no. 10, pp. 1268-1286, Oct 19 2022, doi: 10.5498/wjp.v12.i10.1268.
- [17] K. Humphreys and C. L. Shover, "Twenty-Year Trends in Drug Overdose Fatalities Among Older Adults in the US," (in eng), *JAMA Psychiatry*, vol. 80, no. 5, pp. 518-520, May 1 2023, doi: 10.1001/jamapsychiatry.2022.5159.
- [18] J. Rehm and K. D. Shield, "Global Burden of Disease and the Impact of Mental and Addictive Disorders," (in eng), *Curr Psychiatry Rep*, vol. 21, no. 2, p. 10, Feb 7 2019, doi: 10.1007/s11920-019-0997-0.
- [19] "The global burden of disease attributable to alcohol and drug use in 195 countries and territories, 1990-2016: a systematic analysis for the Global Burden of Disease Study 2016," (in eng), *Lancet Psychiatry*, vol. 5, no. 12, pp. 987-1012, Dec 2018, doi: 10.1016/s2215-0366(18)30337-7.

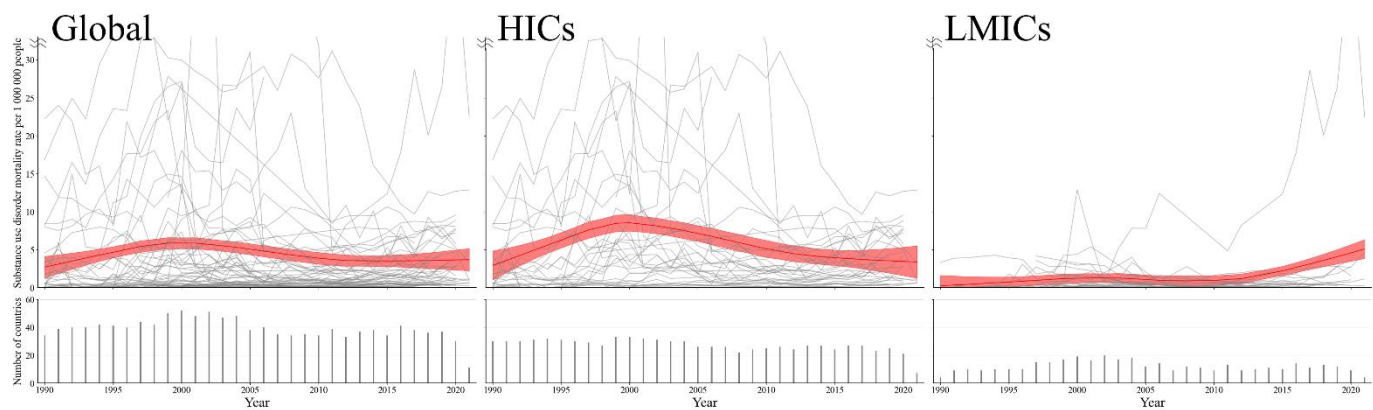


Fig 1. Age-standardized SUD mortality rates for the global, HICs, and LMICs population among 60 countries for the years 1990–2021.

The LOESS mortality rates with 95 % confidence levels, weighted by country population, are shown in red. HICs included 37 countries, including Australia, Austria, Belgium, Canada, Chile, Croatia, Czech Republic, Denmark, Finland, France, Germany, Greece, Hong Kong, Hungary, Iceland, Ireland, Israel, Italy, Japan, Lithuania, Luxembourg, Netherlands, New Zealand, Norway, Panama, Poland, Portugal, Puerto Rico, Romania, Singapore, Slovenia, Spain, Sweden, Switzerland, United Kingdom, United States of America, and Uruguay. LMICs included 23 countries, including Albania, Argentina, Bosnia and Herzegovina, Brazil, Bulgaria, Colombia, Costa Rica, Ecuador, Egypt, Kazakhstan, Kiribati, Malaysia, Mauritius, Mexico, Morocco, North Macedonia, Peru, Philippines, Republic of Moldova, Serbia, South Africa, Thailand, and Venezuela.

Abbreviations: HICs, high-income country; LMICs, low- and middle-income country; LOESS, locally weighted scatterplot smoother; SUD, substance use disorder.

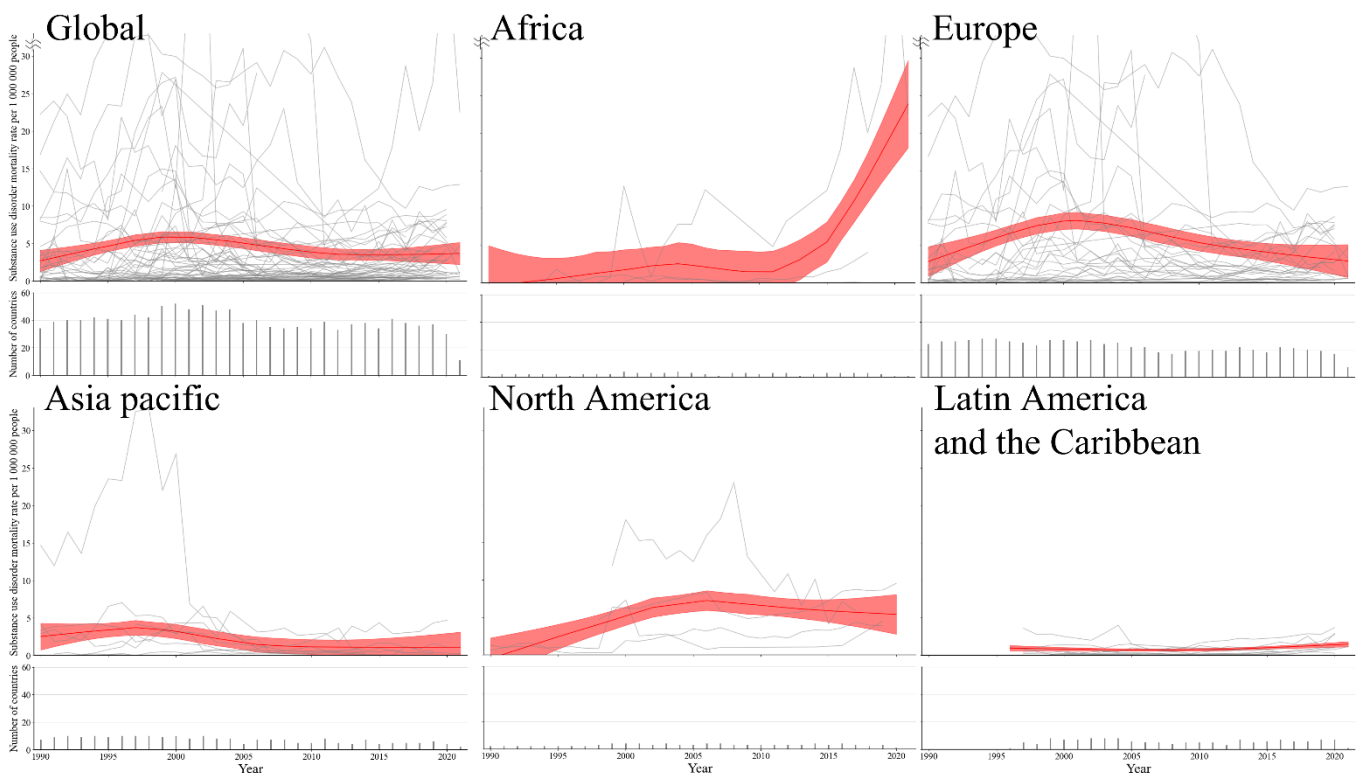


Fig 2. Age-standardized SUD mortality rates across the globe and five continents among 60 countries.

The LOESS mortality rates with 95 % confidence levels, weighted by country population, are shown in red. Africa includes the 4 countries, including Egypt, Mauritius, Morocco, and South Africa. Europe includes the 31 countries, including Albania, Austria, Belgium, Bosnia and Herzegovina, Bulgaria, Croatia, Czech Republic, Denmark, Finland, France, Germany, Greece, Hungary, Iceland, Ireland, Italy, Lithuania, Luxembourg, Netherlands, North Macedonia, Norway, Poland, Portugal, Republic of Moldova, Romania, Serbia, Slovenia, Spain, Sweden, Switzerland, and United Kingdom. Asia Pacific includes the 11 countries, including Australia, Hong Kong SAR, Israel, Japan, Kazakhstan, Kiribati, Malaysia, New Zealand, Philippines, Singapore, and Thailand. North America includes the 4 countries, including Canada, Panama, Puerto Rico, and United States of America. Latin America and the Caribbean include the 10 countries, including Argentina, Brazil, Chile, Colombia, Costa Rica, Ecuador, Mexico, Peru, Uruguay, and Venezuela.

Abbreviations: LOESS, locally weighted scatterplot smoother; SUD, substance use disorder.

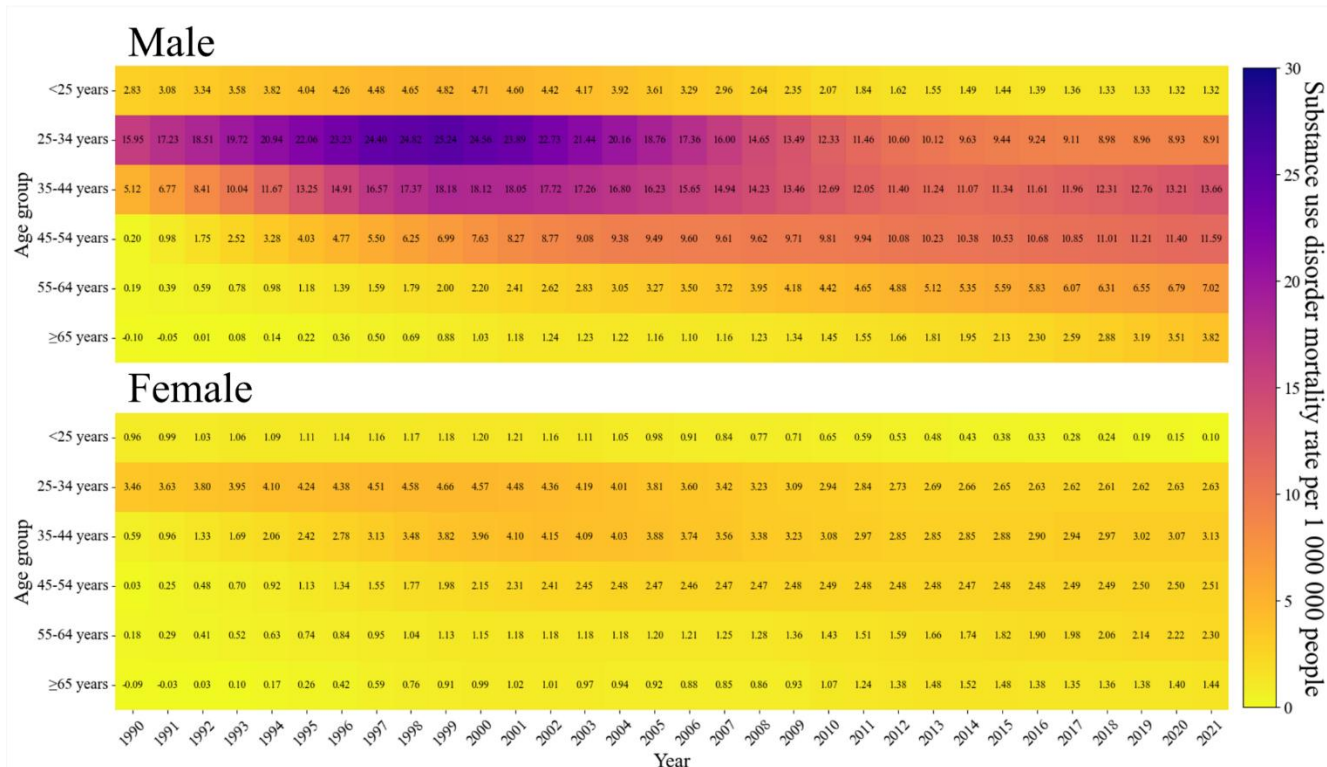


Fig 3. LOESS smoothed SUD mortality rates by sex and age group among 60 countries, 1990–2021. Abbreviations: LOESS, locally weighted scatterplot smoother; SUD, substance use disorder.

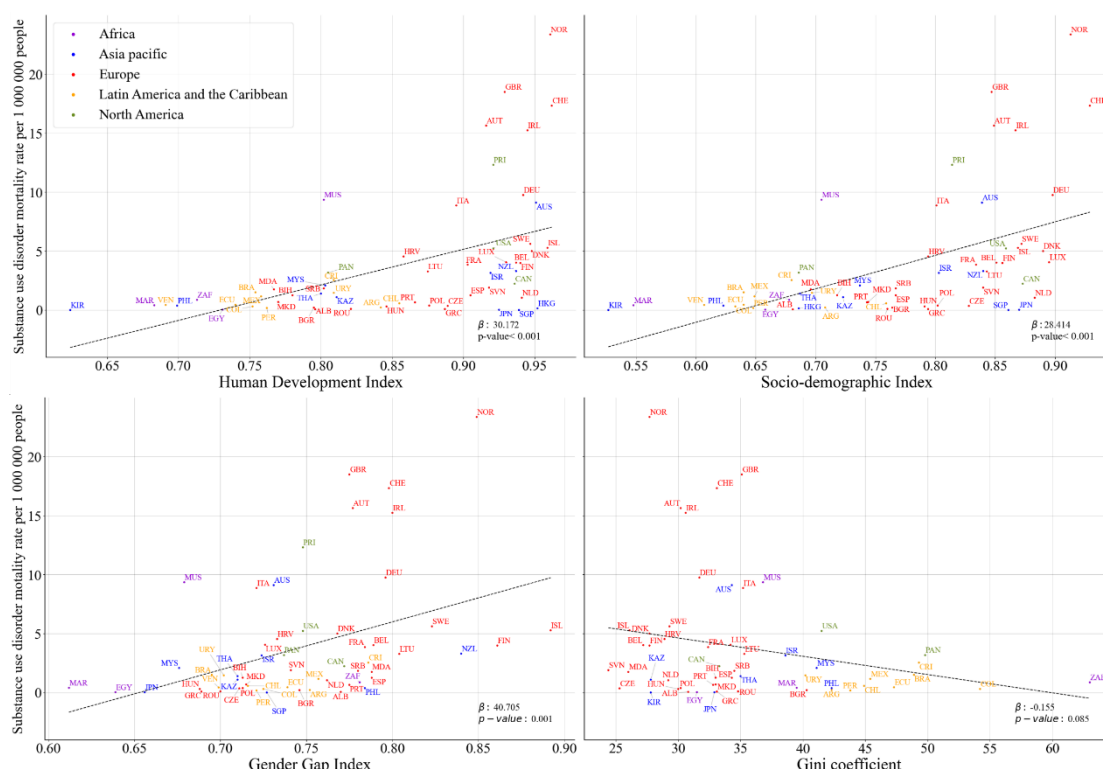


Fig 4. Correlation between age-standardized SUD mortality rates and Human Development Index, Socio-demographic Index, Gender Gap Index, and Gini coefficient. Abbreviations: SUD, substance use disorder.

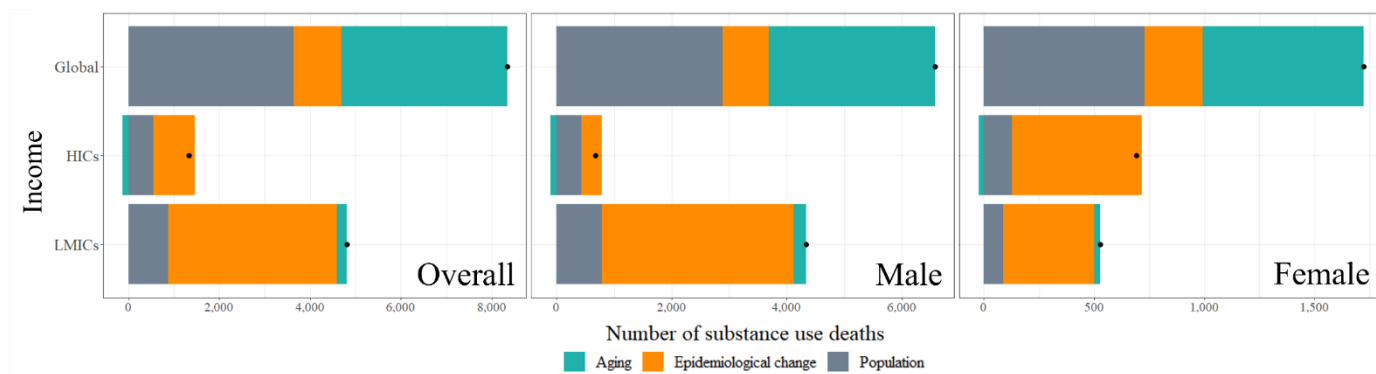


Fig 5. Changes in the number of SUD deaths associated with aging, epidemiological change, and population from 1990 to 2021 by sex. Dots represents the integrated outcome of three factors: aging, epidemiological change, and population. Abbreviations: HICs, high-income country; LMICs, low- and middle-income country; SUD, substance use disorder.

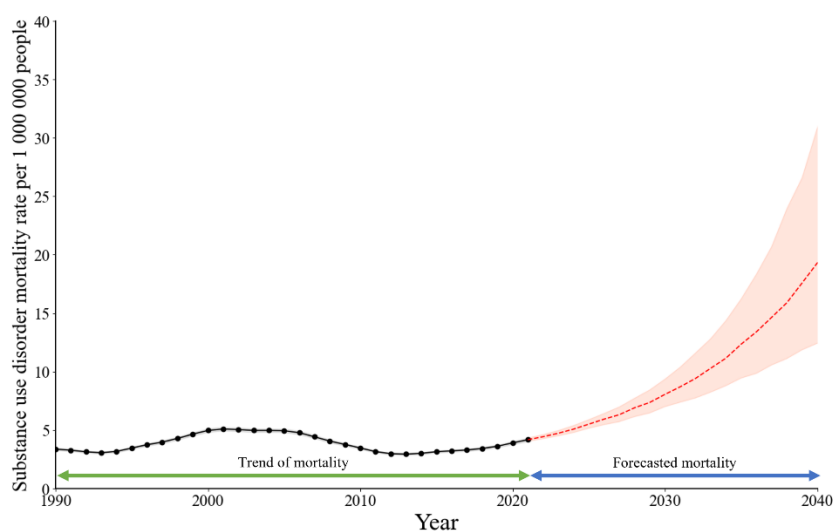


Fig 6. Projections in age-standardized substance use mortality rates from 1990 to 2040 by Bayesian age-period-cohort models. The dashed line represents the Bayesian age-period-cohort value for forecasted mortality, while the shaded area signifies the 95% credible intervals.

# Advanced Behavioral Recognition Models to Improve Early Diagnostic Capabilities for Developmental Disabilities in Children

Insu Jeon<sup>1</sup>, Byeonghun Kim<sup>2</sup>, Chomyong Kim<sup>3</sup>, Jung-Yeon Kim<sup>3</sup>,  
Jiyoung Woo<sup>3,4</sup>, Seungmin Rho<sup>5</sup>, and Jihoon Moon<sup>1,3,4,\*</sup>

<sup>1</sup> Department of Medical Science, Soonchunhyang University, Asan, South Korea

<sup>2</sup> Department of Future Convergence Technology, Soonchunhyang University, Asan, South Korea

<sup>3</sup> Department of ICT Convergence, Shunchunhyang University, Asan, South Korea

<sup>4</sup> Department of AI and Big Data, Soonchunhyang University, Asan, South Korea

<sup>5</sup> Department of Industrial Security, Chung-Ang University, Seoul, South Korea

\*Contact: jmoon22@sch.ac.kr, phone +82-41-530-4956

**Abstract**—Developmental delays, which include a wide range of developmental milestones not achieved within the expected age range, are influenced by genetic, environmental, and socioeconomic factors such as congenital anomalies, infections, and family conditions. The critical importance of early childhood in laying the foundation for future development underscores the need for timely identification and intervention for at-risk infants and toddlers. This paper proposes the use of computational models to refine the early diagnosis of developmental disorders. By examining the behavioral and interactional patterns of children, both typical and developmentally delayed, we aim to establish a robust framework for identifying developmental disorders and delays. Using state-of-the-art convolutional neural network (CNN)-based models, including spatiotemporal graph convolutional network (ST-GCN), two-stream convolutional networks, and R(2+1)D models, our approach focuses on the detection of distinct behavioral patterns. These include self-injurious, stereotypic, and aggressive behaviors categorized under the behavior problems inventory (BPI) framework. Through detailed behavioral analysis and advanced detection technologies, we aim to contribute to early diagnosis and intervention strategies for developmental disabilities in children, potentially mitigating long-term developmental challenges.

## I. INTRODUCTION

In recent years, the integration of computer vision technology into healthcare and developmental monitoring has shown the potential to revolutionize early detection and intervention for several conditions. Identifying developmental delays in children as early as possible and providing early intervention has the greatest impact on positive outcomes [1]. A developmental delay is defined as a 25% delay in two or more of the following areas: gross/fine motor, speech/language, cognitive, social/personality, and activities of daily living, compared to normal expectations for their age. Developmental delays can have a significant negative impact on a child's quality of life. The ultimate goal of treating developmental delays is to maximize a child's potential and minimize secondary complications, enabling them to live as independently as possible, thereby improving their quality of life.

Traditionally, identifying these developmental delays has relied heavily on parental guesswork and professional

assessments based on everyday life experiences. However, these methods can be somewhat subjective and can lead to late or missed diagnoses of developmental delays. Advances in computer vision technology have made it possible to analyze complex human behavior through image analysis. The emergence of sophisticated computer vision techniques, such as spatiotemporal graph convolutional networks (ST-GCNs) [2], two-stream convolutional networks [3], and R(2+1)D models [4], is a promising alternative to relying on humans to assess developmental delays.

By leveraging closed-circuit television (CCTV) footage, a ubiquitous resource in many public and private settings, computer vision techniques such as object tracking, behavior recognition, and interaction understanding can potentially provide an objective, real-time, and non-invasive way to monitor a child's development. This approach could help fill gaps in current screening practices by identifying children early who may benefit from additional diagnostic or early intervention services.

In this paper, we discuss the steps that need to be taken to apply computer vision techniques with this potential to the early screening of children for developmental delays. In particular, we focus on how existing object tracking, motion detection, and interaction recognition techniques can be modified and applied to analyze children's movements and interactions in environments captured by CCTV. The ultimate goal of this research is to use practical, scalable, and non-invasive tools to detect early developmental delays to support early intervention and improve long-term outcomes for children.

## II. BACKGROUND

Computer vision in healthcare covered a wide range of applications, from analyzing medical images (e.g., MRI, CT scans, etc.) to monitoring patient movement in clinical settings, and was not a new concept [5]. Recent advances have enabled more sophisticated applications, such as emotion recognition, gait analysis, and prediction of patient falls [6]. However, the use of computer vision to monitor children's developmental stages remained relatively uncharted territory. Existing studies



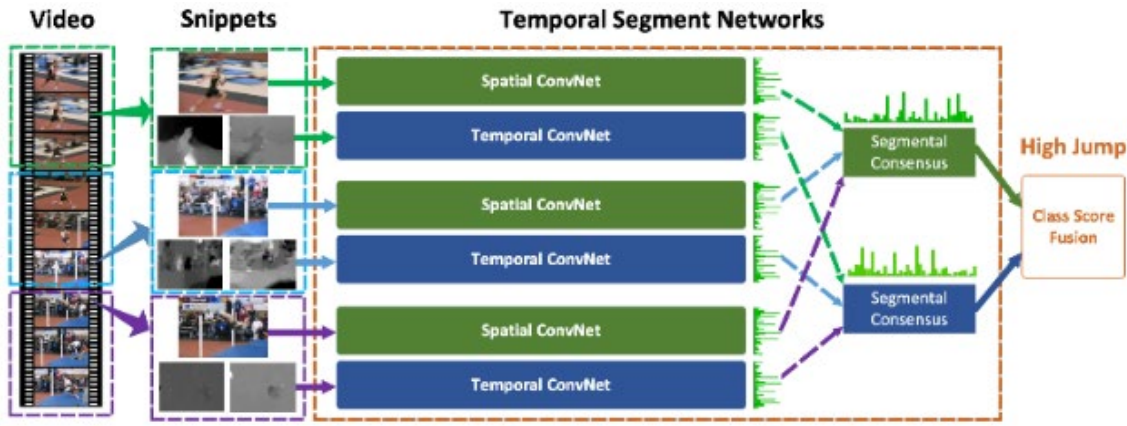


Fig. 1 Temporal segment network (TSN) architecture

have tended to focus on controlled environments and specific tasks, failing to capture the complexity of behavior and environments.

The detection of developmental delays has traditionally relied on standardized developmental screening tools and casual observation [1]. While effective, these methods required significant resources, expertise, and commitment from parents or caregivers. While there has been ongoing research into the use of technologies such as wearable sensors and specialized software applications in this area, these approaches often require active participation or specialized equipment, limiting their scalability and accessibility.

Object tracking and behavior recognition technologies have made significant advances due to deep learning and the availability of large annotated datasets. However, their application to children in the context of developmental monitoring was sparse. Children's movements and interactions were highly variable, context-dependent, and significantly different from those of adults, posing challenges to existing algorithms. In addition, privacy and ethical considerations were paramount in sensitive settings such as homes and schools, especially for vulnerable populations.

Understanding social interactions through computer vision has provided a rich dataset for assessing social and emotional developmental stages. However, current research has primarily focused on adult interactions in contexts such as surveillance and social behavior analysis. Adapting these techniques for children has required recognizing the nuanced dynamics between children and between children and adults that are critical to social and emotional development.

In summary, while the potential of computer vision techniques for early detection of developmental delays in children was significant, numerous challenges remained in applying these techniques to early screening. This thesis aims to address these challenges through an experimental investigation of the applicability, barriers, and potential solutions for using computer vision in the early screening of child developmental delays via CCTV video analysis.

### III. COMPUTER VISION FOR DELAY DETECTION

The experimental goal of this study was to explore the applicability of computer vision techniques for early detection of developmental delays by analyzing children's movements and behaviors based on CCTV footage. For this purpose, we

used a combination of YOLOv8 [7] and DeepSORT [8], the latest technologies in object tracking and behavior recognition, and we used the temporal segment network (TSN) [9] model implemented in the MMAAction2 framework for behavior recognition. The model was pre-trained on the Kinetics-400 dataset, which contains 400 behavior classes and consists of at least 400 video clips for each behavior, providing the data needed to recognize a wide range of human behaviors. The architecture of the TSN is shown in Fig. 1.

#### A. Object Tracking

The results of combining YOLOv8 and DeepSORT for object tracking are shown in Fig. 2 and 3. YOLOv8 was used for object identification, and DeepSORT was used for object tracking. The combination of these technologies provided high accuracy and processing speed for object detection and tracking. This allowed us to accurately track the location and movement of the children in the CCTV footage.

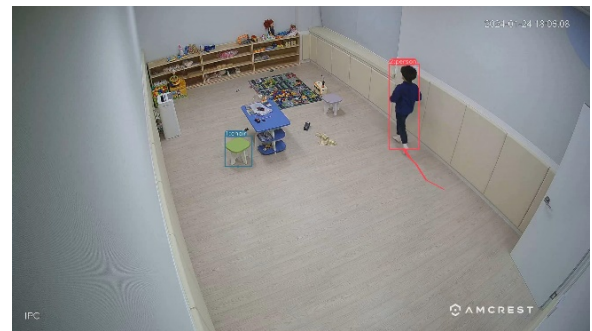


Fig. 2 Detecting and tracking result (children only)



Fig. 3 Detecting and Tracking result (children &amp; adult)

### B. Action and Interaction Recognition

The TSN model within the MMAAction2 framework was used to recognize children's actions and interactions. TSN divided the video into multiple temporal segments and integrated features from each segment to analyze the behavior of the entire video. The results of behavior and interaction recognition are shown in Fig. 4 and 5. When applied to child behavior and interaction videos, TSN showed promise in recognizing child behavior and interaction with additional performance improvements.



Fig. 4 Action and interaction result (pushing car)



Fig. 5 Action and Interaction result (hugging)

Below is a simplified representation of Table I, which shows the accuracy of the TSN model in recognizing specific child behaviors and interactions from the video footage. Each row corresponds to a number (representing a scenario) and lists the behaviors identified by the TSN model along with their recognition accuracy (in parentheses).

TABLE I  
ANALYSIS OF CHILD BEHAVIOR RECOGNITION ACCURACY

Scenario	Behavior 1	Behavior 2	Behavior 3
Pushing car	Pushing car (0.9289)	Crawling baby (0.0562)	Pushing cart (0.0174)
Throwing ball	Tai chi (0.1932)	Throwing ball (0.1350)	Side kick (0.1024)
Hugging	Hugging (0.2612)	Carrying baby (0.0696)	Throwing ball (0.0569)

This table demonstrates the ability of the TSN model to recognize and distinguish between different child behaviors and interactions. The scenarios were referred to by the actions they predominantly represented, and the numbers indicated how accurately each behavior was recognized by the model. For example, in Fig. 4, the model recognized the behavior “pushing

a car” with a high accuracy of 92.89%, while other behaviors such as “crawling baby” and “pushing a cart” were recognized with much lower accuracy. This variation in recognition accuracy underscores the challenges and successes of using computer vision to detect developmental delays by analyzing children’s behaviors.

### IV. CONCLUSIONS

In this study, we explored the possibility of early detection of developmental delays in children by applying computer vision technology to CCTV images. By combining YOLOv8 and DeepSORT, we verified that fast and accurate object tracking was possible, and TSN models within the MMAAction2 framework were used for behavior recognition. The experimental results showed that the object recognition and tracking technology performed well in identifying and tracking children in CCTV footage. This proves that computer vision technology can accurately identify the location and movement of children.

However, it did not perform as expected in behavior and interaction recognition due to the child's unique movement patterns, different forms of interaction, and racial disparity from the pre-training data. To address these issues, additional datasets appropriate for countries that provide early screening services for developmental disabilities need to be built, and algorithms that can accurately analyze children's unique patterns and different interaction modalities need to be developed. Nevertheless, the promise shown in our experiments highlights the importance of further research and technology development in this area.

### ACKNOWLEDGEMENTS

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (No. 2022R1F1A1063134 and No. RS-2023-00218176).

### REFERENCES

- [1] S. Lee, B. Choi, and J. Choi, “Analysis of developmental disabilities news paper articles using text mining: Focused on early screening and early intervention,” *J. Spec. Child. Educ.*, vol. 22, no. 1, pp. 1–27, 2020.
- [2] M.-F. Tsai and C.-H. Chen, “Spatial temporal variation graph convolutional networks (STV-GCN) for skeleton-based emotional action recognition,” *IEEE Access*, vol. 9, pp. 13870–13877, 2021.
- [3] K. Simonyan and A. Zisserman, “Two-stream convolutional networks for action recognition in videos,” *arXiv preprint arXiv:1406.2199*, 2014.
- [4] D. Tran et al., “A closer look at spatiotemporal convolutions for action recognition,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 6450–6459, 2018.
- [5] J. Olveres et al., “What is new in computer vision and artificial intelligence in medical image analysis applications,” *Quant. Imaging Med. Surg.*, vol. 11, no. 8, pp. 3830–3853, 2021.
- [6] N. B. Joshi and S. L. Nalbalwar, “A fall detection and alert system for an elderly using computer vision and Internet of Things,” in *Proc. 2nd IEEE Int. Conf. Recent Trends Electron. Inform. Commun. Technol. (RTEICT)*, pp. 1276–1281, 2017.
- [7] D. Reis et al., “Real-Time Flying Object Detection with YOLOv8,” *arXiv preprint arXiv:2305.09972*, 2023.
- [8] M. I. H. Azhar et al., “People Tracking System Using DeepSORT,” in *Proc. 10th IEEE Int. Conf. Control Syst., Comput. Eng. (ICCSCE)*, pp. 137–141, 2020.
- [9] L. Wang et al., “Temporal Segment Networks for Action Recognition in Videos,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 11, pp. 2740–2755, 2019.

# Towards Gaze Based Method to Quantify Children Interaction with Virtual Reality Contents

Jung-Yeon Kim<sup>1,\*</sup>, Yunyoung Nam<sup>2</sup>

<sup>1</sup>ICT Convergence Research Center, Soonchunhyang University, Asan, Republic of Korea

<sup>2</sup>Department of Computer Engineering and Science, Soonchunhyang University, Asan, Republic of Korea

\*Contact: betterwithme@sch.ac.kr

**Abstract**— Developmental disorders and delays are recognized as a growing concern in South Korea. The exact prevalence rates may vary depending on the specific disorder, but studies indicate that a significant number of children are affected. Developmental disorders and delays are relatively common terms that manifest in 1-3% of children under the age of 5. They do not signify a specific condition or disorder but rather denote slower progress in development compared to their peers of the same age. Early detection and intervention for developmental delays and disorders in children under 3 years old are urgent matters as the developmental pace of infants and toddlers is significantly faster compared to other stages, making the period below 3 years of age the time of highest potential growth in learning and behavioral acquisition. However, while it is currently possible to monitor the growth and developmental trajectory of infants and young children through health check-ups, which include measurements such as brain size, weight, and height, there is a lack of adequate criteria for appropriately evaluating developmental delays and disorders. In addition, recent studies have sought to discover digital biomarkers and utilize them to evaluate certain health conditions including mental health. In this study, our goal is to examine past research on gathering insights from usage data collected from different digital devices and to explore how this data relates to health conditions, focusing specifically on gaze analysis.

## I. INTRODUCTION

In the context where the issue of declining birth rates emerges as a primary concern among the populace, there is evidence indicating a notable increase in the incidence of preterm births relative to the general birth rate. Based on data released by the National Statistics Office, the prevalence of high-risk neonates (preterm infants born before 37 weeks of gestation) has escalated by approximately 1.5-fold over the past decade, rising from 5.8% in 2010 to 9.2% in 2021.

There are several problems related to assessing children for developmental issues. In fact, developmental delay and communication disorders diagnoses and treatments are delayed due to limited accessibility, as hospitals and behavioral development enhancement centers are concentrated in metropolitan area. Parents experience delays in developmental delay diagnoses due to prolonged waiting times averaging more

than three months for child developmental stage assessments, compounded by limited support from national child psychology centers. Additionally, the financial burden posed by private behavioral development enhancement centers, child psychological counseling and therapy centers becomes prohibitive for caregivers.

The developmental pace of infants and young children is notably rapid compared to other periods, rendering the period under three years of age as the most opportune for identifying and intervening in potential developmental delays and disabilities, given its highest potential for learning and behavioral acquisition. Early intervention is deemed necessary due to the diminishing efficacy of interventions as patterns of independent emotional and cognitive processes solidify and behaviors become entrenched beyond the period of infancy and early childhood. Comprehensively addressing developmental delays and disabilities at the national level necessitates the collection of diverse assessment outcomes and the identification of patterns within extensive datasets. Early identification of infants and young children at risk of developmental delays and disabilities within high-risk groups, along with continuous monitoring, is essential. In this study, we focus on gaze movements on assessing developmental delay in children.

## II. GAZE ANALYSIS METHODOLOGIES

Eye tracking has been an active research topic over the past decades to investigate cognitive process and visual attention. Generally, gaze analysis is achieved by computing and evaluating eye movement metrics. In this section, related studies are reviewed in terms of how to identify fixations and saccades, types of eye metrics, and gaze analysis methodologies depending on the types of data, such as images and videos. Finally, studies on how eye tracking can be utilized in medical practice will be revisited.

### A. Fixation Identification Algorithms

Eye movements are mainly composed of saccades and fixations. While the former refers to the fixed eye movement over informative regions of interest, the latter refers to the rapid eye movements between the fixations to move the eye-gaze



from one point to another. Gaze analysis requires to identify fixations and saccades to compute eye movement metrics. There are three different types of identification algorithms based on velocity, dispersion, and area. Velocity-based fixation identification algorithms prioritize the velocity data within eye-tracking protocols, capitalizing on the observation that fixation points exhibit low velocities while saccade points display high velocities. On the other hand, dispersion-based algorithms focus on the spread distance of fixation points, assuming that these points typically cluster together in close proximity whereas area-based algorithms detect points within specified area of interest (AOIs) that depict relevant visual targets. Five representative identification algorithms are summarized in Table 1.

TABLE I  
TYPES OF FIXATIONS IDENTIFICATION ALGORITHMS

Criteria	Representative Algorithms
Velocity	Velocity-Based Identification (I-VT)
Velocity	Hidden Markov Model Fixation Identification (I-HMM)
Area	Area of Interest Identification (I-AOI)
Dispersion	Dispersion-Threshold Identification (I-DT)
Dispersion	Minimum Spanning Trees Based Identification (I-MST)

The five representative identification algorithms were evaluated in terms of accuracy, speed, robustness, ease of implementation, and number of parameters in [1]. Although velocity-based identification algorithms are the simplest method to understand and implement, the result indicates that dispersion-based algorithms are robust and relatively accurate in comparison to velocity and dispersion-based fixation identification algorithms.

#### B. Eye Tracking Metrics Typically Used for Static Images

Over the past decades, eye gaze analysis has focused on images where respondents were presented with the static images and capture the eye movements over the images. Analysis of gaze patterns and eye movements requires to identify saccade and fixation points, and compute eye movement metrics. Common eye movement metrics include fixation durations, saccadic velocities, saccadic amplitudes, and various transition-based parameters between fixations and/or regions of interest. Definition of each metric is given as below.

- Time to first fixation: The amount of time that it takes a respondent to look at a specific AOI from stimulus onset.
- Dwell time: The amount of time that respondents have spent looking at a particular AOI.
- Ratio: The information about how many of respondents actually guided their gaze towards a specific AOI.
- Fixation sequences: The information about when and where a participant looked, which is based on both spatial and temporal information.
- Revisits: The information about how many times a participant returned his/her gaze to a particular spot, defined by an AOI.
- First fixation duration: The information about how long the first fixation lasted for.
- Average fixation duration: The information about how long the average fixation lasted for, which can be determined for either individuals or for groups.
- Heatmaps: The visual information about the general distribution of gaze points, which are typically shown as a color gradient overlay on the presented image or stimulus.

#### C. Gaze Analysis on Videos

Recently, eye tracking technology has been gaining more popularity in analyzing videos. The characteristics of videos are different from static images, and people are different in prioritizing their attention to objects or areas within the presented image. Owing to this, different types of eye tracking metrics are necessary, and researchers proposed to analyze eye metrics that allow comparison between multiple respondents while presenting videos, such as scarf plots that visualize gaze transitions among areas of interest (AOIs) on timelines, a tree visualization to compare duration, frequency, and orderings of fixations on a timeline [2].

#### D. Eye Gaze Analysis in Medical Practice

Eye tracking technology has been tested to evaluate its efficacy in the field of healthcare. Eye tracking can be used to help increase the diagnostic accuracy. Numerous investigations have employed eye tracking methodologies to scrutinize participants' gaze patterns during the execution of visual tasks, facilitating medical practitioners in the identification of nuanced aberrations in visual function through comparative analyses between experimental cohorts and control counterparts. In neurology, the technology has been utilized to help diagnosis precision for attention deficit hyperactivity disorder (ADHD) and findings suggest that the technology is able to provide additional information that may be associated with the ability to maintain focus during visual tasks [3]. In addition, the technology has proven that gaze patterns can be used to analyze eye contact during social interaction for autism spectrum disorder (ASD) [4].

### III. SUMMARY AND CONCLUSIONS

Although eye tracking has been gaining popularity in healthcare domain as the technology can provide valuable information by quantifying eye movements associated with AOIs, the information should be interpreted with caution. This is due to the fact that eye movement metrics are significantly affected by fixation identification algorithms, which determines fixation and saccade points to compute the metrics based on. Evidence suggests that dispersion-based algorithms are relatively robust as compared to velocity- and area-based identification algorithms. Moreover, even though the eye movement metrics can be utilized for both static images and dynamic videos, the metrics must be computed from the continuous images and be expressed on a timeline.

#### ACKNOWLEDGMENT

This research was supported by Korea Institute for Advancement of Technology (KIAT) grant funded by the Korea Government (MOTIE) (P0012724, The Competency Development Program for Industry Specialist).

## REFERENCES

- [1] D. D. Salvucci and J. H. Goldberg, "Identifying fixations and saccades in eye-tracking protocols," in *Proceedings of the 2000 symposium on Eye tracking research & applications*, 2000, pp. 71-78.
- [2] K. Kurzhals, F. Heimerl, and D. Weiskopf, "ISeeCube: visual analysis of gaze data for video," presented at the Proceedings of the Symposium on Eye Tracking Research and Applications, Safety Harbor, Florida, 2014. [Online]. Available: <https://doi.org/10.1145/2578153.2578158>.
- [3] A. Lev, Y. Braw, T. Elbaum, M. Wagner, and Y. Rassovsky, "Eye Tracking During a Continuous Performance Test: Utility for Assessing ADHD Patients," *Journal of Attention Disorders*, vol. 26, no. 2, pp. 245-255, 2022, doi: 10.1177/1087054720972786.
- [4] E. A. Papagiannopoulou, K. M. Chitty, D. F. Hermens, I. B. Hickie, and J. Lagopoulos, "A systematic review and meta-analysis of eye-tracking studies in children with autism spectrum disorders," *Social Neuroscience*, vol. 9, no. 6, pp. 610-632, 2014/11/02 2014, doi: 10.1080/17470919.2014.934966.

# Enhanced Fall Detection: Integrating Near-Fall Events for Safety Surveillance

<sup>1</sup>Nab Mat, <sup>2</sup>Jung-Yeon Kim, <sup>3</sup>Chomyong Kim, and <sup>4</sup>Yunyoung Nam

<sup>1</sup> Department of ICT Convergence, Soonchunhyang University, Asan, South Korea

<sup>2</sup> ICT Convergence Rehabilitation Engineering Research Center, Soonchunhyang University, Asan, South Korea

<sup>3</sup> ICT Convergence Research Centre, Soonchunhyang University, Asan, South Korea

<sup>4</sup> Department of Computer Science and Engineering, Soonchunhyang University, Asan, South Korea

\*Contact: ynam@sch.ac.kr

**Abstract—** Despite the numerous researchers who have conducted studies on fall detection, falls continue to pose a critical concern, particularly among the elderly and individuals with limited mobility. In this study, we propose an enhanced fall detection framework that integrates the detection of near-fall events, thereby providing a more comprehensive safety monitoring solution. We extract dynamic sequences from video streams, capturing the temporal evolution of motion patterns during falls which is commonly used in computer vision and action recognition tasks. The proposed approach merges the strengths of convolutional neural networks (CNNs) architecture specifically tailored for fall detection. The CNNs learn spatial and temporal from dynamic images, enabling robust classification. Moreover, we are leveraging a pre-trained CNN based model which are InceptionV3, and VGG16. We collect new datasets which contains 2124 video clips for falling, 504 video clips for near-fall, and 206 video clips for non-fall incident. The results, coupled with model accuracies of CNN at 97.89%, InceptionV3 at 94%, and VGG16 at 82.63, demonstrate the effectiveness and efficiency of our approach across three classes: fall, near-fall, and non-fall.

## I. INTRODUCTION

Falls represent a major public health concern, especially for older individuals, with those aged 65 and above who live independently at home facing the greatest risk [1]. This issue extends across all age groups and is attributed to factors such as physical weakening, previous falls, reliance on assistive devices, balance, or mobility issues, visual or auditory impairments, and poorly lit environments. Swift fall detection is crucial to minimizing injury severity and ensuring prompt medical intervention.

According to the Public Health Agency of Canada, in 2026, one Canadian older than 65 will be out of five whereas in 2001 the portion was eight to one. Notably, 93% of elderly people stay in their private houses and 29% of them will be led to live a lonely life [2]. There are various types of falling that cause injuries such as loss of balance, falls when sitting down, backward falls, forward falls, and falls by side. These kinds of fall by direction. In addition, physical activities and cardiovascular disorders cause falls.

In recent years, there has been a notable emphasis on developing efficient fall detection systems that leverage technological advancements. These systems integrate various tools such as wearable sensors, sophisticated computer vision techniques, and machine learning algorithms to accurately detect instances of falls, primarily through smartphones [3-4]. However, several limitations persist, particularly regarding the volume of data and challenges associated with the devices used. Wearable sensors face issues like user compliance, discomfort, limited battery life, installation complexity, maintenance requirements, and restricted coverage. Additionally, camera-based systems encounter challenges such as data resolution, scalability, synchronizing data from multiple cameras, and

limitations in input features. Moreover, while fall accidents may not always lead to severe health consequences, healthy individuals are still susceptible to accidental falls, and they often lack access to expensive 24-hour monitoring devices [5]. As a result, there is growing interest in utilizing computer vision for fall detection, particularly through camera video analysis.

Despite the progress made in fall detection using cameras, several limitations persist in existing approaches. These limitations include challenges related to the camera's field of view, the quality of background lighting, the distances between objects and cameras, and the potential for falls occurring outside the designated fall coverage area. Additionally, the complexity of human actions poses a significant challenge in accurately detecting falls, especially since much of the recent research has primarily focused on distinguishing between falls and non-falls only [6].

Furthermore, current research on fall detection often neglects the critical aspect of detecting falls rapidly. While accuracy is important, the speed of detecting a fall is equally crucial. Many existing systems lack the capability to promptly alert caregivers or emergency services, leading to potential worsening of the consequences of falls, particularly for older adults. Recognizing and addressing this gap is essential for advancing fall detection technology and ensuring timely assistance for those in need.

In this study, our primary objective is to detect falls by incorporating near-fall events. Near-fall occurrences serve as early indicators of an impending fall, enabling us to identify and intervene before the fall transpires.

We have proposed a Convolutional Neural Networks (CNN) model that utilizes a single camera to capture video sequences. A key aspect of our approach is the conversion of these video sequences into dynamic images using the rank pooling method with static windows. This technique aims to enhance the model's ability to detect falls efficiently. Our passion lies in leveraging computer vision, as it eliminates the need for individuals to wear any devices on their bodies. We have meticulously collected a high-quality dataset with a sophisticated environment setup to support our research endeavors.

Furthermore, we have structured our dataset to encompass three distinct classes: non-fall, near-fall, and fall. To construct this dataset, we enlisted the participation of twenty-four healthy individuals from both adult and youth demographics, comprising both male and female subjects. Videos were recorded across various environmental settings such as hospitals, homes, roads, and nursing homes, each video segment lasting no more than 10 seconds. This diverse dataset facilitates robust training of our fall detection model, enabling it to effectively differentiate between different scenarios and environmental conditions.

Building upon this dataset, we propose a novel approach for fall detection using Convolutional Neural Networks (CNNs). Our proposed model incorporates a pre-trained CNN base model, complemented by custom layers for feature processing and classification. This strategy allows us to leverage the rich feature representations learned by the base model while tailoring the model specifically for the task of fall detection.

By integrating the pre-trained base model with custom layers, our model learns to extract pertinent features from input images and make precise predictions regarding fall

occurrences. Through meticulous experimentation and evaluation, we showcase the efficacy of our proposed model in accurately estimating falls from visual data, thereby contributing significantly to the advancement of fall detection technology, and enhancing the safety and well-being of vulnerable individuals.

Our contribution of this research is focusing on improving fall prevention technology in two main ways: first, by collecting real-world data that includes different lighting, flooring, and noise conditions while ensuring ethical data acquisition, and second, by incorporating near-fall even which can be the early sign to detect a fall.

This study is organized into five sections. Section II presents a review of related work, Section III outlines our proposed methodology for fall detection using a single camera viewpoint, Section IV presents the experimental results and system discussion, and finally, Section V concludes the work and discusses avenues for future research.

## II. RELATED WORK

In recent years, there has been a notable surge in research focusing on fall detection, reflecting the growing awareness of this issue. While falls may not always lead to severe health problems, they pose a significant risk even to healthy individuals, especially when access to expensive 24-hour monitoring devices is limited. To tackle this concern, researchers and engineers have been developing new fall detection systems that leverage camera video to swiftly detect and alert caregivers or medical personnel in the event of a fall. Numerous studies in fall detection research have contributed to advancing our understanding of this critical field.

Xiao et al. [7] propose a novel approach that extends dynamic imaging techniques from RGB video to depth video, introducing multi-view dynamic images constructed from raw-depth video at different virtual imaging viewpoints within 3D space. This method enhances action representation by capturing motion and temporal evolution information.

Chhetri et al. [8] present a system that utilizes dynamic optical flow to summarize video content, employing the Enhanced Dynamic Optical Flow technique for encoding temporal data from optical flow videos using rank pooling.

Bilen et al. [9] introduce the concept of dynamic images, offering a compact representation of videos that is particularly advantageous for video analysis, especially when employing convolutional neural networks (CNNs).

Rastogi, Shikha, et al. [10] address fall detection (FD) and activity monitoring (AM) using vision-based methods, evaluating various techniques across different camera types and complex indoor and outdoor scenes. Their comparative analysis provides insights into suitable FD and AM techniques.

Singh et al. [11] focus on human activity recognition (HAR) using vision-based methods, proposing a deeply coupled ConvNet that combines RGB and dynamic motion images. Their approach achieves high accuracy on standard RGB-D datasets for single and multiple-person activities, surpassing state-of-the-art methods.

Hazelhoff et al. [12] introduce a real-time fall-detection system for unobserved home environments, utilizing two uncalibrated cameras and employing principal component

analysis and a Gaussian multi-frame classifier for fall detection, achieving over 85% accuracy.

Fernando et al. [13] introduce a rank pooling technique to represent video sequences robustly, enhancing action recognition performance in computer vision and pattern recognition tasks.

Leite et al. [14] propose a multi-stream approach for fall detection using optical flow, saliency map, and RGB data fed into a VGG-16 architecture and classified by an SVM, achieving impressive accuracy rates on URFD and FDD datasets.

Putra et al. [15] introduce an event-triggered machine learning (Event-T-ML) approach for fall detection, aligning fall stages precisely to enhance feature recognition and assessing its effectiveness in accurately detecting fall events.

### III. METHOD

In this work, we explore the effectiveness of various deep learning models, including Convolutional Neural Networks (CNNs), InceptionV3, and VGG16 have been individually trained and evaluated, for fall detection. CNNs have demonstrated remarkable success in image classification tasks due to their ability to learn hierarchical representations from raw pixel data. Additionally, by utilizing pre-trained models like InceptionV3, and VGG16, which have been trained on large-scale image datasets, we leverage transfer learning to harness the knowledge encoded in these models and adapt it to our specific fall detection task.

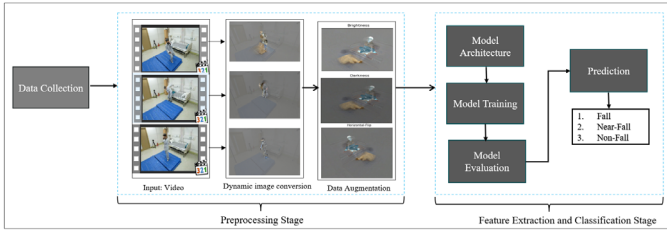


Fig. 1 An Overview of The Proposed Method

The process starts by gathering data from 24 participants across varied scenarios, including differences in room layout, furniture arrangement, and more, where input in the form of video frames is acquired. Next, the preprocessing stage involves dynamic image creation and data augmentation. Dynamic image creation selects relevant frames from the video and convert them into a single image, effectively capturing essential information from the entire sequence as shown in Fig. 2, while data augmentation techniques enhance the dataset's diversity.

Finally, the feature extraction and classification stage employ a model architecture trained on the augmented data to predict fall events. The model provides three possible outcomes: 'Fall,' 'Near-Fall,' and 'Non-Fall.' By evaluating these models, we aim to identify the most suitable architecture for accurately categorizing fall events into 'Fall,' 'Near-Fall,' or 'Non-Fall' outcomes. This approach allowed for a comprehensive evaluation of each model's performance in accurately detecting fall events within video data.

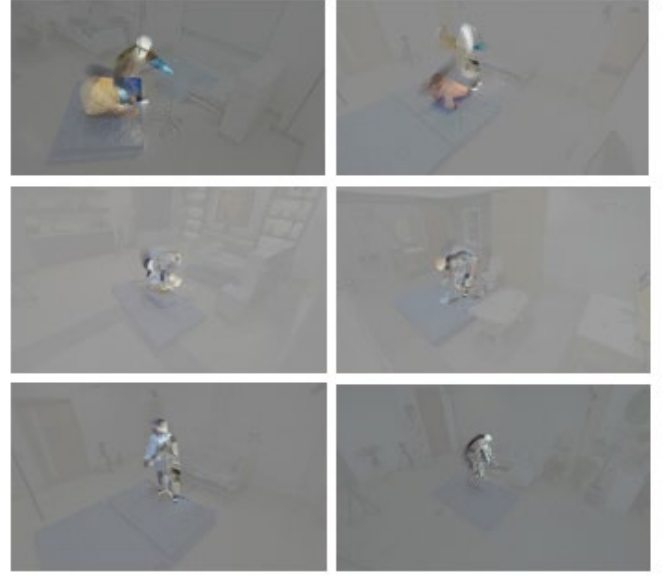


Fig. 2 Illustrates dynamic images extracted from short video sequences, condensed into static images. These images provide a clear, impactful, and efficient representation of videos, which is especially advantageous for fall detection applications.

The examples of dynamic images are generated by summarizing short video sequences into still images by utilizing a rank pooling technique that captures essential moments of action from a sequence of videos. These dynamic images serve as a simplified and efficient representation of the video content. The process of converting videos into dynamic images will be further detailed in a subsequent section.

#### A. Dynamic Image Creation

In this stage, we describe the processing of dynamic creation. A dynamic image is a standard RGB image that summarizes the appearance and dynamics of a whole video sequence.

CNNs are great at learning complex data representations automatically, but they're limited to specific pre-designed architectures. When designing CNNs for video data, we need to consider how to present the video information to the CNNs. Therefore, dynamic image creation plays an integral part in this work.

Our proposed approach capitalizes on the framework introduced by Fernando et al [13]. To generate dynamic images for fall detection. The paper proposed representing a video using the ranking function for its frames ( $I_1, \dots, I_T$ ). Each frame ( $I_t$ ) contributes a feature vector  $\in R^d$ . They compute the time average of these features up to time  $t$ :  $V_t = \frac{1}{t} \sum_{\tau=1}^t \psi(I_\tau)$ . The ranking function assigns a score  $S(t|d)$  to each time  $t$ . The score is based on a vector of parameters  $d \in R^d$ . The goal is to ensure that the scores reflect the rank of frames in the video. Frames occurring later in the video receive larger scores. Specifically, if  $q > t$ , then  $S(q|d) > S(t|d)$ . Learning  $d$  is posed as a convex optimization problem using the RankSVM [16] formulation:

$$d^* = \rho(I_1, I_T; \psi) = \operatorname{argmin}_d E(d),$$

$$E(d) = \frac{\lambda}{2} \|d\|^2 + \frac{2}{T(T-1)} \times \sum_{q>t} \max\{0, 1 - S(q|d) + S(t|d)\} \quad (1)$$

The first term is a standard quadratic regularize commonly used in SVMs. The second term is a hinge-loss soft-counter that tracks how many pairs of frames are incorrectly ranked by the scoring function. It's important to note that a pair is only considered correctly ranked if their scores are separated by at least a unit margin ( $S(q|d) > S(t|d) + 1$ ).

Optimizing equation (1) defines a function  $\rho(I_1, \dots, I_T; \psi)$  that transforms a sequence of T video frames into a single vector  $d^*$ . Since this vector contains sufficient information to rank all frames in the video, it aggregates information from all frames and can serve as a video descriptor. Throughout the paper, we refer to the process of constructing  $d^*$  from a sequence of video frames as rank pooling.

In [13] the mapping  $\psi(\cdot)$  utilized in this construction is based on the Fisher Vector coding of various local features (such as HOG, HOF, MBH, TRJ) extracted from individual video frames.

The  $\psi(I_t)$  function now combines the RGB components of each pixel in the image  $I_t$  into a large vector. Alternatively,  $\psi(I_t)$  may incorporate a simple component-wise non-linearity, such as the square root function  $\sqrt{\cdot}$  (which corresponds to using the Hellinger's kernel in the SVM). In either case, the resulting descriptor  $d^*$  is a real vector with the same number of elements as a single video frame. Thus,  $d^*$  can be interpreted as a standard RGB image. The result of dynamic image creation for fall near-fall and non-fall showed in Fig.3

Additionally, we resize each frame to dimensions of 1920 by 1080 during converting video to dynamic image, this helps in creating visual representations of the data since the original frames are too large.

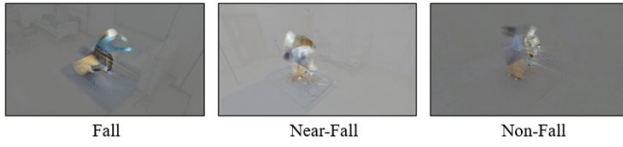


Fig. 3 Sample Dynamic image from each class

### B. Data Augmentations

In this section, we first work on data augmentation. In the data augmentation, we applied three techniques which are brightness with `brighter_factor = 1.2` and darkness with `darker_factor = 0.8` are for lighting condition. Additionally, we also adjusted the horizontal flip data augmentation for camera angle. The sample of augmentation result shown in in the Fig 4.

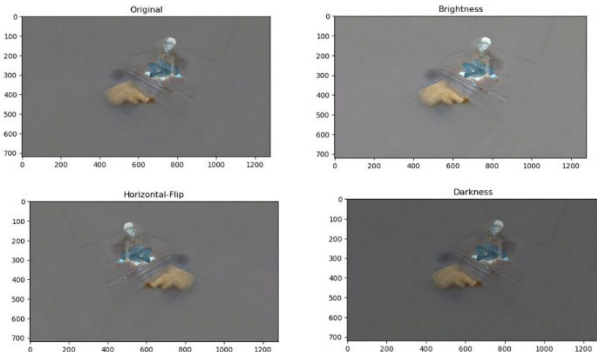


Fig. 4 Sample Dynamic image for data augmentation

### C. Oversampling

The dataset exhibited significant class imbalance, as illustrated (a) in Fig. 5. To address this issue, a selective duplication approach was employed for oversampling. This involved duplicating and selecting specific instances from the minority class to match the frequency of the majority class. By strategically duplicating instances, we aimed to address class imbalance while minimizing redundancy and preserving the integrity of the dataset.

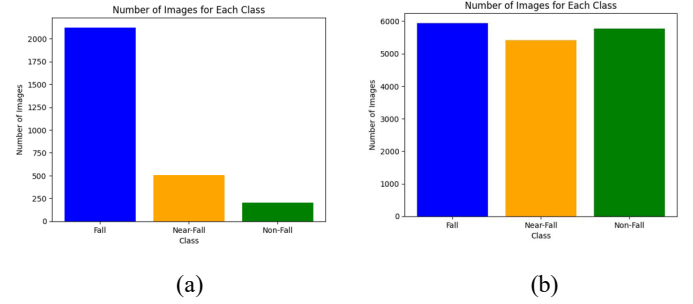


Fig. 5 Before (a) and After(b) oversampling

### D. Proposed Model Architecture

#### 1. Convolutional Neural Networks (CNNs)

Fig. 3 showed the proposed CNN model architecture. CNNs are designed for image-related tasks, including action detection. Within the intricate architecture of CNNs, each component plays a vital role in dissecting and comprehending the complexities of image data.

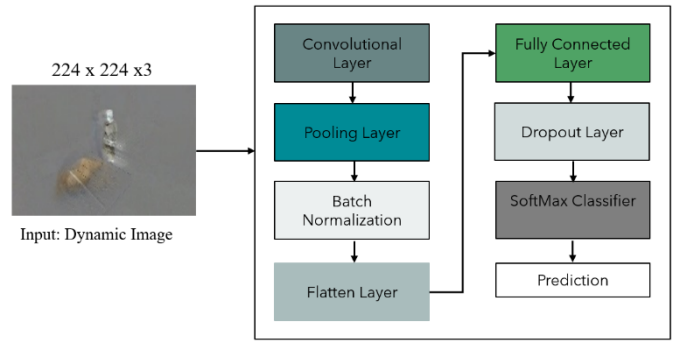


Fig. 6 The architecture of the proposed CNN

The proposed CNN model, outlined in Fig. 6, starts with an input layer for RGB images (224x224x3 pixels) and consists of three convolutional layers followed by ReLU activation functions. The first layer utilizes 128 filters (5x5) to capture low-level features, followed by layers with 64 and 32 filters (3x3), respectively. Regularization techniques like L2 regularization are employed to prevent overfitting. Max pooling layers downsample feature maps using 2x2 pooling windows. The flattened outputs are then fed into a fully connected neural network with two dense layers. The first dense layer has 256 units with ReLU activation and dropout regularization (dropout rate: 0.5). The output layer, with three units and softmax activation, computes class probabilities for classification tasks.



small penalty coefficient, are incorporated into these layers to mitigate overfitting.

## 2. Pre-trained CNN-based Model

Following both models layers, a MaxPooling2D layer downsamples spatial dimensions of feature maps, retaining vital information. The output is flattened into a one-dimensional tensor for subsequent dense layers. A dense layer with 512 units and ReLU activation enables learning complex patterns. Dropout with a rate of 0.5 prevents overfitting by deactivating neurons during training. The output layer, with softmax activation, classifies into fall, near-fall, and non-fall. Compiled with RMSprop optimizer (learning rate: 0.0001) and categorical cross-entropy loss, accuracy is monitored as the evaluation metric.

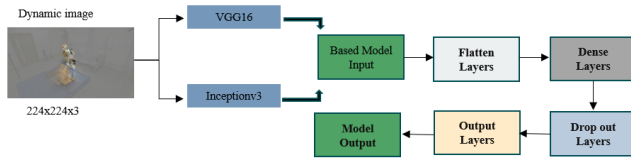


Fig. 7 The modified architecture of proposed Pre-trained CNN-based model

## IV. RESULT

The experimental process such as experimental result from proposed model, data splitting is discuss in this section. we collect own dataset and divide into 70:15:15. Its mean that 70% dataset is used for the tranining, 15% dataset is used for validation and testing, respectively. In the training phase, we set the epoch to 200 for CNN model and 300 epochs for pre-trained models and batch size is 32. The learning rate is 0.0001 and we utilized the Adam optimizer for the learning process only CNN and RMSprop for pretrained model. Multiple classifies are used and each classifies evaluated using four metrics including Recall Rate, F1-Score, Precision, and Accuracy.

TABLE I  
THE COMPARISON OF THE MODEL EVALUATIONS

Model	Accuracy	F1-Score	Recall	Precision
	(%)	(%)	(%)	(%)
CNN	97.89	98	98	98
InceptionV3	94	94	94	94
VGG16	82.63	85	83	90

The evaluation results of various convolutional neural network (CNN) architectures demonstrate significant variances in performance metrics. CNN model exhibits exceptional accuracy, with an impressive rate of 97.89%, alongside a balanced F1-score, recall, and precision, all maintaining a high threshold at 98%. InceptionV3, while slightly trailing behind the CNN model, still achieves commendable results with an accuracy of 94%, along with consistent F1-score, recall, and precision metrics at 94%. However, VGG16 demonstrates the lowest performance among the evaluated models, with an accuracy of 82.63%, albeit with a relatively higher precision score of 90%. These results underscore the importance of

selecting an appropriate neural network architecture tailored to specific tasks, with the CNN model emerging as the frontrunner in this comparative analysis.

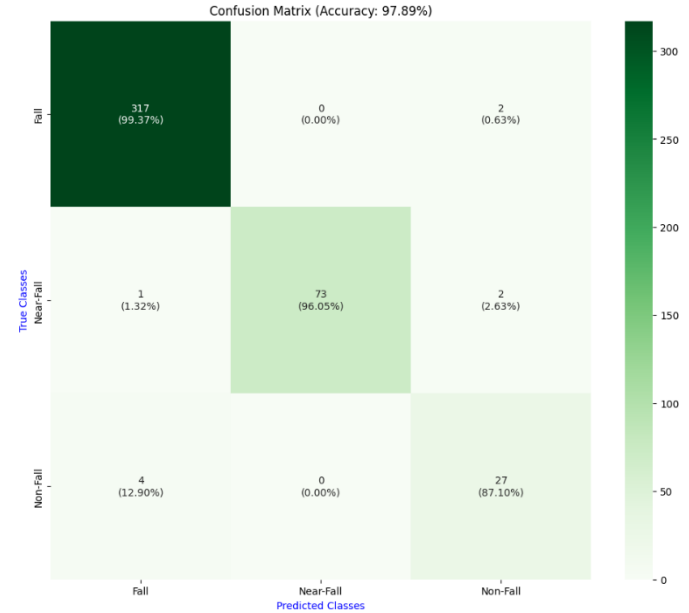


Fig. 8 Confusion Matrix of Proposed CNN Model

The model's performance varies across different classes. For the "Fall" class, it accurately predicted 99.37% of falls, with only 2 misclassifications. In the "Near-Fall" class, it correctly identified 96.05% of instances, with 2 misclassifications. However, in the "Non-Fall" class, it had an 87.10% accuracy, with 4 misclassifications.

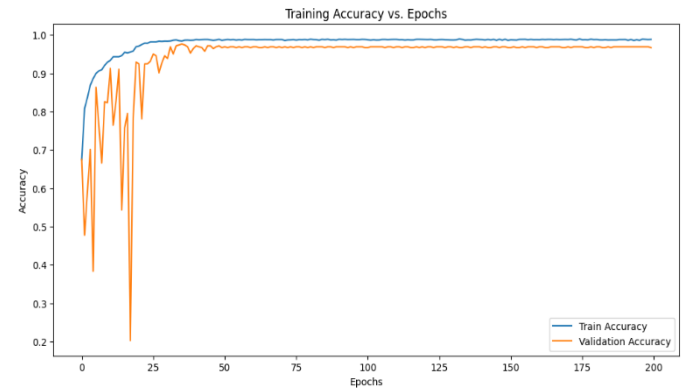


Fig. 9 Training and Validation Accuracy Vs. Epochs of CNN

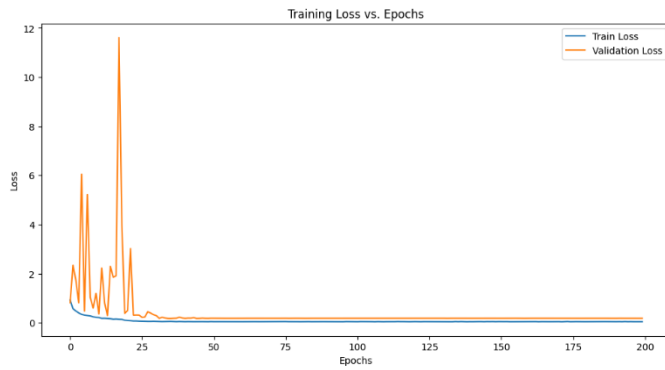


Fig. 10 Training and Validation Accuracy Loss Vs. Epochs of CNN

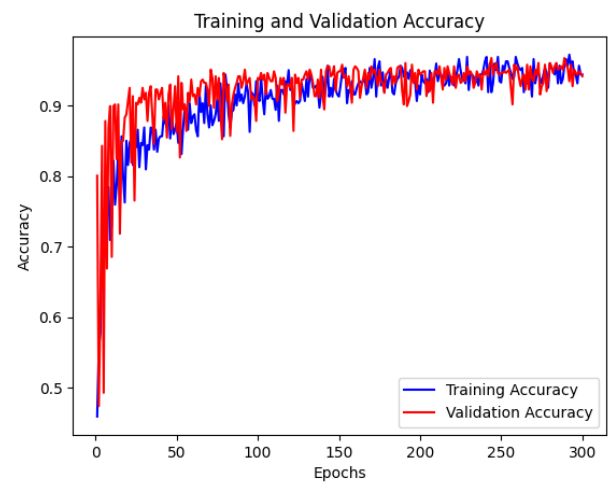


Fig. 12 Training and Validation Accuracy Vs. Epochs of InceptionV3

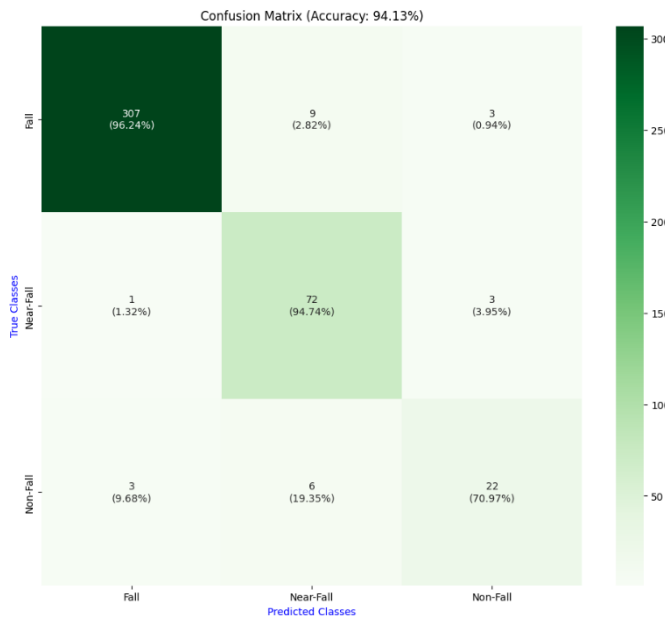


Fig. 11 Confusion Matrix of Proposed InceptionV3 Model

In class-specific results, the model exhibits robust performance in identifying falls and near-falls, showing minimal misclassifications. Specifically, for the "Fall" class, it achieved a high true positive rate of 96.24%, with only a small percentage of instances misclassified as non-falls. Similarly, in the "Near-Fall" class, the model attained a true positive rate of 94.74%, with only a few instances incorrectly labeled. However, in the "Non-Fall" class, the model's performance was comparatively lower, with a true positive rate of 70.97% and a higher rate of misclassifications.

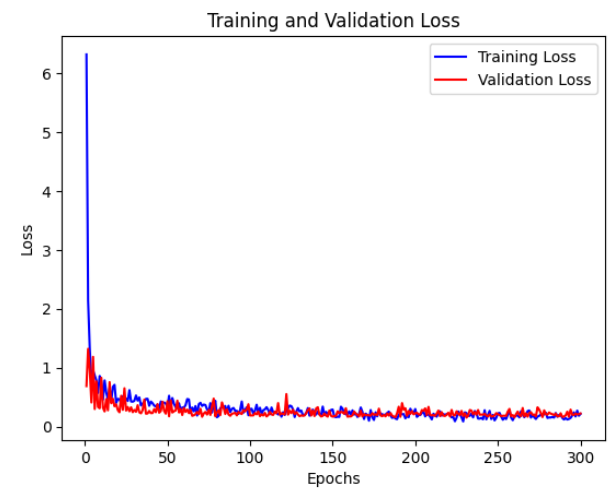


Fig. 13 Training and Validation Accuracy Loss Vs. Epochs of InceptionV3

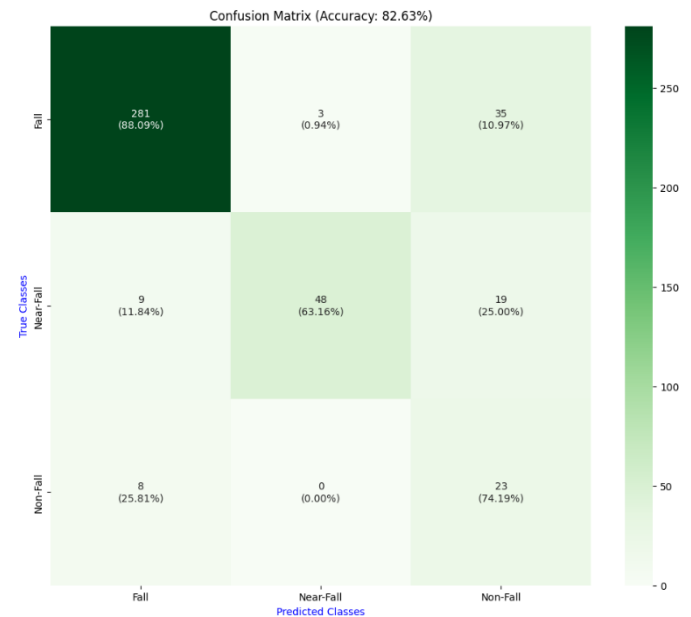


Fig. 14 Confusion Matrix of VGG16 Model



The model achieves an overall accuracy of 82.63%. Across different categories, it correctly identifies instances of 'Fall' with a true positive rate of 88.09%, while experiencing false positives at a rate of 0.94% and false negatives at 11.84%. For 'Near-Fall', the true positive rate is 63.10%, with false positives at 25.00% and false negatives at 11.84%. In the 'Non-Fall' category, the model demonstrates a true positive rate of 74.19% and a false negative rate of 25.81%. These metrics provide a detailed understanding of the model's performance across varied classes, guiding potential improvements and optimizations.

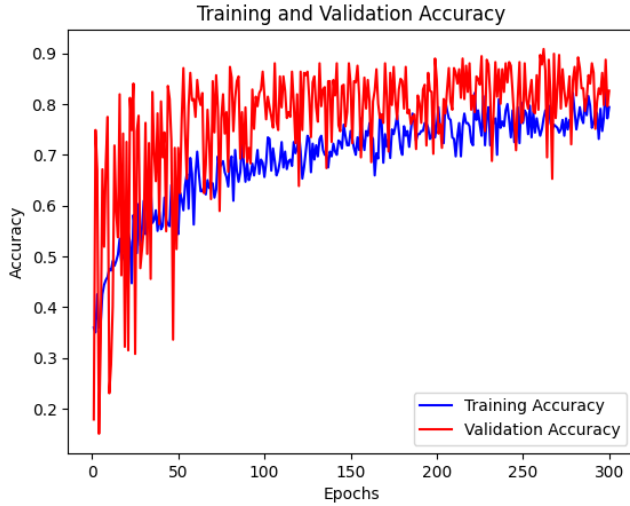


Fig. 15 Training and Validation Accuracy Vs. Epoch of VGG16

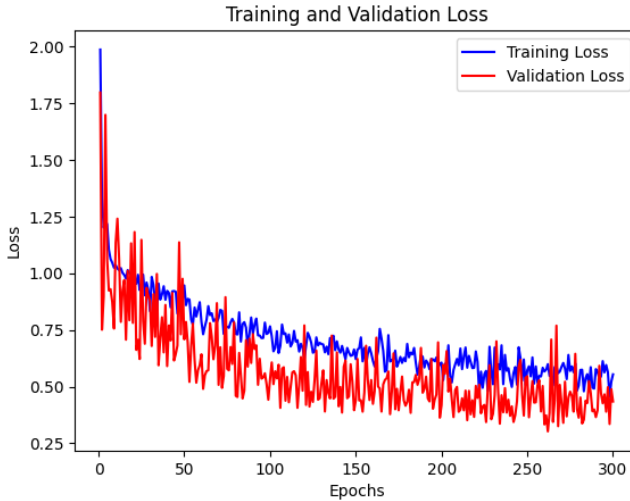


Fig. 16 Training and Validation Accuracy Loss Vs. Epochs of VGG16

Below are the equations and descriptions for key performance metrics and tools often used to evaluate and display the results of a classification model:

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (2)$$

$$Precision = \frac{TP}{TP+FP} \quad (3)$$

$$Recall = \frac{TP}{TP+FN} \quad (4)$$

$$F1\_score = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (5)$$

Where:

True Positive (TP)

- The predicted value matches the actual value, or the predicted class matches the actual class.
- The actual value was positive, and the model predicted a positive value.

True Negative (TN)

- The predicted value matches the actual value, or the predicted class matches the actual class.
- The actual value was negative, and the model predicted a negative value.

False Positive (FP)

- The predicted value was falsely predicted.
- The actual value was negative, but the model predicted a positive value.

False Negative (FN)

- The predicted value was falsely predicted.
- The actual value was positive, but the model predicted a negative value.

## V. CONCLUSION AND FUTURE WORK

This study aimed to detect early signs of falls which is near-fall by leveraging the rank pooling technique to capture the temporal evolution of motion patterns during falls effectively. Our approach combines convolutional neural networks (CNNs) and pre-trained models such as VGG16 and Inceptionv3, customized for fall detection, to learn spatial and temporal features from dynamic images, thus enhancing classification accuracy.

Our future research will prioritize advancements in fall detection systems through targeted strategies. Firstly, we will explore innovative of decreasing time between detection and actual fall and expand the dataset to include various environmental settings to better simulate real-life falling scenarios, thereby refining the model's ability to discern nuanced patterns in real-time fall situations. Additionally, we plan to explore the potential of time-series data models for dynamic sequence analysis. These endeavors are aimed at propelling the evolution of fall detection technology, ultimately ensuring heightened safety and well-being for individuals prone to falls.

## ACKNOWLEDGMENT

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2023-00218176)

## REFERENCES

- [1] Chandak, Ayush, and Nitin Chaturvedi. "Machine-learning-based human fall detection using contact-and noncontact-based sensors." *Computational intelligence and neuroscience* 2022 (2022).
- [2] Shu, F., Shu, J. An eight-camera fall detection system using human fall pattern recognition via machine learning by a low-cost android box. *Sci Rep* 11, 2471 (2021).
- [3] Rakhman, Arkham Zahri, and Lukito Edi Nugroho. "Fall detection system using accelerometer and gyroscope based on smartphone." In *2014 The 1st International Conference on Information Technology, Computer, and Electrical Engineering*, pp. 99-104. IEEE, 2014.
- [4] Igual, Raul, Carlos Medrano, and Inmaculada Plaza. "Challenges, issues and trends in fall detection systems." *Biomedical engineering online* 12, no. 1 (2013): 66.
- [5] Shu, F., Shu, J. An eight-camera fall detection system using human fall pattern recognition via machine learning by a low-cost android box. *Sci Rep* 11, 2471 (2021).
- [6] Shotkit, Marsupial Drive, Pottsville, NSW Australia
- [7] Xiao, Yang, Jun Chen, Yancheng Wang, Zhiguo Cao, Joey Tianyi Zhou, and Xiang Bai. "Action recognition for depth video using multi-view dynamic images." *Information Sciences* 480 (2019): 287-304.
- [8] Chhetri, Sagar, Abeer Alsadoon, Thair Al-Dala'in, P. W. C. Prasad, Tarik A. Rashid, and Angelika Maag. "Deep learning for vision-based fall detection system: Enhanced optical dynamic flow." *Computational Intelligence* 37, no. 1 (2021): 578-595.
- [9] Bilen, Hakan, Basura Fernando, Efstratios Gavves, Andrea Vedaldi, and Stephen Gould. "Dynamic image networks for action recognition." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3034-3042. 2016.
- [10] Rastogi, Shikha, and Jaspreet Singh. "Human fall detection and activity monitoring: a comparative analysis of vision-based methods for classification and detection techniques." *Soft Computing* 26, no. 8 (2022): 3679-3701.
- [11] Singh, Tej, and Dinesh Kumar Vishwakarma. "A deeply coupled ConvNet for human activity recognition using dynamic and RGB images." *Neural Computing and Applications* 33 (2021): 469-485.
- [12] Hazelhoff, Lykele, Jungong Han, and Peter HN de With. "Video-based fall detection in the home using principal component analysis." In *Advanced Concepts for Intelligent Vision Systems: 10th International Conference, ACIVS 2008, Juan-les-Pins, France, October 20-24, 2008. Proceedings* 10, pp. 298-309. Springer Berlin Heidelberg, 2008.
- [13] Fernando, Basura, Efstratios Gavves, José Oramas, Amir Ghodrati, and Tinne Tuytelaars. "Rank pooling for action recognition." *IEEE transactions on pattern analysis and machine intelligence* 39, no. 4 (2016): 773-787.
- [14] Leite, Guilherme, Gabriel Silva, and Helio Pedrini. "Fall detection in video sequences based on a three-stream convolutional neural network." In *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*, pp. 191-195. IEEE, 2019.
- [15] Putra, I. Putu Edy Suardiyana, James Brusey, Elena Gaura, and Rein Vesilo. "An event-triggered machine learning approach for accelerometer-based fall detection." *Sensors* 18, no. 1 (2017): 20.
- [16] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression", *Statistics and computing*, vol. 14, pp. 199-222, 2004.

# Children-adult speaker Diarization using Multi-modal Model

YunJung Hong<sup>1\*</sup>, and Jiyoung Woo<sup>2</sup>

<sup>1</sup> Dept. ICT convergence, University of SoonChunHyang, Asan, South Korea

<sup>2</sup> Dept. AI and Bigdata, University of SoonChunHyang, Asan, South Korea

\*Contact: mobu6765@gmail.com

**Abstract**— Speaker recognition technology has emerged as an important technology for identification in various systems such as e-commerce, forensic science, judicial enforcement, and artificial intelligence based conversation interfaces in modern society. This study proposes a multi-modal model that extracts speech intervals for each speaker by extracting speech intervals using Silero vad from audio data and by classifying them based on age from speech intervals. The model is based on the representative speech signal features Spectral Contrast, Spectral Roll-Off, and Spectral Bandwidth, statistical values for the signal feature peaks, and Mel-Spectrograms images representing speaker characteristics. The multi-modal model is designed based on Vision Transformer, which is excellent at learning the overall spatial features and complex patterns of images through a linear embedding process by dividing the images into patch units. This study is expected to contribute to further the foundation of a more comprehensive and adaptive system that can control response and interaction with age.

## I. INTRODUCTION

The voice serves as a significant medium capable of extracting diverse information about the speaker, including gender, age, health, and psychological state, among others. Consequently, voice analysis and recognition technology holds substantial applicability across various domains and is readily encountered in everyday life. However, most existing speech recognition models tend to be based on databases of adult speakers, leading to limited diversity in recognizing children's age groups who exhibit differences from adults in pitch and pronunciation, resulting in lower accuracy. Additionally, when recognition fails, children, compared to adults, find it challenging to accept clear pronunciation and demands for tone and expression changes. Therefore, in this era where speech systems are essential, from education to therapeutic assistance, the need for speech recognition systems focused on children is becoming increasingly imperative. Moreover, information about how much and how children speak in interactions includes information about their developmental status and can serve as a crucial clue for diagnosis, especially in cases like autism spectrum disorder where early detection allows for effective treatment. In order to automatically extract speech features (or interactions) for the development of such child-related speech systems, a classification step is essential to distinguish between the voice of a guardian or medical staff and the child's voice in conversational speech.

In this study, we develop a two-step speech recognition model targeting preschool children using conversation data sets between children and adults. By using both voice features and mel spectrum images mainly used in existing research, we propose a multi-modal model that learns changes and patterns of voice data over time in addition to signal features. This study is expected to contribute to the development of a system that provides customized treatment and services according to the child's language development level, disposition, emotions, and condition by automatically extracting the child's voice segments.

## II. RELATED WORK

Tursunov et al.[1] proposed a CNN model containing two multi-focus modules (MAMs) to effectively extract spatial and temporal information, achieving 76% and 90% accuracy on Common Voice and Korean Speech Recognition datasets. However, unlike this study, the used Korean data consists of high-quality recordings inputted from quiet rooms, which is expected to be difficult to use in real life with a lot of noise. Safavi et al.[2] applied mainly used methods in speech signal analysis, such as the Gaussian Mixture Model-Universal Background Model (GMM-UBM), GMM-Support Vector Machine (GMM-SVM), and i-vector-based approaches, focusing on the voices of children with relatively few studies. The considerations for children's voices, such as changes in spectral information due to the high frequency of children's voices, such as the transformation of adolescence, were summarized, and the useful frequencies of gender and speaker identification were discovered through several experiments. Ghahremani et al. [3] used an x-vector deep neural network (DNN) architecture for age estimation based on speaker's speech signals by mapping variable-length utterances to fixed-dimensional embedding vectors that hold relevant sequential information. Zazo et al. [4] attempted to supplement the existing methods such as i-vector extraction and low accuracy on short speech samples of artificial neural networks. Accordingly, we proposed a real-time age estimation system based on LSTM recurrent neural networks (RNNs) that can handle short utterances based on speech features.

## III. METHOD

### A. Data

The dataset used in this study was obtained from the "Free Conversation Speech (Adults, Pre-schoolers)" and "Command Speech (Pre-schoolers)" public datasets provided by AI Hub. Specifically, we focused on speech samples from adults aged 25 to 45 and children aged 3 to 6, considering the language development stages. From each dataset, we utilized 38,150 samples of free conversation (Pre-schoolers), 140,068 samples of command speech (Pre-schoolers), and 141,803 samples of free conversation (Adults). Each audio file consists of a single sentence or word.

The average duration of the files is approximately 5 seconds for adults and 3 seconds for children. However, for the developmental measurements intended to be applied in this study, the recorded data typically consists of longer durations, often lasting several tens of minutes in a question-and-answer format. Additionally, compared to the typical minimum duration of 10 seconds for windowing speech files, the AI Hub data contains durations that are too short. To address this, we randomly extracted 8 files (3 for adults, 5 for pre-schoolers) without duplication and combined them to create speech data with a minimum duration of 30 seconds.

Among the generated 28,000 synthetic data samples, speakers are categorized only as adults or children. A gap of 0.5 to 3 seconds is added when transitioning between speakers, while a random gap of 0.5 to 1 second is inserted consecutively for the same speaker. Subsequently, the data was structured according to the Kaldi documentation format.

### B. Pre-processing

The audio data contains excessive information unsuitable for direct model learning, with a daunting rate of 16,000 samples per second, making it impractical for effective training. Therefore, in this study, the audio data underwent pre-processing, involving the extraction of features pertaining to the characteristics of the speech signal, such as pitch, specific energy ratios, and frequency distribution, in order to prepare it for model training. When importing files, a random decibel adjustment within the range of  $[-5, 5]$  was applied to each voice file to account for the diverse recording conditions observed in real-world scenarios.

### C. Step 1 : Voice Activity Detection(VAD)

VAD is a technology that distinguishes between the part containing voice and the part containing only silence/noise in a voice signal. By detecting the starting and ending points of speech in the input signal and extracting only accurate speech data, the input data of the speaker classification model can be reduced. Additionally, imbalances between silence and speech data can be prevented and learned efficiently.

This study used Silero VAD, a pre-trained model released in 2021. Silero v4 16k version was used, and vad was extracted by setting the sample size to 512, the minimum

utterance length to 0.3ms, and the threshold to 0.9. The results are shown in Figure 1.

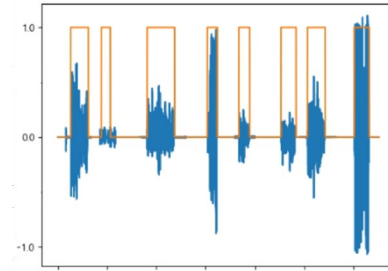


Fig. 1 Visualization of VAD results.

### D. Step 2 : Speaker classification

#### (1) Feature

A 64-dimensional Mel-Spectrogram image, voice features, and signal features were extracted from the speech section extracted in Step 1 in 1-second increments.

The Mel-Spectrogram is a visual representation that maps frequencies to the Mel scale, reflecting the human auditory system, allowing you to see spectral changes over time. Because frequencies are correlated, it shows better performance in limited-domain problems, and features such as the speaker's age and gender can be extracted through image patterns. In this study, the size of the FFT is 2048, the Hop length is 256, and the Mel filter The number was set to 64, and in order to improve calculation efficiency and make it easier to input patches of a fixed size into the Vision Transformer, zero padding was applied to unify the size to  $[63*64]$ .

As the voice signals, Spectral Centroid, Spectral Roll-off, Spectral Bandwidth, Zero Crossing Rate, Tempo(BPM) were used. Spectral Centroid is an index that represents the spectral center of the voice signal by calculating the weighted average of the frequencies, and means the height of the voice signal. Spectral roll-off refers to a frequency below a specific ratio of the total spectral energy, and the higher the value, the richer and brighter the sound can be known. Spectral Bandwidth is the width of the frequency range within the spectrum, and the spectral bandwidth tends to decrease with age. The Zero Crossing Rate is a measure of the number of times the amplitude crosses zero in a voice signal and is used as an indicator of the rate of change of the voice signal. Tempo (BPM) is the number of bits per minute and refers to the rhythm and speed of the music. If the indicators have continuous values, representative values such as mean, maximum, and minimum values were extracted and used.

Signal functions include rising and falling times, inter-peak amplitude, RMS(Root mean square), wave, skewness, kurtosis, mean, absolute minimum, absolute maximum, shape ratio, low-frequency/high-frequency filtering, and maximum values. After applying filtering, the power spectral density was calculated using the Welch method, and the maximum peak value was obtained.

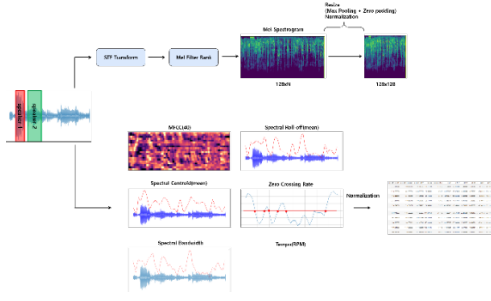


Fig. 2 Feature extraction process

## (2) Multi-modal Model

The model proposed in this study is a multi-modal structure that combines a VIT model using input data and a CNN model using voice/signal features using input data for Mel spectral images, which can be seen in Figure 3. With this structure, we tried to learn both the detailed spectral characteristics and temporal patterns of voice signals.

Vision Transformer is an architecture for image processing and applies Transformer, which was originally used in the natural language processing domain, to the image processing field. The entire image is divided into small patches of fixed size and unfolded in a vector format, and location embedding are added to integrate location information into each patch. These combined vectors pass through the Transformer encoder layer to extract features of an image containing location information. As for the voice data, the VIT model was selected in that the change over time is important.

Finally, the results from the VIT model are concatenated with the CNN model results to classify the speaker of that sample into multi-label forms.

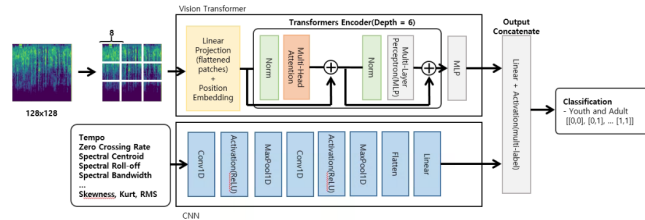


Fig. 3 Multi-modal model architecture

## IV. RESULTS

In this section, we present the results of using the multimodal model for speaker segmentation of preschool-adult conversations and compare it with the performance of the SA-EEND model, one of the representative models of speaker segmentation.

Unlike previous studies using only log-mel spectra or MFCC, other acoustic features and signal statistics were additionally used in this study, and the output results of the SA-EEND model and this study model are shown in Figure 4, and the performance values for each model can be found in Table 1.

The DER of the multimodal model and SA-EEND model in this study achieved 0.64 and 0.78, respectively, and the multimodal model showed higher performance.

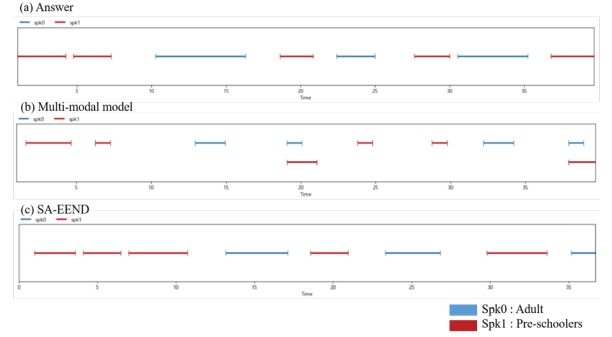


Fig. 4 Model output visualization

TABLE I . DER SCORE BY MODEL

Model	DER
Multi-modal Model	0.64
SA-EEND	0.78

## V. CONCLUSIONS

In this study, an experiment was conducted to classify speaker speech sections by age by extracting various features from audio data and applying them to a multi-modal model. It showed higher performance than the existing model, and the possibility of a multi-modal model was confirmed.

The data used in this study created a speech section with the length of the file from the original data of the synthesized data, and in the process, a part other than the voice section is included, so a vad process for the original file is required. In future work, it is expected that the performance can be improved by additionally proceeding with solutions to these problems.

## ACKNOWLEDGMENT

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. RS-2023-00218176).

## REFERENCES

- [1] Tursunov, Anvarjon, et al. "Age and gender recognition using a convolutional neural network with a specially designed multi-attention module through speech spectrograms." *Sensors* 21.17 (2021): 5892.
- [2] Safavi, Saeid, Martin Russell, and Peter Jančovič. "Automatic speaker, age-group and gender identification from children's speech." *Computer Speech & Language* 50 (2018): 141-156.
- [3] Ghahremani, Pegah, et al. "End-to-end Deep Neural Network Age Estimation." *Interspeech*. Vol. 2018. 2018.
- [4] Zazo, Ruben, et al. "Age estimation in short speech utterances based on LSTM recurrent neural networks." *IEEE Access* 6 (2018): 22524-22530.



# Speech Emotion Classification Using Acoustic and Spectral Features with Machine Learning

Gati L. Martin, and Jiyoung Woo\*

*Department of Future Convergence Technology, Soonchunhyang University, Asan, South Korea*

*\*Corresponding author: [jywoo@sch.ac.kr](mailto:jywoo@sch.ac.kr)*

**Abstract**—Emotions play an important role in physiological activity, social interaction, and decision-making. Emotions can be expressed in speech, facial expressions, writing, or vital signs. Compared to facial expressions, speech is less affected by attributes such as movement, glasses, and beards. Speech emotion recognition (SER) has gained significant attention in recent years due to its wide range of applications, such as depression diagnosis, online education, and mental health monitoring. This study proposes a children’s emotion detection model based on the acoustic and spectral features of a speaker’s verbal cues. Architectures such as support vector machine (SVM), random forest (RF), vision transformers (ViT), and convolutional neural networks (CNN) have been used to test the emotion-capturing capability.

## I. INTRODUCTION

Emotions play a crucial role in our daily lives and can be reflected in our routines and behaviours. With the advancement of technology, emotion recognition has become widely applicable in various fields such as human-computer interaction, medical health, Internet education, security monitoring, and psychological analysis. Therefore, emotion recognition can be assessed by analyzing facial expressions, speech, behaviour, or physiological signals [1]. Speech is the primary means of human communication. Every utterance contains information about the message, the speaker, the emotion, and the language. Speech emotion recognition (SER) is an important research area with wide applications. It involves detecting speakers’ emotions from their voices. Hence, it is necessary to develop an algorithm that is human-like and can accurately detect emotions.

Several studies have used machine learning models to detect emotions from auditory data. The features that SER can extract are categorised into spectral (mel frequency cepstral coefficients (MFCCs), linear prediction cepstral coefficients (LPCCs), perceptual linear prediction (PLP), gammatone frequency cepstral coefficients (GFCCs), Formants), prosodic (energy, pitch frequency), and voice quality (jitter, shimmer, harmonics to noise ratio) [2]. Different approaches can be used to derive features from audio signals. The most common is to divide the signals into speech frames and extract low-level features. Acoustic features are widely used and effective for most existing ML-based SER studies [3]. They include voice quality, prosodic, and spectral features [4]. Many studies have been carried out using different features. The popular classifiers adopted range from traditional models such as SVM, Gaussian Mixture Model (GMM) to deep learning models

such as CNN and LSTM. Recently, attention has been paid to adopting transformer-based models (e.g., Wav2Vec, Hubert, ViT). In the study [5], authors apply hierarchical SVM with linear and RBF kernels using MFCC features and classify seven emotions. Additionally, prosodic features were compared against MFCC; however, the performance with prosodic was only 48% compared to 68% with MFCCs. [6] used a combination of MFCC, pitch, and energy features for emotion recognition tasks on three datasets. Moreover, the performance of different feature combinations and SVM classifier settings was compared. The study concluded that a combination of various features provides different results, and the sensitivity of emotional features in different languages is also different. The advent of deep learning techniques has radically changed the way audio-speech data is processed. Architectures such as CNN and ViT have been adopted to classify audio signals. Although the architectures are borrowed from computer vision, the performance has greatly improved by using the image-form spectral feature. [7] presents an in-depth review of deep learning techniques, limitations, and feature processing. [8] proposed a SER framework using ID CNN and combining five features (Chromagram, MFCCs, Mel-Spectrogram, Tonnet, and Contrast) as input. The model was evaluated on three datasets and achieved a good accuracy of over 60%. Combinations of features such as Mel-Spectrogram, MFCC, and raw spectrogram magnitudes were analyzed to identify the best combination of features for CNN and LSTM [9]. The results show that the MFCC feature provided the best performance and reported high accuracy rates. Recently, attention-based models have gained popularity, and few studies have applied them to SER tasks. [10] proposes self-attention-based deep learning (two-dimensional CNN and LSTM). The authors performed extensive experiments with different combinations of spectral and rhythmic features. MFCCs emerged as the best-performing features, with an average accuracy of 90% using a combination of three datasets. Kumar, CS Ayush, et al. [11] proposed a comparative study between CNN-LSTM and ViT for speech emotion detection. The performance was 88.50% and 85.36% respectively. The Mel spectrogram was used as the input feature for the ViT model. Table I summarises related works based on their features and algorithms.

Through the literature review, several research gaps were identified, as follows: First, most researchers focused on SER from adult speech, and few are from children’s speech. This is due to the lack of publicly annotated children datasets and the

TABLE I  
SUMMARY OF THE RELATED WORKS ON SPEECH DETECTION

Study	Data (class)	Features	Methodology	Algorithm	Acc(%)
[5], 2013	EmoDB (7)	MFCC	The 3-stage hierarchical SVM is proposed to separate 7 emotions individually.	SVM	68.00
[6], (2015)	EmoDB, MalayalamDB (4)	Pitch, energy, spectral features	Built-in Binary tree, one against one and one versus the rest methods using both linear and RBF kernel.	SVM	75.00, 95.83
[9], (2019)	EmoDB, IEMOCAP (4)	mel spectrogram, magnitude spectrogram, MFCC	Various features are tested with the architectures to reveal the best feature-architecture combination.	CNN, LSTM	82.35
[8], (2020)	RAVDESS (8), EmoDB (7), IEMOCAP (4)	MFCC, Chromagram, Mel-Spectral, contrast, Tonnetz	Combine multiple sound spectral representations features (pitch, timbre, harmony, etc.). Tune model by adding, removing, and modifying some layers.	CNN	71.61, 86.10, 64.40
[10], (2023)	RAVDESS, SAVEE, TESS (7)	Spectral features	To identify the best-performing features on different combinations of spectral and rhythmic information (ZCR, RMS, Tempogram, Chroma).	Attention-based CNN-LSTM	90.0
[11], (2022)	EmoDB (4)	MFCC, Mel-spectrogram	Attention-based deep learning techniques to extract feature.	CNN-LSTM, ViT	88.50, 85.36
Ours	AIHub (5)	Mel-spectrogram	Feature extraction and emotion detection on children speech.	CNN, ViT	75.00, 79.81

difficulty of emotional expression in acting speech. Many preschoolers either do not have enough vocabulary to identify feelings or find it difficult to read long texts. Second, most studies focus on spectral features (particularly MFCC), and few have used acoustic features. The features that are found to be important in emotion recognition include MFCC, LPCC, PLP, GFCC, pitch, energy, loudness, frequency, jitter, harmonicity, shimmer, etc., which together form acoustic features.

In this study, we classify five emotions: confrontation, curiosity, denial, positive, and neutral, by analyzing the spectral and acoustic features extracted from the children's speech. We proposed machine-learning models, SVM, RF, and CNN as classifiers. In addition, we used the Mel-spectrogram feature, a visual representation of the spectrum of the audio signal, as input to CNN and ViT. We achieved an accuracy of 79.81% and 75.00% for ViT and CNN, respectively.

## II. METHODOLOGY

In our methodology, we perform two steps: feature extraction and classification. Moreover, we performed feature selection for the input features of the traditional machine learning classifiers (SVM and RF). The deep-learning-based models require minimal feature processing as the model learns the global features.

### A. Features

The right features are the key to a good-performing model. In speech, there are different features, with spectral and rhythmic features being the most prominent. We focused on acoustic features that describe the variation in the pronunciation of speech patterns, such as pitch intensity, spectral features, and the vibration frequency of the voice. The waveforms in the audio files are represented in the time domain. To extract features, the speech waveform is transformed into a parametric representation at a low data rate for further processing and analysis. First, the audio signal is divided into frames (20-40ms), and the Fast Fourier Transform is applied to obtain the spectrum. The spectrum shows the strength of each frequency band from which sound characteristics can be extracted.

In addition, speech signals can also be represented in graphical forms, such as time-frequency spectrograms. MFCCs can be used to create spectrograms, which allow the transfer of a sound waveform into the image domain. A Mel spectrogram is a visual representation of the spectrum of frequencies in an audio signal over time. It provides a 2D representation where the color intensity represents the amplitude or energy of each frequency component at different time intervals. MFCCs are derived from the Mel Spectrogram but are further processed to extract relevant information. After extracting the MFCC features, a normalization function is applied to normalize the data. Acoustic features were extracted using the eGeMAPS parameters set (88 features) provided by the openSMILE toolkit.

### B. Model

We employ traditional and deep learning models, SVM, RF, CNN, and ViT, as classifiers

a) *SVM*: uses a kernel function to project training data into feature space and finds a suitable hyperplane to separate the data by the highest margin. We used a linear kernel and regularization parameter (C) of 0.1 for optimal performance.

b) *RF*: is a simple but powerful model comprised of trees. It creates a decision based on the randomly selected data sample, gets a prediction from each tree, and votes for the best solution.

c) *CNN*: has been shown, by extensive research, to be very useful in extracting information from raw signals in various applications. The main benefit of CNN is that it automatically identifies the relevant features without any human supervision. The model comprised three convolution layers with 16, 32, and 64 filters, respectively, and a kernel size of three, followed by an activation and pooling layer after each layer. Finally, the fully connected layer receives the low-level features and creates the high-level abstraction. The classification scores are generated using the ending layer with the softmax activation function.

d) *ViT*: Transformer was originally proposed for natural language processing (NLP), and it has been demonstrated to achieve much better performance than CNNs. It was introduced to computer vision using image patch sequences and attention mechanisms. The Mel-spectrogram of the input audio signal is regarded as an image and converted into patches. The Adam optimizer with a learning rate of 0.0004 and 20 epochs was used.

The SVM and RF models were developed using optimal parameters, that were found more than five times with a grid search. The number of estimators ranged from 50 to 350. For every experiment, we use 5-fold cross-validation

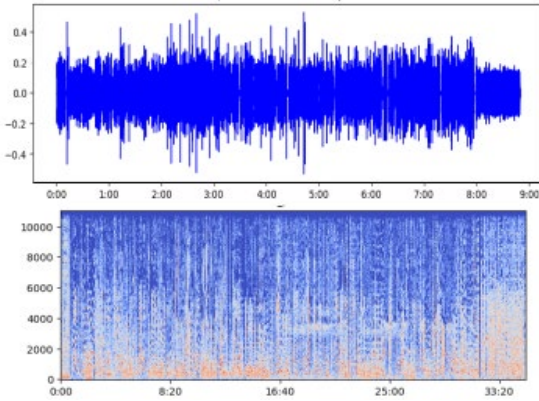


Fig. 1 A sample of wave-plot and spectrogram representation for curiosity emotion.

### III. EXPERIMENTS AND RESULTS

#### A. Dataset

The dataset was obtained from the AI Hub platform [12]. This platform offers a wide range of resources and open

datasets. The data was constructed using children's voices collected from educational broadcast videos (EBS, KBS). They are available in four categories: Preschool Development, Children's Education, Children's Drama, and Entertainment. The data was released in 2023 and is available in Korean with a size of about 862 GB. For the experiments, we used the categories preschool development education and drama (a few were selected) with a total of 994 audio files. The audio utterances contain five emotions: confrontation, curiosity, denial, positive, and neutral. The data was divided into training and testing. Table II shows the number of samples in each set per emotion category.

TABLE II  
DATA STATISTICS

Emotions	Training	Testing
Confrontation (0)	188	21
Curiosity (1)	129	18
Denial (2)	181	21
Neutral (3)	177	20
Positive (4)	215	24
Total	890	104

#### B. Results and Discussions

The input features for the classification models were MFCC coefficients (40 features), 88 acoustic features, and a mel-spectrogram. Table III summarizes the performance of our classifiers in performance metrics: precision, recall, and F1-score of each class. Generally, CNN models outperformed SVM and Random Forest, with SVM showing competitive performance with RF. The results of CNN demonstrate its ability to extract features automatically. This is demonstrated in class 1 (curiosity) by the F1-score increase from 0.0% to 41.38% and 9.52% to 40.00% for acoustic and MFCC features, respectively. The overall accuracy without feature selection for SVM was MFCC = 45.19%, acoustic features = 53.85%, and for RF, the overall accuracy was MFCC = 48.08%, acoustic features = 51.92% respectively. Overall, acoustic features outperform MFCC due to their broader scope in capturing the nuances of speech. While MFCC focuses primarily on spectral characteristics, acoustic features encompass a wider range, including pitch intensity and vocal vibration frequency, which are vital for accurately representing speech patterns. This richer feature set provides a more comprehensive understanding of the nuances of speech, leading to improved model performance.

A total of 44 and 20 features were selected from acoustic and MFCC features. As noted in Table III, the feature selection on MFCC features adds a significant value to the overall accuracy. For the SVM and RF classifiers, the accuracy increased to 50.96% and 52.88% for acoustic and MFCC



TABLE III  
PERFORMANCE WITH/WITHOUT FEATURE SELECTION USING ACOUSTIC AND MFCC FEATURES, WHERE THREE VALUES INDICATE PRECISION, RECALL, AND F1-SCORE.

Model	Label	OpenSMILE						MFCC					
		No FS			Feature Selection (FS)			No FS			Feature Selection (FS)		
		P	R	F1	P	R	F1	P	R	F1	P	R	F1
SVM	0	46.15	28.57	35.29	33.33	42.86	37.50	39.13	42.86	40.91	31.82	33.33	32.56
	1	00.00	00.00	00.00	57.14	22.22	32.00	33.33	05.56	09.52	71.43	27.78	40.00
	2	47.06	76.19	58.18	59.09	61.90	60.47	37.50	42.86	40.00	46.15	57.14	51.06
	3	70.37	95.00	80.85	66.67	80.00	72.73	68.75	55.00	61.11	68.42	65.00	66.67
	4	50.00	62.50	55.56	45.83	45.83	45.83	44.74	70.83	54.84	53.33	66.67	59.26
RF	0	45.45	23.81	31.25	45.45	23.81	31.25	35.29	28.57	31.58	46.15	28.57	35.29
	1	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	60.00	16.67	26.09
	2	51.72	71.43	60.00	50.00	71.43	58.82	38.71	57.14	46.15	41.94	61.90	50.00
	3	61.54	80.00	69.57	59.26	80.00	68.09	65.38	85.00	73.91	71.43	75.00	73.17
	4	48.65	75.00	59.02	45.71	66.67	54.24	51.72	62.50	56.60	52.94	75.00	62.07
CNN	0	35.29	57.14	43.64	-			33.33	47.62	39.28	-		
	1	54.55	33.33	41.38	-			50.00	33.33	40.00	-		
	2	68.42	61.90	65.00	-			54.55	57.14	55.81	-		
	3	75.00	90.00	81.82	-			64.29	45.00	52.94	-		
	4	50.00	39.33	40.00	-			50.00	54.17	52.00	-		

features, respectively. However, we noticed a drop of accuracy

TABLE IV  
RESULTS ON MEL-SPECTROGRAM FEATURES

Label	ViT			CNN		
	P	R	F1	P	R	F1
0	94.74	85.71	90.00	87.50	66.67	75.68
1	47.06	44.44	45.71	64.29	50.00	56.25
2	56.52	61.90	59.09	44.83	61.90	52.00
3	95.24	1.00	97.56	86.36	95.00	90.48
4	1.00	1.00	1.00	1.00	95.83	97.87

from 53.85% to 50.96% for SVM and 51.92% to 50.00% for RF in acoustic features. The degradation could be attributed by the removal of potentially relevant information during the selection process. Feature selection aims to enhance model efficiency by reducing dimensionality and eliminating redundant or noisy features. But, in some cases, this process may inadvertently discard crucial information, leading to a slight decrease in accuracy.

On the other side, in terms of spectral features, the ViT performed the classification task best, with an accuracy of 79.81% while CNN achieved an accuracy of 75.00%. Table IV shows the results on mel-spectrogram features. The attention mechanism helps to weigh and ensure concentration on the important features, which leads to good overall accuracy. This ensures that the prediction does not happen randomly but rather through understanding the pattern in the data. Although the number of samples used for training was small, the performance indicates that the model possesses a high level of capability with a self-attention mechanism. Generally, the experimental results show that both features can perform well under different conditions.

#### IV. CONCLUSIONS

SER is a complex task that includes the detection of feelings conveyed in voice data. This study classified speech emotions using spectral and acoustic features. Two traditional machine learning and deep learning-based classifiers were employed for the classification task. Moreover, we used the spectral image with ViT and CNN to classify five speech emotions. Our experiment results show that the proposed approach performs

well under different conditions. In the future, we intend to improve the accuracy of the automatic SER by using feature combinations and extending the corpus of children's emotional speech to be able to train deep neural networks of different architectures; and also by using multiple modalities such as facial expressions, body movements, etc.

#### ACKNOWLEDGMENT

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2023-00218176).

#### REFERENCES

- [1] L. Shu, J. Xie, M. Yang, Z. Li, Z. Li, D. Liao, X. Xu, and X. Yang, "A review of emotion recognition using physiological signals," *Sensors*, vol. 18, no. 7, p. 2074, 2018.
- [2] A. S. Alluhaidan, O. Saidani, R. Jahangir, M. A. Nauman, and O. S. Neffati, "Speech emotion recognition through hybrid features and convolutional neural network," *Applied Sciences*, vol. 13, no. 8, p. 4750, 2023.
- [3] T. Liu and X. Yuan, "Paralinguistic and spectral feature extraction for speech emotion classification using machine learning techniques," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2023, no. 1, p. 23, 2023.
- [4] S. Madanian, T. Chen, O. Adeleye, J. M. Templeton, C. Poellabauer, D. Parry, and S. L. Schneider, "Speech emotion recognition using machine learning—a systematic review," *Intelligent systems with applications*, p. 200266, 2023.
- [5] A. Milton, S. S. Roy, and S. T. Selvi, "Svm scheme for speech emotion recognition using mfcc feature," *International Journal of Computer Applications*, vol. 69, no. 9, 2013.
- [6] M. Siniith, E. Aswathi, T. Deepa, C. Shameema, and S. Rajan, "Emotion recognition from audio signals using support vector machine," in *2015 IEEE recent advances in intelligent computational systems (RAICS)*. IEEE, 2015, pp. 139–144.
- [7] R. Jahangir, Y. W. Teh, F. Hanif, and G. Mujtaba, "Deep learning approaches for speech emotion recognition: State of the art and research challenges," *Multimedia Tools and Applications*, pp. 1–68, 2021.
- [8] D. Issa, M. F. Demirci, and A. Yazici, "Speech emotion recognition with deep convolutional neural networks," *Biomedical Signal Processing and Control*, vol. 59, p. 101894, 2020.
- [9] S. K. Pandey, H. S. Shekhawat, and S. M. Prasanna, "Deep learning techniques for speech emotion recognition: A review," in *2019 29th International Conference Radioelektronika (RADIOELEKTRONIKA)*. IEEE, 2019, pp. 1–6.
- [10] J. Singh, L. B. Saheer, and O. Faust, "Speech emotion recognition using attention model," *International Journal of Environmental Research and Public Health*, vol. 20, no. 6, p. 5140, 2023.
- [11] C. A. Kumar, A. D. Maharana, S. M. Krishnan, S. S. S. Hanuma, G. J. Lal, and V. Ravi, "Speech emotion recognition using cnn-lstm and vision transformer," in *International Conference on Innovations in Bio-Inspired Computing and Applications*. Springer, 2022, pp. 86–97.
- [12] AI-Hub, "Ai-hub," 2024. [Online]. Available: <https://www.aihub.or.kr/>

# YOLOv8-ND: New Detection Algorithm for Lightweight and Fast Object Detection

Neunggyu Han<sup>1\*</sup>, Seungmin Rho<sup>2</sup>, Yunyoung Nam<sup>3</sup>

<sup>1</sup>*Department of ICT Convergence, Soonchunhyang University, Asan 31538, Korea*

<sup>2</sup>*Department of Industrial Security, Chung-Ang University, Seoul, South Korea*

<sup>3</sup>*Department of Computer Science and Engineering, Soonchunhyang University, Asan 31538, Korea*

\*Contact: az0422@naver.com

**Abstract**—Recent deep learning algorithms are conducting a lot of research to optimize the structure and make it lightweight. Existing deep learning algorithms have been studied with the goal of running in high-performance computer environments, but are currently aimed at running in low-power computer environments such as smartphones. Therefore, it is very important to develop lightweight algorithms while maintaining high performance. A representative example in the field of deep learning algorithms is the object detection algorithm. And among object detection algorithms, there is the YOLO (You Look Only Once) algorithm. This algorithm was created to improve the slow operation speed of the existing R-CNN. The latest version of YOLOv8 can be seen to be very fast and highly accurate compared to the early version. However, there is one problem with this algorithm as well. It is a structure of layers containing algorithms for detecting objects. Although this structure is simple, it has the problem of low efficiency. In particular, there are many parameters, so it requires a lot of computing resources. In this paper, the goal is to improve the object detection algorithm and construct a faster model with fewer parameters. One way is to change the structure of the layer containing the object detection algorithm. While the existing structure was divided into two parts for classification and coordinate inference, this paper changes it to a simpler structure. As a result, the object detection layer in this paper has about 2.37 times fewer parameters than before, and the overall model was able to achieve an inference speed improvement of up to 20% with 1.19 times fewer parameters. However, the inference accuracy is lowered by about 15% compared to the previous version, so additional research in this area appears to be necessary.

## I. INTRODUCTION

Recent deep learning algorithms are conducting a lot of research to optimize the structure and make it lightweight. Existing deep learning algorithms have been studied with the goal of running in high-performance computer environments, but are currently aimed at running in low-power computer environments such as smartphones. An example of this is Samsung's smartphone Galaxy S24, released in January 2024 [1]. This smartphone provides a phone function that allows real-time interpretation using a deep learning algorithm and a picture editing function using a deep learning algorithm. As such, numerous studies are being conducted to ensure that deep learning algorithms can be widely used in everyday life, rather than being the exclusive domain of high-performance

computers. However, many studies are still being conducted with high-performance computer environments as the main target, rather than low-power environments such as smartphones. Therefore, there are still many difficulties in applying it directly in an environment such as a smartphone.

There are various fields of deep learning algorithms. Among them, a widely known one is the object detection algorithm. An object detection algorithm is an algorithm that finds and displays objects in an image. The most representative of these algorithms is YOLO (You Only Look Once) [2-9]. This algorithm is designed to detect objects at a faster rate by simplifying the inference process of R-CNN[10-12], an existing object detection algorithm. Therefore, it is often used as an algorithm to detect objects in real time. There are various versions of this YOLO algorithm. The first YOLO [2] has a low accuracy compared to R-CNN, but its representative feature is that it has a very fast inference speed. In YOLOv2[3], anchor box[12] was introduced to achieve higher accuracy than the previous version. This anchor box is a pre-calculated value for the object's coordinates, and the object's coordinates are calculated based on this value. Except for the latest version, YOLOv8[9], research has been conducted to ensure that the remaining versions[4-8] are faster and have higher accuracy than before. In the latest version, YOLOv8, the anchor box introduced in YOLOv2 was removed to enable smoother object recognition. Additionally, faster inference speed can be achieved by improving the structure. As such, much research has been conducted on the YOLO algorithm, and a fast and accurate algorithm has been developed. However, problems still exist even in the latest version, YOLOv8. After the process of extracting features from the image was completed, inefficiencies were discovered in the structure of the layer that classifies the coordinates and classes of objects. The structure of this layer is a structure that derives results by dividing it into parts that classify objects and infer coordinates. In addition, although this structure is quite simple, there are about 2 million parameters, accounting for 18% of the total parameters of the YOLOv8s standard model.

In this paper, we propose a new object detection layer. The structure of this layer is simpler than before. Additionally, the object detection layer has 2.37 times fewer parameters.

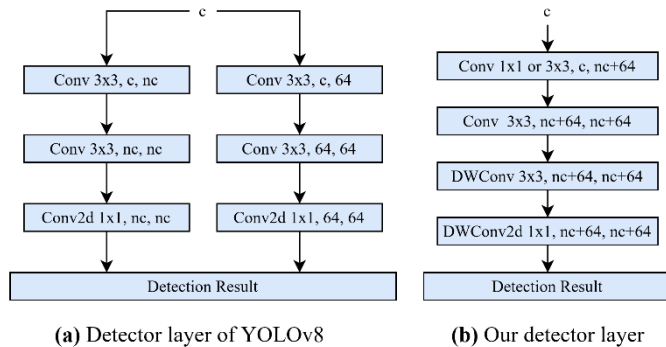
Additionally, when this layer is applied to the YOLOv8 model, the parameters of the entire model can be reduced by 19%. By using such a layer, the inference speed can be improved by up to 20% compared to the existing one.

## II. ARCHITECTURE

### A. Lightweight Detector Layer

The structure of the existing YOLOv8 object detection layer is the same as (a) in Figure 1. This structure consists of a part that infers the class of an object and a part that infers the class are divided into two parts to derive results. Thanks to this structure, YOLOv8 was able to infer the coordinates and classes of objects at high speed. But there is a problem here. The input layer is divided into two parts with a 3x3 convolutional layer. When changing the number of channels using a 3x3 convolutional layer, there is an advantage in that more diverse features can be summarized and extracted. However, due to the nature of the 3x3 convolutional layer, there is a problem that it requires many parameters. In particular, if the number of channels in the input layer is greater than that in the output layer, more parameters are required. In this respect, there is a problem that it inevitably consumes a lot of computing resources.

Therefore, in this paper, the structure was changed to (b) in Figure 1 to further simplify it. This structure is structured so that classification and coordinate inference are carried out simultaneously rather than separately. In addition, if the number of channels for the input layer is more than (number of classes + 64), a 1x1 convolutional layer is used, and otherwise, a 3x3 convolutional layer is used. Lastly, the output layer has bias values set for coordinate inference for 64 channels, and bias values for class inference for the remaining channels. Therefore, when outputting with a general convolutional layer, the loss value calculation is very complicated, so there is a problem that the speed is very slow. Therefore, the problem was solved using a depth wise convolutional layer. With this, the parameters for the YOLOv8s standard object detection layer could be reduced by 2.37 times compared to before.



**Figure 1** Diagram of the structure of the object detection layer. The difference between the existing method and the method of this paper is explained.

## III. EXPERIMENTS AND RESULTS

### A. Environmental Setup

The experimental environment is shown in Table 1. In this environment, the model is trained and the inference speed is calculated. First, the MS-COCO [13] dataset is used as the dataset to learn the model. This dataset is a general-purpose dataset consisting of 80 classes. Additionally, existing YOLO versions provide pre-trained weights using this dataset, so these weights and inference performance are compared. Additionally, hyper-parameters including epoch and batch for learning are set to the same settings as the pre-trained weights of YOLOv8.

**Table 1.** Experimental Environment

Experimental Environment	
CPU	Intel Xeon 4216 x2
RAM	192GB
GPU	RTX A5000 x3
OS	Ubuntu 22.04

### B. Performance Evaluation

The results of the performance evaluation are shown in Table 2. This includes evaluation of not only the existing YOLOv8 models, but also YOLOv5[6], YOLOv6[7], and YOLOv7[8]. Among the evaluation criteria, inference speed includes batch 1 and batch 32. Among these, batch 1 is a value assuming that an object is detected with a single camera. And batch 32 is a value assuming that objects are detected with multiple cameras. Here, the inference speed was measured a total of 10 times, and the result was averaged after deleting the maximum and minimum values one by one.

In this result, improvement refers to the amount of improvement compared to the existing equivalent YOLOv8. If this figure is better than before, it is marked in black, and if it is not, it is marked in red.

In these results, you can see that most models are about 20% faster than before in the batch 1 experiment. However, the mAP50-95 performance indicator showed that performance decreased by up to 10% compared to before. Additionally, when applied to YOLOv8x, it was confirmed that it had similar accuracy to the existing YOLOv8m but was slightly slower. Additionally, when applied to YOLOv8l, performance was seen to be lower and slower than YOLOv7. From this, we can confirm that it is not suitable when the scale of the model is large.

**Table 2.** Model Performance Evaluation

Model	Size	FLOPs (G)	Params (M)	mAP50-95 (%)	Inference Speed (ms; per image)	
					batch 1	batch 32
YOLOv5n	640	4.5	1.9	27.7	7.5	1.4
YOLOv6n	640	11.4	4.7	37.6	10.0	1.4
YOLOv8n	640	8.7	3.2	37.5	8.9	1.5
YOLOv8NDn	640	9.9	3.1	34.2	7.2	1.5
Improvement	-	+1.2	-0.1	-3.3	-1.7	=
YOLOv5s	640	16.4	7.2	37.1	7.6	1.5
YOLOv6s	640	45.3	18.5	45.0	10.9	2.0
YOLOv7-tiny	640	13.7	6.2	36.0	6.6	1.3
YOLOv8s	640	28.6	11.2	44.7	9.1	2.1
YOLOv8NDs	640	25.8	9.8	40.8	7.8	2.0
Improvement	-	-2.8	-1.4	-3.9	-1.3	-0.1
YOLOv5m	640	48.9	21.2	45.4	9.9	3.0
YOLOv6m	640	85.8	34.9	50.0	20.6	4.1
YOLOv7	640	104.5	36.9	49.7	10.0	3.6
YOLOv8m	640	78.9	25.9	50.1	11.3	4.0
YOLOv8NDm	640	66.6	22.8	45.7	9.3	3.5
Improvement	-	-12.3	-3.1	-4.4	-2.0	-0.5
YOLOv5l	640	109.0	46.5	48.7	11.5	5.0
YOLOv6l	640	150.7	59.6	52.8	21.3	6.3
YOLOv8l	640	165.2	43.7	52.9	13.4	6.2
YOLOv8NDl	640	142.8	38.8	48.4	10.7	5.7
Improvement	-	-22.4	-4.9	-4.5	-2.7	-0.5
YOLOv5x	640	205.5	86.7	50.3	14.0	8.5
YOLOv7x	640	189.7	71.3	51.5	11.1	5.4
YOLOv8x	640	257.8	68.2	54.0	13.7	9.6
YOLOv8NDx	640	221.0	60.2	49.7	12.3	8.7
Improvement	-	-36.8	-8.0	-5.3	-1.4	-0.9

#### IV. CONCLUSIONS

In this paper, research was conducted to improve the object detection layer of the existing YOLOv8. The existing object detection layer of YOLOv8 was structured to conduct inference by dividing it into two parts for class and coordinate inference. However, inefficiencies were found in this structure, and to solve them, a new object detection layer was proposed in this paper. Unlike before, this object detection layer simultaneously infers the class and coordinates of an object through a single path. This increases the inference speed by up to 20% compared to before. However, inference performance has been reduced by up to 10%. Additionally, when applied to YOLOv8l and YOLOv8x, which have large model scales, the performance was similar to that of other models with small scale models. Therefore, additional research appears to be needed to increase speed while minimizing performance loss.

#### ACKNOWLEDGEMENT

This research was supported by Culture, Sports and Tourism R&D Program through the Korea Creative Content Agency

grant funded by the Ministry of Culture, Sports and Tourism in 2024(Project Name: Development of distribution and management platform technology and human resource development for blockchain-based SW copyright protection, Project Number:RS-2023-00228867, Contribution Rate: 100%)

#### REFERENCES

- [1] Enter the New Era of Mobile AI With Samsung Galaxy S24 Series, <https://news.samsung.com/global/enter-the-new-era-of-mobile-ai-with-samsung-galaxy-s24-series> (accessed on 2024.02.13)
- [2] Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 779-788).
- [3] Redmon, J., & Farhadi, A. (2017). YOLO9000: better, faster, stronger. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 7263-7271).
- [4] Redmon, J., & Farhadi, A. (2018). YOLOv3: An incremental improvement. arXiv preprint arXiv:1804.02767.
- [5] Bochkovskiy, A., Wang, C. Y., & Liao, H. Y. M. (2020). YOLOv4: Optimal speed and accuracy of object detection. arXiv preprint arXiv:2004.10934.
- [6] Jocher, G. (2020). YOLOv5 by Ultralytics (Version 7.0) [Computer software]. <https://doi.org/10.5281/zenodo.3908559>

- [7] Li, C., Li, L., Jiang, H., Weng, K., Geng, Y., Li, L., ... & Wei, X. (2022). YOLOv6: A single-stage object detection framework for industrial applications. arXiv preprint arXiv:2209.02976.
- [8] Wang, C. Y., Bochkovskiy, A., & Liao, H. Y. M. (2023). YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 7464-7475).
- [9] Jocher, G., Chaurasia, A., & Qiu, J. (2023). Ultralytics YOLO (Version 8.0.0) [Computer software]. <https://github.com/ultralytics/ultralytics>
- [10] Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 580-587).
- [11] Girshick, R. (2015). Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision* (pp. 1440-1448).
- [12] Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28
- [13] Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., ... & Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13* (pp. 740-755). Springer International Publishing.

# Exploring Deep Learning for Emotion Classification in Avatar-Generated Image Dataset Collected via Meta Quest Pro

Ahsan Aziz<sup>1</sup>, Chomyong Kim<sup>2</sup>, Yunyoung Nam<sup>3</sup>

<sup>1</sup>Department of ICT Convergence, Soonchunhyang University, Asan, 31538, Republic of Korea

<sup>2</sup>ICT Convergence Research Centre, Soonchunhyang University, Asan, South Korea

<sup>3</sup>Department of Computer Science and Engineering, Soonchunhyang University, Asan, South Korea

\*Contact: [ynam@sch.ac.kr](mailto:ynam@sch.ac.kr)

**Abstract**— Facial emotion recognition is crucial in numerous domains such as human-computer interaction and affective computing. This research introduces an advanced method for facial emotion recognition, presenting an innovative approach to this field. The first stage of our investigation included gathering emotional data through a Virtual Reality (VR) device. We engaged four participants in collecting this data, instructing them to express predetermined emotions while viewing a series of sample images. These samples covered four specific emotional categories: happiness, anger, neutrality, and surprise.

This research delves into the domain of emotion classification through the analysis of a unique dataset composed of avatar-generated images obtained using Meta Quest Pro. The dataset is characterized by five distinct emotional classes, each containing an extensive set of approximately 3000 images. These images, portraying avatars crafted by the participants, represent a rich source of diverse emotional expressions. Our investigation revolves around the development and training of deep learning models tailored for the precise classification of emotions encapsulated within these images. The primary aim is to attain a heightened level of accuracy in detecting and categorizing emotional states expressed by the avatars. By leveraging deep learning methodologies, this study contributes valuable insights into the efficacy of such models in effectively capturing and classifying a wide spectrum of emotions present in the dataset. The outcomes of this research hold significance in advancing the realms of affective computing and human-computer interaction, paving the way for enhanced emotional understanding and interpretation in virtual environments.

## I. INTRODUCTION

In a world progressively dominated by digitalization, the convergence of technology and human emotion presents a promising frontier with potential applications in various domains like healthcare, education, entertainment, and human-computer interaction. Among the various paths of investigation, facial emotion recognition emerges as a powerful tool for understanding and enhancing human experiences. The capability to interpret and react to human emotions is of great importance, not only for the development of more empathetic and responsive technology but also due to its profound implications in fields such as psychology, marketing, and artificial intelligence.

Traditional techniques for facial emotion recognition mainly depend on examining images and videos, extracting information from two-dimensional depictions of facial

expressions. However, these methods have inherent constraints in capturing the nuances and subtleties of human emotions, which frequently manifest in the three-dimensional dynamics of facial features. This is where Virtual Reality (VR) becomes crucial—a technology that has transformed how we interact with digital content and, more significantly, provides a unique avenue for observing and analyzing emotions in an immersive and authentic context. In [1] the study explores the effectiveness of virtual reality (VR) in evaluating and enhancing emotion recognition within social contexts. The research employs realistic and dynamic stimuli, conducting a comparative analysis of three emotion recognition tasks: VR, video, and photo tasks. Utilizing a Virtual Reality device with facial point tracking technology and making a whole avatar as per given points in conjunction with machine learning capabilities, the paper delves into the captivating realm of facial emotion recognition. By surpassing the limitations of flat imagery and videos, this innovative approach opens up new dimensions in the interpretation, comprehension, and responsive handling of human emotions. The introduction sets the stage for the exploratory journey ahead, outlining the foundational concepts and driving forces that underlie this research.

In [2] the author recognizes the unique characteristics of data collection in virtual reality (VR) environments, especially through head-mounted displays, in comparison to conventional classroom or online learning settings. This highlights the need to develop a recognition approach specifically tailored to VR contexts, and the author introduces a novel method for identifying learning concentration within VR environments. By immersing users in computer-generated settings, VR surpasses the constraints of traditional screens, providing a sense of presence and embodiment that was once confined to the realm of science fiction. With the emergence of increasingly advanced VR devices, equipped with features like positional tracking, haptic feedback mechanisms, and precise motion capture, the potential applications have expanded exponentially. In [3] the paper describes the creation of an inventive VR-based system designed for presenting facial emotional expressions. This system not only enables the monitoring of emotional expressions but also tracks eye gaze and physiological signals connected to the identification of

emotions, introducing more efficient therapeutic methodologies. One particularly promising avenue within the VR spectrum lies in its capacity to mirror and manipulate human emotions. By furnishing a platform capable of replicating real-world scenarios and interpersonal interactions, VR provides an ideal environment for eliciting and scrutinizing emotions under controlled and customizable conditions. This paper navigates through the utilization of VR as a conduit for capturing the intricate choreography of facial expressions in three dimensions, thereby amplifying our ability to identify and appreciate emotions more fully.

The process of facial emotion recognition is a complex endeavor that heavily relies on analyzing facial expressions, which are composed of various facial landmarks such as the positions of the eyes, eyebrows, nose, and mouth. These landmarks play a crucial role in conveying emotions and subtle cues, which often require a high degree of precision for accurate detection.

In recent years, advancements in the domains of computer vision and machine learning have staged a revolution in the field of facial emotion recognition. Machine learning algorithms, particularly the formidable deep learning models, have exhibited remarkable prowess in autonomously extracting features from facial data, thus enabling the accurate recognition of emotions. This paper delves into the methodologies and techniques employed to harness the power of machine learning in deciphering the wealth of data gleaned from VR devices and Avatars.

In this proposal, we present a novel approach to facial emotion recognition the initial phase of our research involved the acquisition of emotional data utilizing a Virtual Reality (VR) device. We enlisted four participants to contribute to this data collection process, tasking them with displaying predetermined emotions while exposed to a set of sample images. These sample images encompassed five distinct emotional classes, including happiness, anger, neutrality, surprise, and fear.

The paper contributes a pioneering approach to facial emotion recognition, leveraging Virtual Reality (VR) technology for data acquisition and analysis. Through the utilization of a unique dataset comprising avatar-generated images obtained via Meta Quest Pro, the study encompasses five distinct emotional classes: happiness, anger, neutrality, surprise, and fear. Notably, the inclusion of fear expands the emotional spectrum, providing a rich source of diverse expressions for analysis. This novel methodology marks a significant contribution to the field, offering a fresh perspective on data collection and classification within affective computing.

Central to the research is the development and training of deep learning models tailored specifically for emotion classification. By employing deep learning methodologies, the study aims to achieve a heightened level of accuracy in detecting and categorizing emotional states depicted by the avatars. The exceptional performance of the machine learning model, particularly the cubic Support Vector Machine (SVM) with DenseNet-201, underscores the robustness and efficacy of the proposed approach. Achieving a remarkable accuracy rate of 99.7% when applied to the VR dataset, these results highlight the potential for practical applications across various domains, particularly in advancing human-computer interaction through more intuitive and empathetic interfaces.

In essence, this research contributes significantly to the realms of affective computing and human-computer interaction by enhancing emotional understanding and interpretation in virtual environments. By demonstrating the feasibility and effectiveness of deep learning models in capturing and classifying a wide spectrum of emotions, the study opens avenues for further advancements in technology-driven emotional recognition systems. The outcomes hold promise for the development of more immersive and responsive human-computer interfaces, ultimately leading to enhanced user experiences and interactions in virtual spaces.

## II. RELATED WORK

This research [1] is to investigate the utility of virtual reality (VR) in the assessment and training of emotion recognition within social contexts, utilizing realistic and dynamic stimuli and involves a comparative analysis of three emotion recognition tasks: VR, video, and photo tasks. In [1] 100 healthy participants completed all three emotion recognition tasks and during the VR task, participants evaluated emotions expressed by virtual characters (avatars) in a VR urban setting, while their eye movements were tracked. The overall recognition accuracy stands at 75%, aligning closely with the accuracy observed in the photo and video tasks. Nevertheless, distinctions emerge in recognizing specific emotions; VR performs less effectively in identifying disgust and happiness but outperforms the video task in recognizing surprise and anger. Participants allocate varying amounts of time to different emotions during the VR task. Notably, disgust, fear, and sadness receive more attention compared to surprise and happiness. Additionally, participants exhibit a preference for focusing on the eyes and nose regions of the avatars rather than their mouths [1].

In [2] the author acknowledges the distinctive nature of data collection within virtual reality (VR) environments, particularly through head-mounted displays, when contrasted with traditional classroom or online learning settings. This underscores the necessity for developing a recognition approach tailored specifically to VR contexts and they have a novel method for recognizing learning concentration within VR environments. This innovative approach hinges on the integration of multi-modal features, incorporating data derived from learner interactions (e.g., interactive assessments, text interactions, clickstream data) and visual cues (e.g., pupil facial expressions and eye gaze). These combined features facilitate the evaluation of learners' concentration from cognitive, emotional, and behavioural perspectives and the study of this research reveals a positive association between heightened levels of concentration and enhanced learning achievements. Additionally, it highlights the significant role played by learners' perceived sense of immersion within the VR environment in shaping their level of concentration.

In [19] EVOKE (Emotion enabled Virtual avatar mapping using Optimized KnowledgeE distillation), a lightweight emotion recognition framework tailored for seamless integration into 3D avatars within virtual environments. EVOKE addresses the rising demand for immersive and emotionally engaging experiences in virtual environments. Leveraging knowledge distillation on the DEAP dataset, which encompasses valence, arousal, and dominance as primary emotional classes, the framework achieves



competitive results with reduced computational resources. Remarkably, the distilled model, a CNN with only two convolutional layers and significantly fewer parameters than the teacher model achieves an accuracy of 87%. This balance between performance and deployment ability positions EVOKE as an optimal choice for virtual environment systems. Moreover, the multi-label classification outcomes are utilized to map emotions onto custom-designed 3D avatars, further enhancing the immersive experience for users.

### III. PROPOSED METHOD

In this section, we will discuss the proposed methodology of facial emotion recognition using the Avatar-generated dataset gathered from VR devices. The below figure is the proposed architecture.

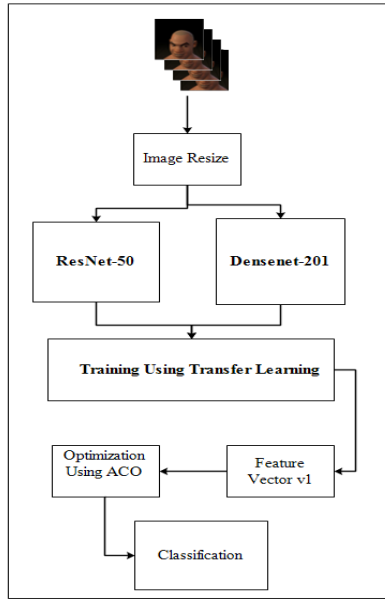


Figure 1 The Proposed Flow Diagram

#### A. Data Pre-processing and Standardization

In the initial phase of our study, the dataset for emotion recognition was acquired in the form of avatar facial expressions captured through videos. To facilitate further analysis and model training, we performed a crucial data preprocessing step by converting these videos into individual frames or images. This transformation allowed us to extract and isolate distinct facial expressions over time, providing a granular view of the emotional states exhibited by the avatars.

To enhance computational efficiency and standardize the dataset, we resized the obtained images. This resizing step ensures consistency in the dimensions of all facial frames, minimizing computational complexity during subsequent processing stages. By standardizing the image sizes, we aimed to create a uniform and optimized dataset that can be effectively utilized for training and evaluating emotion recognition models.

The meticulous process of converting videos into frames and resizing images not only facilitates the practical application of deep learning techniques but also ensures the preservation of essential facial features critical for accurate emotion classification. This preprocessing methodology lays a robust foundation for the subsequent stages of our research, allowing for a comprehensive exploration of emotion recognition in avatar-generated facial expressions.

#### B. Deep Learning (DL)

Deep Learning (DL) stands as a formidable subset of machine learning, involving the creation and training of artificial neural networks for intricate tasks. These networks, with their hierarchical architecture, autonomously learn and represent features from raw data. Their capacity for automatic feature extraction makes them highly effective in various domains, including image recognition, natural language processing, and emotion classification. DL models exhibit the unique ability to progressively capture complex patterns within the data.

ResNet, or Residual Network, stands out as a transformative deep learning architecture renowned for overcoming challenges in training very deep neural networks. Introduced by Kaiming He et al. in 2015 [17], ResNet revolutionized the field by introducing skip connections or residual blocks. These connections allow the network to skip one or more layers during forward and backward propagation, addressing the vanishing gradient problem and enabling the training of significantly deeper networks. In our research, we harnessed the power of ResNet-50, a variant of ResNet, as the foundational architecture for our emotion recognition task.

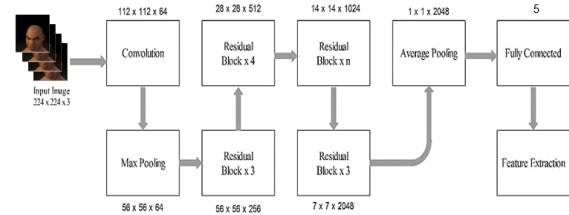


Figure 2 Modified ResNet-50

DenseNet, or Densely Connected Convolutional Network, represents another influential deep learning architecture designed to tackle the limitations of traditional convolutional neural networks (CNNs). Devised by Gao Huang et al. in 2017 [18], DenseNet optimizes information flow and feature reuse by connecting each layer to every other layer in a feed-forward fashion. This dense connectivity enhances feature propagation and mitigates the vanishing gradient problem. In our investigation, we opted for DenseNet-201, a variant of DenseNet, to assess its effectiveness in the realm of emotion recognition.

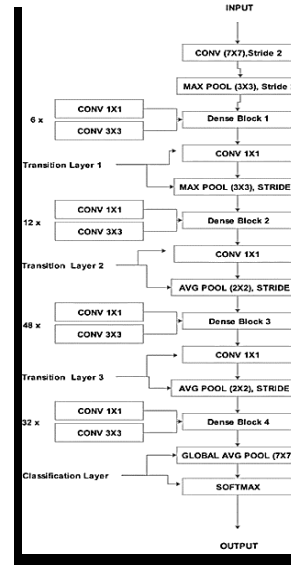


Figure 3 Modified DenseNet-201

Within both ResNet-50 and DenseNet-201 architectures, we introduced average pooling layers after the models. The addition of these average pooling layers serves to down-sample the spatial dimensions of the feature maps produced by the preceding layers. This modification not only alleviates the overall computational burden but also fosters a more concise representation of the learned features. By incorporating average pooling layers, we intend to enhance the efficiency and interpretability of the deep learning models deployed for emotion recognition within the specific context of avatar-generated facial expressions.

### C. Transfer Learning (TL)

In our pursuit of refining the model's performance for emotion recognition in avatar-generated facial expressions, we leveraged the powerful technique of transfer learning. Transfer learning involves pre-training a deep neural network on a large dataset and then fine-tuning it on a smaller, task-specific dataset. This approach capitalizes on the knowledge gained during pre-training, enabling the model to generalize better to the target task with a limited amount of task-specific data. In our study, we adopted a transfer learning strategy with a pre-trained model, specifically ResNet-50 and DenseNet-201, which had been trained on a vast dataset for image classification. By fine-tuning these models on our emotion recognition dataset comprising five distinct classes, we aimed to capitalize on the learned features from the extensive pre-training phase. This transfer learning methodology not only expedited the training process but also facilitated the adaptation of the model to our specific emotion recognition task, demonstrating its efficacy in handling the nuances and complexities inherent in the diverse emotional expressions depicted in avatar-generated facial images and the below figure 4 shows the modification of transfer learning using five classes.

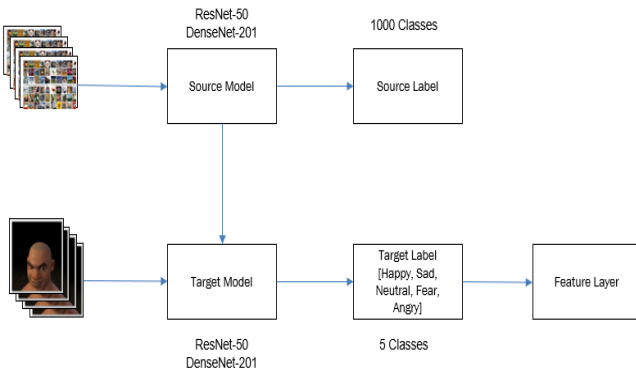


Figure 4 Modified Transfer Learning

### D. Machine Learning (ML)

With the standardized and categorized dataset in hand, we proceeded to employ various machine learning algorithms, with a predominant focus on Support Vector Machines (SVM). SVMs, renowned for their versatility and proficiency in classification tasks, were instrumental in the context of emotion recognition. These algorithms were trained on the pre-processed data, leveraging the extracted features to make predictions regarding the emotional states expressed by the subjects.

## IV. RESULTS AND DISCUSSION

### A. Dataset

To obtain a comprehensive understanding of human emotions in immersive contexts, we leveraged VR technology as a potent tool for data collection. Our study involved Four subjects who actively participated in emotion elicitation sessions within the VR environment. Each subject was presented with a set of sample images, meticulously designed to evoke five distinct emotional states: happiness, anger, neutrality, surprise, and fear. The VR device captured their facial expressions and responses as they viewed and reacted to these visual stimuli. Below Figure 5 is an explanation of all the Avatars that we have used to gather the dataset.



Figure 5 Avatar Emotions

The following points are kept in mind while collecting the Avatar emotion dataset which are as follow:

- Telling the subject about the emotions.
- Visualizing the subject, the images that the subject will see and act according to.
- After wearing the VR subject should make a face according to an image.
- Start recording the Avatar while the subject is in the current emotional state. And stop it while the subject is still in the same emotional act state.
- And continue this for several images for the same emotion and then save those Avatar videos at the end.
- Then do the same procedure for the rest of the Five emotions.
- No garbage value would be recorded if we follow step four correctly.

Below Figure 6 is the proposed data collection flow chart using the VR device.

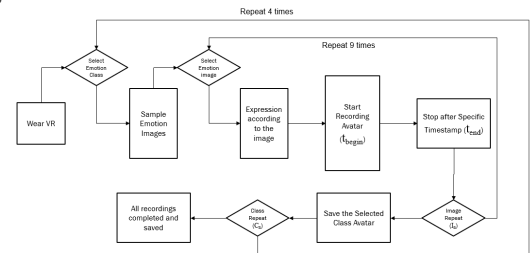


Figure 6: Proposed Data Collection Flow Chart

This paper introduces several significant contributions to the field of facial expression analysis. One of the primary advancements is the creation and utilization of an extensive dataset capturing precise facial points and displaying the avatar accordingly. This dataset comprises a wide array of facial action units, encompassing various facial muscles, each with its unique roles in shaping facial expressions. To

our knowledge, there is no prior research that provides such a comprehensive and detailed Avatar facial dataset. The provision of this dataset significantly enriches our comprehension of facial expressions and fosters the exploration of the intricate realm of human emotions and expressions. The dataset is unprecedented in its depth and granularity, opening new avenues for studying and understanding human emotions. It not only deepens our knowledge of facial expressions but also provides opportunities for a more nuanced and accurate analysis of emotional states.

### B. Experimental results

In our experimental trials, we attained remarkable levels of accuracy when employing our machine-learning model on the dataset obtained from the VR device. This substantial accuracy underscores the model's proficiency in effectively discerning and categorizing emotions using facial key points of avatar as features. These outcomes portend favourable opportunities for the utilization of this approach in diverse applications, notably within the domain of human-computer interaction. Cross-validation is a pivotal method in the realm of machine learning, facilitating a robust appraisal of model effectiveness and ensuring its ability to generalize to unseen data. In our research, we applied 10-fold cross-validation to ascertain the reliability of our dataset. This technique entails partitioning the dataset into ten equally sized segments, known as 'folds.' In each cycle, one-fold serves as the test set, while the remaining folds constitute the training set. This process iterates ten times, ensuring that each data point undergoes testing exactly once. The results from each fold are then aggregated to offer a comprehensive evaluation of the model's performance, while simultaneously reducing potential biases stemming from a single train-test division. Table 1 below provides a comprehensive summary of the performance metrics for three distinct classifiers employed in our study: Cubic SVM, Quadratic SVM, Cubic KNN on ResNet-50.

Table I Results of the Different Classifiers on ResNet-50

Classifiers	Recall Rate (%)	Precision Rate (%)	F1 Score (%)	AUC (%)	Time (sec)	Accuracy (%)
Cubic SVM	98.0%	98.02%	97.90 %	0.996 %	18.396 %	98.0%
Quadratic SVM	98.56 %	98.58%	98.57 %	0.996 %	14.949 %	<b>98.6%</b>
Cubic KNN	92.58 %	94.58%	93.57 %	0.992 %	23.709 %	92.6%

And Table 2 below provides a comprehensive summary of the performance metrics for three distinct classifiers employed in our study: Cubic SVM, Quadratic SVM, Cubic KNN on DenseNet-201.

Table II Results of the Different Classifiers on DenseNet-201

Classifiers	Recall Rate (%)	Precision Rate (%)	F1 Score (%)	AUC (%)	Time (sec)	Accuracy (%)
Cubic SVM	99.72 %	99.72%	99.72 %	1%	15.085 %	<b>99.7%</b>
Quadratic SVM	99.72 %	99.72%	99.72 %	1%	57.026 %	99.7%
Cubic KNN	92.58 %	94.58%	93.57 %	0.996 %	27.829 %	92.6%

These models were rigorously evaluated across various metrics, including recall rate, precision rate, F1 score, processing time, and accuracy, to assess their efficacy in the context of emotion recognition based on Avatar.

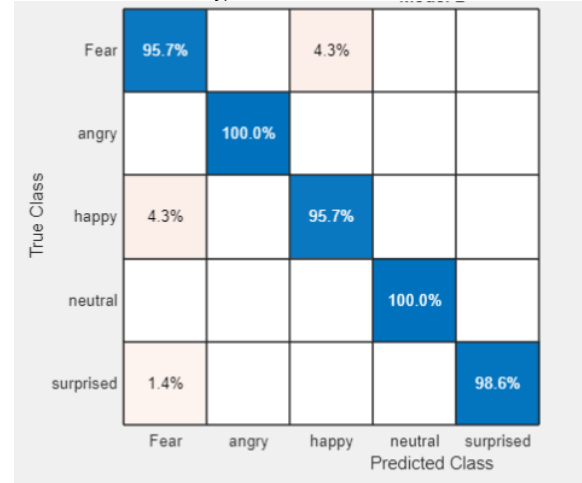


Figure 7 Cubic SVM Recall Rate for ResNet-50

Notably, the cubic SVM classifier emerged as the top-performing model, achieving the highest accuracy among the classifiers tested.

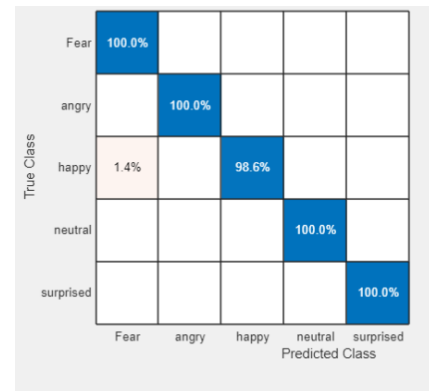


Figure 8 Cubic SVM Recall Rate for DenseNet-201

Furthermore, it is imperative to note that the success of the cubic SVM classifier extends beyond accuracy, as it consistently demonstrated superior performance in multiple metrics, including recall rate and F1 score, signifying its ability to effectively identify and classify emotions. The classifier's efficiency in processing time also underscores its real-world applicability for interactive systems. For a more detailed understanding of the classification performance, we provide the confusion matrix for the recall rate for ResNet-50 in Figure 7 and for DenseNet-201 in Figure 8.

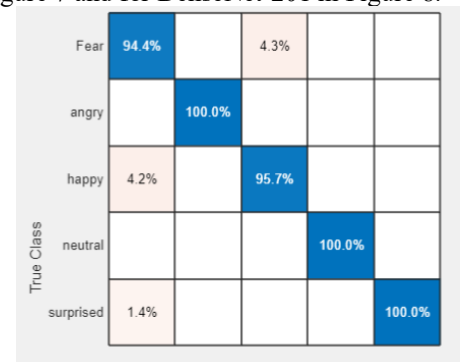


Figure 9 Cubic SVM Precision Rate for ResNet-50

For the precision rate, we have provided in Figure 8 below for ResNet-50, and the precision rate for DenseNet-201 is provided in Figure 10.

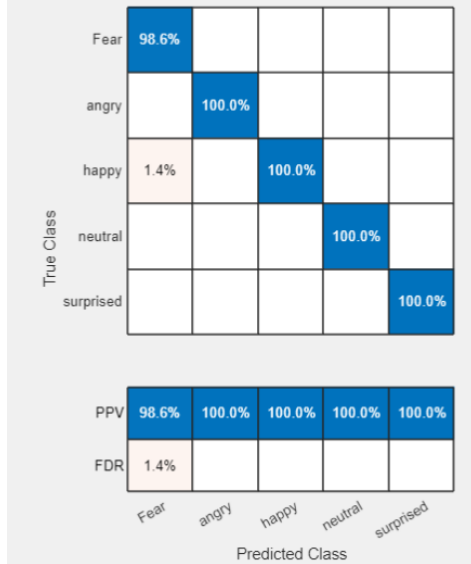


Figure 10 Cubic SVM Precision Rate for DenseNet-201

The confusion matrix offers a comprehensive view of the classifier's true positive, true negative, false positive, and false negative predictions, allowing for a deeper insight into its precision and recall rates. These results collectively emphasize the cubic SVM classifier's suitability for applications necessitating accurate and efficient emotion recognition.

Table III Comparative Analysis of Proposed and Previous Classification System

Reference	Model Technique	F1 Score	Accuracy
EVOKE [19]	CNN	0.88	87.62%
Proposed Methodology	Hybrid CNN-SVM	0.97	98.60%
		0.99	99.70%

The comparison Table III illustrates the performance of two emotion recognition methodologies: EVOKE and the Proposed Methodology. EVOKE, employing a CNN model, achieves an F1 Score of 0.88 and an accuracy of 87.62%. In contrast, the Proposed Methodology combines a hybrid CNN-SVM approach, yielding significantly higher results. Initially, it achieves an impressive F1 Score of 0.97 with an accuracy of 98.60%, showcasing its robustness. Furthermore, the methodology undergoes refinement, resulting in an exceptional F1 Score of 0.99 and an accuracy of 99.70%. These results underscore the superiority of the Proposed Methodology over EVOKE in terms of accuracy and performance, positioning it as a more effective solution for emotion recognition tasks within virtual environments.

## V. CONCLUSION

In this paper, Facial emotion recognition plays an essential role in human-computer interaction, affective computing, and various other domains. In this study, we propose a state-of-the-art approach to facial emotion recognition the initial phase of our research involved the acquisition of emotional data utilizing a Virtual Reality (VR) device. We enlisted four participants to contribute to this data collection process,

tasking them with displaying predetermined emotions while exposed to a set of sample images. These sample images encompassed five distinct emotional classes, including happiness, anger, neutrality, surprise, and fear.

This research introduces several significant contributions to the field of facial expression analysis and delves into the domain of emotion classification through the analysis of a unique dataset composed of avatar-generated images obtained using Meta Quest Pro. The dataset is characterized by five distinct emotional classes, each containing an extensive set of approximately 3000 images. These images, portraying avatars crafted by the participants, represent a rich source of diverse emotional expressions. Our investigation revolves around the development and training of deep learning models tailored for the precise classification of emotions encapsulated within these images. The primary aim is to attain a heightened level of accuracy in detecting and categorizing emotional states expressed by the avatars. By leveraging deep learning methodologies, this study contributes valuable insights into the efficacy of such models in effectively capturing and classifying a wide spectrum of emotions present in the dataset. The outcomes of this research hold significance in advancing the realms of affective computing and human-computer interaction, paving the way for enhanced emotional understanding and interpretation in virtual environments. Our machine learning model, particularly the cubic Support Vector Machine (SVM) with DenseNet-201, demonstrated exceptional performance. We achieved a remarkable accuracy rate of 99.7% when applying this model to the VR dataset for emotion recognition based on Avatar images. This extraordinary level of accuracy underscores the robustness and efficacy of our approach. Such outstanding results hold great promise for the application of our methodology across a broad spectrum of fields, with particular relevance to the advancement of human-computer interaction. Our research then proceeded to employ diverse machine learning algorithms, with a primary focus on Support Vector Machines (SVMs), for emotion recognition within this extensive dataset.

## ACKNOWLEDGMENT

This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the ICAN(ICT Challenge and Advanced Network of HRD) support program(IITP-2024-2020-0-01832) supervised by the IITP(Institute for Information & Communications Technology Planning & Evaluation)

## REFERENCE

- [1] Geraets, C. N. W., S. Klein Tunte, B. P. Lestestuiwer, M. Van Beilen, S. A. Nijman, J. B. C. Marsman, and W. Veling. "Virtual reality facial emotion recognition in social environments: An eye-tracking study." *Internet interventions* 25 (2021): 100432.
- [2] Hu, Renhe, Zihan Hui, Yifan Li, and Jueqi Guan. "Research on Learning Concentration Recognition with Multi-Modal Features in Virtual Reality Environments." *Sustainability* 15, no. 15 (2023): 11606.
- [3] Bekele, Esubalew, Zhi Zheng, Amy Swanson, Julie Crittendon, Zachary Warren, and Nilanjan Sarkar. "Understanding how adolescents with autism respond to facial expressions in virtual reality environments." *IEEE transactions on visualization and computer graphics* 19, no. 4 (2013): 711-720.
- [4] Gutiérrez-Maldonado, José, Mar Rus-Calafell, and Joan González-Conde. "Creation of a new set of dynamic virtual reality faces for the

- assessment and training of facial emotion recognition ability." *Virtual Reality* 18 (2014): 61-71.
- [5] Bisogni, Carmen, Aniello Castiglione, Sanoar Hossain, Fabio Narducci, and Saiyed Umer. "Impact of deep learning approaches on facial expression recognition in healthcare industries." *IEEE Transactions on Industrial Informatics* 18, no. 8 (2022): 5619-5627.
- [6] Tang, Yichuan. "Deep learning using linear support vector machines." arXiv preprint arXiv:1306.0239 (2013).
- [7] Goodfellow, Ian J., Dumitru Erhan, Pierre Luc Carrier, Aaron Courville, Mehdi Mirza, Ben Hamner, Will Cukierski et al. "Challenges in representation learning: A report on three machine
- [8] Bekele, Esubalew, Zhi Zheng, Amy Swanson, Julie Crittendon, Zachary Warren, and Nilanjan Sarkar. "Understanding how adolescents with autism respond to facial expressions in virtual reality environments." *IEEE transactions on visualization and computer graphics* 19, no. 4 (2013): 711-720.
- [9] Sun, Yi, Yuheng Chen, Xiaogang Wang, and Xiaoou Tang. "Deep learning face representation by joint identification-verification." *Advances in neural information processing systems* 27 (2014).
- [10] Ruan, Delian, Yan Yan, Shenqi Lai, Zhenhua Chai, Chunhua
- [11] Shen, and Hanzi Wang. "Feature decomposition and reconstruction learning for effective facial expression recognition." In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7660-7669. 2021.
- [12] Hua, Wentao, Fei Dai, Liya Huang, Jian Xiong, and Guan Gui. "HERO: Human emotions recognition for realizing intelligent Internet of Things." *IEEE Access* 7 (2019): 24321-24332.
- [13] Zhu, Qing, Qirong Mao, Hongjie Jia, Ocquaye Elias Nii Noi, and Juanjuan Tu. "Convolutional relation network for facial expression recognition in the wild with few-shot learning." *Expert Systems with Applications* 189 (2022): 116046.
- [14] Miranda, Catarina Runa, and Verónica Costa Orvalho. "Assessing Facial Expressions in Virtual Reality Environments." In *VISIGRAPP (3: VISAPP)*, pp. 488-499. 2016.
- [15] Zheng, Lim Jia, James Mountstephens, and Jason Teo. "Four-class emotion classification in virtual reality using pupillometry." *Journal of Big Data* 7 (2020): 1-9.
- [16] Dalili, M.N., Penton-Voak, I.S., Harmer, C.J., Munafo, ` M.R., 2015. Meta-analysis of emotion recognition deficits in major depressive disorder. *Psychol. Med.* 45 (6), 1135–1144
- [17] He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Spatial pyramid pooling in deep convolutional networks for visual recognition." *IEEE transactions on pattern analysis and machine intelligence* 37, no. 9 (2015): 1904-1916.
- [18] Huang, Gao, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. "Densely connected convolutional networks." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700-4708. 2017.
- [19] Nadeem, Maryam, Raza Imam, Rouqaiah Al-Refai, Meriem Chkir, Mohamad Hoda, and Abdulmotaleb El Saddik. "EVOKE: Emotion Enabled Virtual Avatar Mapping Using Optimized Knowledge Distillation." In *2024 IEEE International Conference on Consumer Electronics (ICCE)*, pp. 1-6. IEEE, 2024.



# Human Fall Direction Classification using Hybrid Deep Learning and Machine Learning Models

Awais Khan<sup>1</sup>, Jung-Yeon Kim<sup>2</sup> and Yunyoung Nam<sup>3\*</sup>

<sup>1</sup> Department of ICT Convergence, Soonchunhyang University, Asan 31538, Republic of Korea

<sup>2</sup> ICT Convergence Research Center, Soonchunhyang University, Asan 31538, Korea

<sup>3</sup> Emotional and Intelligent Child Care System Convergence Research Center, Soonchunhyang University, Asan 31538, Republic of Korea

\*Corresponding Author: Yunyoung Nam. Email: ynam@sch.ac.kr

**Abstract— the human fall detection system plays a crucial role in assistive technology, providing essential support for a substantial population. Its significance extends to monitoring and alerting seizures and sudden falls in individuals with epilepsy and related conditions. Our proposed methodology involves a systematic four-step process. Initial image extraction from videos is followed by data pre-processing, subsequent data labelling, and, in the fourth step, feature extraction utilizing deep learning models. The culmination of these steps involves classification through a machine learning model. Notably, our approach employs advanced transfer learning techniques to enhance the performance of ResNet-50 pre-trained models. The dataset, sourced from Soonchunhyang University Cheonan Hospital, Asan, demonstrates a commendable average accuracy of 97.9%. This research contributes to the advancement of fall detection technology, addressing critical needs in healthcare and assistive systems.**

## I. INTRODUCTION

The human fall detection system holds significant importance within assistive technology, addressing the essential need for living assistance for a considerable population. The elderly demographic in Bangladesh and Western countries has notably increased in recent years. Fall-related incidents have been identified as a leading cause of severe injuries and fatalities among individuals aged 79 and above [1]. The National Institutes of Health in the United States reported approximately 1.6 million elderly individuals suffering from fall-related injuries annually [2]. China, experiencing the fastest aging population globally, is projected to reach around 35% by 2050 [2] from the base year 2020. A study highlights that 93% of elders, with 29% living alone, face the risk of falling [3]. Disturbingly, it was verified that around 50% of elderly individuals who remain on the floor due to fall events for over an hour may succumb within six months, even without apparent injuries [4].

According to data from the Public Health Agency of Canada [5], a significant demographic shift is anticipated by 2026, with one in five Canadians aged 65 and older, compared to the ratio of eight to one in 2001. Notably, 93% of elderly individuals prefer to reside in their private homes, and among them, 29% lead solitary lives [5]. Moreover, almost 62% of injury-related hospitalizations for the elderly result from falls [6].

In recent years, various methods utilizing advanced devices such as wearable sensors, accelerometers, gyroscopes, magnetometers, etc., have been proposed to identify falls accurately. However, the impracticality of wearing such devices for extended periods renders this solution ineffective [7]. Recognizing the limitations of wearable devices, there is a pressing need for an in-depth surveillance system for senior individuals that can autonomously and promptly detect falls within a room, notifying caretakers instantly. Achieving this goal necessitates the implementation of a sensor-based alarm generation system in the living space.

In recent years, among various types of deep learning models [8], the Convolutional Neural Network (CNN) has demonstrated remarkable success in diverse computational tasks such as image segmentation, object recognition, natural language processing, image understanding, and machine translation [9]. However, the efficacy of CNN in identifying different poses, as explored in [10], encounters challenges in varying illumination conditions. The background subtraction method used in this context tends to misclassify datasets due to shadows, leading to false predictions, especially in bending, crawling, and sitting positions.

An alternative approach to human fall detection was introduced by Tamura et al. [11], utilizing a gyroscope and an accelerometer to trigger a wearable airbag upon detecting a fall. The system's design involved producing 16 subjects to simulate falls, and a thresholding technique was applied for fall detection.

In the realm of real-life action recognition systems, a comprehensive overview is presented in [12], combining Deep Bidirectional LSTM (DB-LSTM) and CNN. While DB-LSTM recognizes hidden sequential patterns, CNN extracts data from video frames. However, limitations are observed in scenarios with identical backgrounds and occluded environments, leading to false projections.

### A. Major Challenges

The main problems are as follows: Fall detection systems help prevent patient falls in hospitals and nursing homes, reducing injuries and complications. These systems ensure the safety of elderly individuals living independently by detecting and alerting to falls. Fall detection technology can monitor and alert to seizures and sudden falls in patients

with epilepsy and other seizure-related conditions. It aids in assessing changes in a patient's gait and mobility, useful for tracking physical function and detecting underlying medical issues.

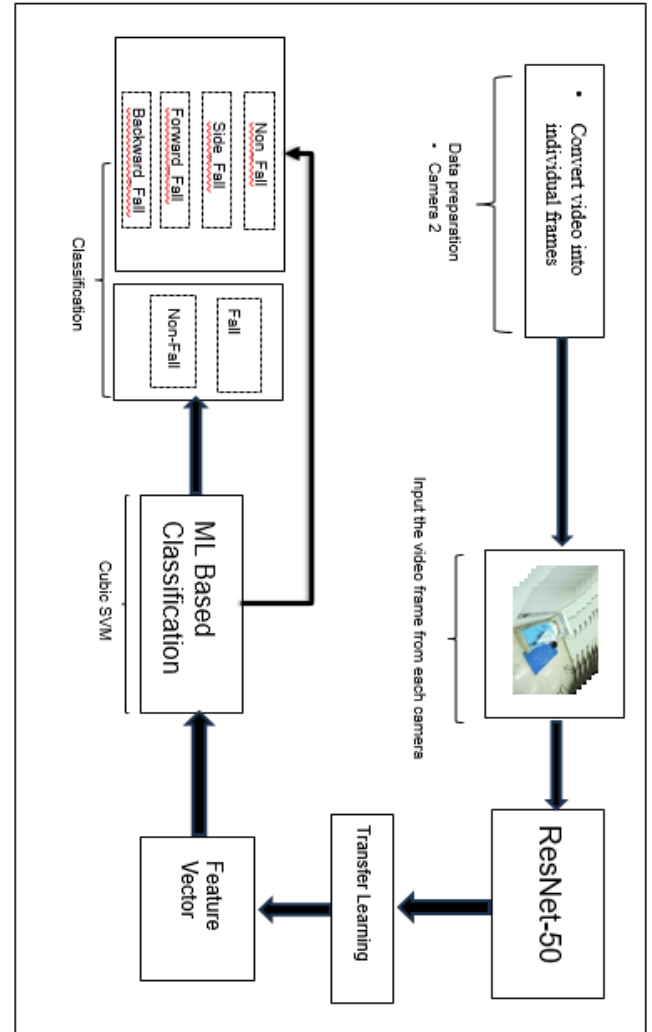
### B. Major Contributions

The aim of this study is to overcome the limitations of existing methods by presenting an innovative framework for accurate human fall direction image classification through a novel deep learning approach. The proposed framework encompasses the following sequential steps:

- Extraction of images or frames from videos.
- Modification of ResNet-50 pretrained deep learning models, incorporating an additional layer. Introduction of a new layer connecting the preceding layers using fully connected (FC) layers.
- Utilization of the features extracted from the modified models for subsequent classification.

## II. METHODOLOGY

In the methodology section, we introduced a novel approach for classifying the direction of human falls, combining elements of deep learning and machine learning, illustrated in Figure 1. Our proposed methodology comprises four distinct steps. In the initial step, we extracted images from the videos, followed by data pre-processing in the second step. Subsequently, data labelling was conducted in the third step, and in the fourth step, features were extracted using deep learning models. Ultimately, classification was executed through a machine learning model. This approach leverages advanced techniques in transfer learning to improve the working performance of ResNet-50 pre-trained models specifically. Following the extraction of features from these modified model, utilizing fall direction dataset, the resultant feature vectors from the proposed dataset through the modified models were passed to the Cubic SVM machine learning classifiers for the conclusive classification. Our proposed classifiers for this task include Cubic SVM.



**Figure 1.** Proposed diagram of human fall direction classification

### A. Data Collection

The dataset utilized in our study encompasses two primary classes: Non-fall and Fall. For the Non-fall class, we collected data from 416 videos, each captured at a rate of 59.94 frames per second (fps). The videos were recorded over a 10-second interval, with each frame featuring a size of 3840 by 2160 pixels. The Fall class, initially comprising a general classification, was further subdivided into three subclasses: forward fall, backward fall, and side fall. This refinement allows for a more nuanced analysis of fall patterns. The dataset was obtained from Soonchunhyang University Cheonan Hospital, Republic of Korea, ensuring a diverse and representative collection. For the Fall class, we gathered data from 1232 videos, also recorded at 59.94 fps and spanning a 10-second duration. Similar to the Non-fall class, the frame size for Fall class videos is 3840 by 2160 pixels. This comprehensive table outlines the key details regarding the video count, frame rate, time interval, and frame size for both classes and their respective subclasses in our dataset, as presented in Table 1.

**Table 1.** Detail description of proposed dataset.

Class	Video Number	Video fps	Time interval	Frame Size	Video Size



Non-fall	416	59.94	10s	3840* 2160	12.5M B
Fall	1232	59.94	10s	3840* 2160	12.5M B

### B. Modified Resnet50

The ResNet architecture is highly esteemed for its remarkable ability to facilitate a more direct flow of information within the network, effectively resolving the challenge of disappearing gradients during backpropagation. Its efficiency in handling deep networks is well-acknowledged, attributed to the incorporation of shortcut connections that address the vanishing gradient problem. In our Fall direction study, we harnessed the power of ResNet-50, boasting 23 million parameters. To tailor it to our research needs, we made modifications by replacing the final classification layer and employing advanced transfer learning techniques for model fine-tuning. This tailored adaptation excels particularly in feature extraction, yielding feature vectors with dimensions  $N \times 2048$ .

## III. RESULTS AND DISCUSSION

This section details the exhaustive experimental procedures applied to evaluate the proposed framework. The outcomes are visually depicted through graphs, accompanied by meticulously defined performance metrics. Our investigation involved the segmentation of the HFD (Human Fall Direction) image dataset into training and testing sets at an 80:20 ratio, details are shown in Table 2. The training configuration specified 100 iterations, 100 epochs, a minibatch size of 34, and a learning rate fixed at 0.0001. We implemented a 5-fold cross-validation, systematically evaluating various classifiers across a spectrum of performance metrics, encompassing precision, rate, recall, and accuracy. MATLAB 2022a was employed for all simulations, executed on a system equipped with a Core i7 processor and 8 GB of RAM.

**Table 2.** Splitting of training and testing images for evaluations.

Classes	Total Images	Training Images	Testing Images
Non_Fall	600	480	120
Side_Fall	600	480	120
Forward_Fall	600	480	120
Backword_Fall	600	480	120
Total Images	2400	1920	480

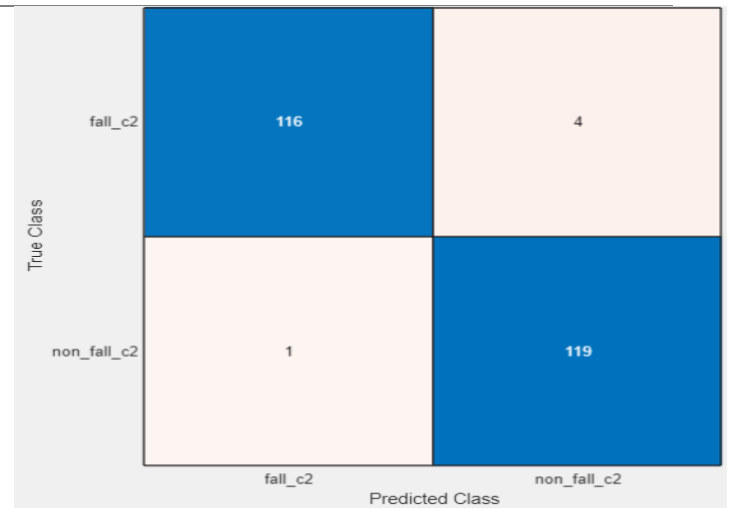
### C. Experimental results for Fall and Non-fall dataset

Table 3. Presents the outcomes of the human fall direction classification achieved through the ResNet50 model. The test features for Fall and Non-fall extracted from ResNet50 underwent evaluation using two distinct machine learning classifiers. Remarkably, the CSVM classifier

exhibited the highest accuracy, reaching an impressive 97.9%. Notably, this model demonstrated commendable performance across various metrics, with computational time standing at 30 seconds, recall rate at 97.6, and precision rate at 95.8. Following closely, the LSVM classifier secured the second-highest accuracy at 96.4%, accompanied by corresponding values of 35.2 seconds for computational time, 96.2 for recall rate, and 96 for precision rate and AUC at last Wide neural network (WNN) got the accuracy of 95.9%. Further insights into the classification performance can be gleaned from the confusion matrix, specifically for the ResNet-50 and CSVM model, as shown in Figure 3.

**Table 3.** Proposed results of human Fall and Non-fall direction classification using ReNet-50 and CSVM.

DL Models	ML Classifiers	Accuracy	Time (sec)	Precision Rate	Recall Rate
ResNet-50	1- CSVM	97.9%	30	95.8	97.6
	2- LSVM	96.4%	35.2	96	96.2
	3- WNN	95.9%	30	95.9	96



**Figure 2.** Confusion matrix of proposed results of human Fall and Non-fall direction classification using ReNet-50 and CSVM.

### D. Performance Analysis across Four Classes of Fall Directions

Table 4 presents the outcomes of the human fall direction classification achieved through the ResNet50 model. The test features for Back-fall, Side-fall, Forward-fall and Non-fall extracted from ResNet50 underwent evaluation using two distinct machine learning classifiers. Remarkably, the CSVM classifier exhibited the highest accuracy, reaching an impressive 97.9%. Notably, this model demonstrated commendable performance across various metrics, with computational time standing at 40.2 seconds, recall rate at 97.6, and precision rate at 98. Following closely, the Wide neural network (WNN) classifier secured the second-highest accuracy at 96.3%, accompanied by corresponding values of 25.2 seconds for computational time, 96.3 for recall rate, and 95.2 for precision rate and at last LSVM got the accuracy of 95.%. Further insights into the classification

performance can be gleaned from the confusion matrix, specifically for the ResNet-50 and CSVM model, as shown in Figure 4. The comparison with existing techniques is presented in Table 5.

**Table 4.** Proposed results of human fall direction classification with four number of classes using ReNet-50 and CSVM.

DL Models	ML Classifiers	Accuracy	Time (sec)	Precision Rate	Recall Rate
ResNet-50	1- CSVM	97.9%	40.2	98	97.6
	2- WNN	96.3%	25.2	96.2	96.3
	3- LSVM	95%	33.2	95.1	96.2

True Class	back_fall_c2	116	2	1	1
	forward_fall_c2	1	119		
	non_fall_c2	1		119	
	side_fall_c2	4			116
		back_fall_c2	forward_fall_c2	non_fall_c2	side_fall_c2
		Predicted Class			

**Figure 3.** Confusion matrix of human fall direction classification with four number of classes using ReNet-50 and CSVM.

**Table 5.** Comparison with existing techniques.

Reference	Method	Accuracy	Classes	Year
[12]	DCNN	97.6%	2 Fall/non-Fall	2024
[13]	Open-pose + LSTM	92%	2 Fall/non-Fall	2020
[14]	2D-pose estimation	95%	2 Fall/non-Fall	2017
Proposed	TL based deep learning models	97.9%	2 and 4 No of classes	2024

#### IV. CONCLUSIONS

In this paper, we present a hybrid deep learning and machine learning model for fall direction classification. Utilizing data from Soonchunhyang University Cheonan Hospital, we apply transfer learning and ResNet-50 models, achieving an outstanding 97.9% accuracy. Our feature work involves incorporating an alarm system for timely action and optimizing computational efficiency by focusing on salient regions for feature extraction. This enhancement aims to improve responsiveness and streamline fall direction classification.

#### ACKNOWLEDGMENT

This research was supported by the MSIT (Ministry of Science and ICT), Korea, under the ICAN (ICT Challenge and Advanced Network of HRD) support program (IITP-2024-2020-0-01832) supervised by the IITP (Institute for Information & Communications Technology Planning & Evaluation).

#### REFERENCES

- [1] Mubashir, M.; Shao, L.; Seed, L. A survey on fall detection: Principles and approaches. *Neurocomputing* 2013, 100, 144–152.
- [2] Yang, L.; Ren, Y.; Hu, H.; Tian, B. New fast fall detection method based on spatio-temporal context tracking of head by using depth images. *Sensors* 2015, 15, 23004–23019.
- [3] Rougier, C.; Meunier, J.; St-Arnaud, A.; Rousseau, J. Robust video surveillance for fall detection based on human shape deformation. *IEEE Trans. Circuits Syst. Video Technol.* 2011, 21, 611–622.
- [4] Lord, S.; Smith, S.; Menant, J. Vision and falls in older people: Risk factors and intervention strategies. *Clin. Geriatr. Med.* 2010, 26, 569–581.
- [5] Canada's Aging Population. Public Health Agency of Canada, Division of Aging and Seniors. 2002.
- [6] Reports on Senior's Falls in Canada. Public Health Agency of Canada, Division of Aging and Seniors. 2005.
- [7] Chen, Y.; Li, W.; Wang, L.; Hu, J.; Ye, M. Vision-Based Fall Event Detection in Complex Background Using Attention Guided Bi-Directional LSTM. *IEEE Access* 2020, 8, 161337–161348.
- [8] Liaqat, S.; Dashtipour, K.; Arshad, K.; Assaleh, K. A Hybrid Posture Detection Framework: Integrating Machine Learning and Deep Neural Networks. *IEEE Sens. J.* 2021, 21, 9515–9522.
- [9] Jahanjoo, A.; Naderan, M.; Rashti, M.J. Detection and Multi-class Classification of Falling in Elderly People by Deep Belief Network Algorithms. *J. Ambient Intell. Humaniz. Comput.* 2020, 11, 4145–4165.
- [10] LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* 2015, 521, 436–444.
- [11] Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Region-based convolutional networks for accurate object detection and segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* 2016, 38, 142–158.
- [12] <https://link.springer.com/article/10.1007/s11042-023-16476-6>
- [13] Lin, C.B.; Dong, Z.; Kuan, W.K.; Huang, Y.F. A framework for fall detection based on openpose skeleton and lstm/gru models. *Appl. Sci.* 2021, 11, 329.
- [14] Núñez-Marcos, A.; Azkune, G.; Arganda-Carreras, I. Vision-based fall detection with convolutional neural networks. *Wirel. Commun. Mob. Comput.* 2017, 2017, 9474806

# Posture Recognition System for Adults and Children using YOLO

Neunggyu Han<sup>1\*</sup> and Yunyoung Nam<sup>2</sup>

<sup>1</sup>Department of ICT Convergence, Soonchunhyang University, Asan 31538, Korea

<sup>2</sup>Department of Computer Science and Engineering, Soonchunhyang University, Asan 31538, Korea

\*Contact: az0422@naver.com

**Abstract**— A variety of information can be obtained through analysis of child behavior. In particular, children with developmental disabilities may display behavioral patterns that are different from typical children. In addition, rapid diagnosis and treatment of diseases such as these are important. For diseases like this, the guardian's observation and judgment are generally important. Additionally, since diagnosis at a hospital relies on the judgment of the guardian, an accurate diagnosis may be difficult. Therefore, direct observation by an expert can obtain very accurate results. However, there are great difficulties in having experts always visit the places where children are. Also, creating a space for child behavior analysis in a hospital may not be used as a rejection by children or parents. Therefore, in this paper, we implement a system that analyzes behavior after creating a space for children in a place other than a hospital. This space is in the form of a cafe for children, and four cameras on the ceiling observe without blind spots. And based on the images from these cameras, adults and children are classified using YOLO and their postures are also analyzed. At the current stage, it was possible to recognize the posture for each frame for one angle. Therefore, in the next step, the exact posture of the object will be recognized through the coordinate change amount, and the distance between objects will be calculated to analyze whether or not there is interaction.

## I. INTRODUCTION

Recently, the problem of children with developmental disabilities has been intensifying in Korean society. In particular, the incidence of premature and low birth weight babies is increasing due to advanced pregnancy and childbirth. Therefore, the rates of developmental delays and communication disorders are increasing. Additionally, various activities of children have been restricted due to COVID-19. As a result, children's opportunities to develop large and small muscles are reduced, and the frequency of aggressive behavior is increasing due to increased stress. Lastly, there is a problem that children's language development is also delayed due to the use of masks. For developmental disorders such as these, rapid diagnosis and treatment are very important. However, the current diagnostic method has the problem of being somewhat inaccurate as it relies solely on parental observation and judgment. In particular, in the case of dual-income families, there is a problem in that as the time to care for children decreases, there are fewer opportunities to detect children's

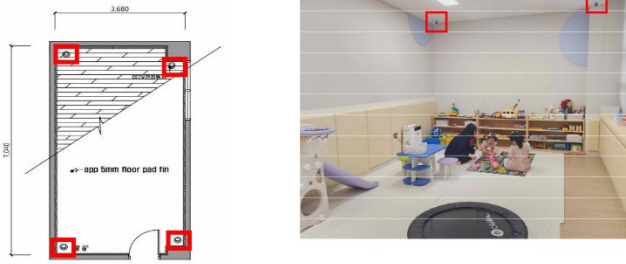
abnormal behavior. Therefore, current diagnostic methods are an obstacle to accurate diagnosis. The most accurate way to diagnose with current methods is for an expert to directly analyze the child's behavior. To diagnose using this method, an expert must visit the child in person. However, because there is a limit to the number of specialized personnel, this method has its limitations. Another way is to create a space in the hospital where children and parents can participate in activities together. This method has the advantage of solving the problem that experts have to travel directly and utilizing professional manpower more efficiently. This method also has limitations. In particular, parents feel reluctant about going to the hospital. In cases like these, quick diagnosis and treatment may not be possible. Therefore, in this paper, we implement a system to analyze behavior by creating a space outside the hospital where the behavior of children and parents can be analyzed. First, this space is equipped with toys and books for children to play with. Additionally, in this space, parents and children can act without resistance, as if using a kids cafe. Next, in this space, four cameras on the ceiling observe without blind spots. And based on the images from these cameras, adults and children are classified and their postures are recognized. There are a total of 5 recognized postures: 2 for adults and 3 for children. Adult postures are divided into stand and sit, and children's postures are divided into stand, sit, and log (lying, rolling, or crawling postures). At the current stage, we have succeeded in recognizing posture for one channel. In the next step, we will analyze scores through interactions between children and parents and establish a diagnostic system for suspicious symptoms regarding toys handled by children.

## II. SYSTEM IMPLEMENTATION

### A. Room Setup

A space was prepared to analyze children's behavior. This space is prepared separately from the hospital, so children and parents can visit without feeling uncomfortable. Additionally, this space has toys and books, similar to a kids cafe. This space is prepared so that parents and children can behave as usual. In this space, four cameras are installed on the ceiling, as shown in Figure 1. This camera observes all spaces in real time without blind spots. In addition, the sound is also

recorded so that the conversation between the child and the parent can be analyzed. This paper conducts experiments based on videos filmed in this space.



**Figure 1.** Room setup. The red boxes mean cameras position.

### B. Dataset Preparation

The dataset for the experiment is conducted using directly collected video data. This video data records children and adults appearing in the space for this experiment and behaving normally. A dataset was prepared based on this recorded video. This paper conducted an experiment as a preliminary verification for future experiments. Therefore, a dataset was prepared for one camera out of four cameras. First, the size of the recorded video data is UHD (3840x2160) and is recorded at 100fps. Nine of these video data were prepared, six were used for training and three were used for validation. This video data was used by extracting frame data at approximately 16 frames per second. A dataset was prepared by labeling the extracted frame data as shown in Figure 2.



**Figure 2.** Example for labelling. 2 adults were labelled ‘adult stand,’ 1 adult was labelled ‘adult sit,’ and 1 child was labelled ‘child stand.’

### C. Train for obtaining weights

Using the prepared dataset, we derive weights for posture recognition. Weights were trained using the x-large model, which has the highest accuracy among YOLOv8 [1]. The number of data used for learning is small. Therefore, in this paper, accuracy was secured through transfer learning. Transfer learning is a method used when there is little data and is a method of learning our dataset using existing weights learned from data similar to our dataset. The main feature of this method is that higher accuracy can be obtained compared to learning with Scratch. In this paper, transfer learning was performed using the weights provided by YOLOv8. This weight is learned with the MS-COCO [2] dataset. Since this weight also includes humans, high accuracy can be achieved. The hyper-parameters used for learning are as follows. First,

the epoch was set to 50. Also, batch was set to 32. The remaining hyper-parameters used default values. In the case of batch, if the default value of 16 was used, there was a problem that a lot of noise occurred in recognition, so the higher value of 32 was used.

## III. EXPERIMENTAL RESULT

### A. Evaluation of weights

First is the evaluation of weights. This weight was evaluated using the best weight after learning was completed. The evaluation dataset used here is three video data for validation. The model evaluation results are shown in Table 1. It can be seen that the accuracy based on mAP50 is over 80%. With this, the performance can be considered sufficiently high. However, with regard to the sitting posture of adults, the data included in the validation is very small, so the accuracy appears to be low.

Table 1. Evaluation of weights

Class	Image	Instance	P	R	mAP50
All	171	192	84.8%	74.8%	74.9%
adult stand	171	142	96.8%	96.5%	91.8%
adult sit	171	4	45.8%	25.0%	20.6%
child stand	171	12	92.0%	96.3%	95.9%
child sit	171	29	89.5%	96.6%	87.4%
child log	171	5	100%	59.7%	78.7%

### B. Posture recognition

Figure 3 shows the results of posture recognition on verification video data. As you can see from these results, you can see that all postures are recognized normally.



(a) An adult stand



(b) An adult sit, and a child stand



(c) An adult sit, and a child lie

**Figure 3.** Examples of posture recognition

#### IV. CONCLUSIONS

In this paper, a system for analyzing children's behavior was prepared. A space was prepared to analyze the behavior of children and parents, and posture was recognized using one of four cameras installed in the space. Additionally, YOLOv8x was used for posture recognition. As a result, an accuracy of 80% was achieved in the verification data set, and as a result of recognizing the posture with the verification data, it was

confirmed that it was recognized without problems. Based on the results of this experiment, we will analyze the interaction between children and parents and implement a system that represents this as a score. In addition, we will implement a system that analyzes the interaction between children and toys to diagnose the expected type of child's developmental disability.

#### ACKNOWLEDGEMENT

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. RS-2023-00218176)

#### REFERENCES

- [1] Jocher, G., Chaurasia, A., & Qiu, J. (2023). Ultralytics YOLO (Version 8.0.0) [Computer software]. <https://github.com/ultralytics/ultralytics>
- [2] Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., ... & Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13 (pp. 740–755). Springer International Publishing.



# Improved Sampled-Data Control of Human-Assisting Drone Systems with an Asynchronous Sampling Time

Seok Young Lee<sup>1,\*</sup>

<sup>1</sup>Electronic Engineering, Soonchunhyang University, Asan, Republics of Korea

\*Contact: suk122@sch.ac.kr

**Abstract**— This paper is concerned with the stability analysis and stabilization problem of a network-controlled drone system having asynchronous sampling times. Owing to the inevitable limitations in hardware resources and network bandwidths, choosing sampling times that can guarantee system stability has played an important role in controlling remote systems. Based on an input-delay approach, a sampled-data system can be modeled into a linear system with a time-varying delay. Utilizing the input delay approaches, we derive stability criteria for the drone system with an asynchronous sampling time. Also, control gains are developed with the stability criteria. Numerical examples demonstrate the effectiveness of the proposed approaches.

## I. INTRODUCTION

This paper is concerned with the stability analysis and stabilization problem of a network-controlled drone system having asynchronous sampling times. Between two successive sampling times, a sampled-data system is remotely controlled by previous sampled input and thus can be reformulated into a linear system with a saw-tooth delay. Naturally, system information at the sampling moment have played a key role in stability analysis and stabilization of the sampled-data systems. This paper utilizes the results of [2] for the stability analysis and stabilization of a network-controlled drone system with an asynchronous sampling times. Stability and stabilization conditions are obtained via linear matrix inequality (LMI) conditions. Numerical examples demonstrate the effectiveness of [2] in terms of allowable sampling intervals.

## II. MAIN RESULTS

Consider the following sampled-data system of [1].

$$\dot{x}(t) = Ax(t) + A_d x(t_k), \forall t \in [t_k, t_{k+1}) \quad (1)$$

where  $x(t) \in R^n$  is the system state,  $A, A_d \in R^{n \times n}$  are system matrices, and  $t_k$  is the sampling instant such that  $\cup k \in N[t_k, t_{k+1}) = [0, +\infty)$ . The sampling interval is defined as

$$t_{k+1} - t_k = h_k \quad (2)$$

Based on Theorem 2 of the paper [2], this paper discusses the Quanser AERO 2-DOF drone system [3] with a asynchronous sampling time. This helicopter system can be modelled as a free body diagram in Figure 1. This system consists of two identical rotors that produce the thrust forces  $F_p(t)$  and  $F_y(t)$  acting at two points with distances  $r_p$  and  $r_y$  from the z-axis along the x-axis, respectively. Thus, a propeller generates a torque around the y-axis leading to a pitch motion, while the other handles a yaw motion. This system can be described as follows:

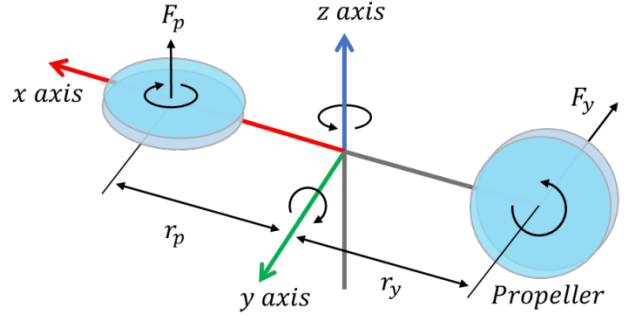


Fig 1 A free-body diagram of the 2-DOF drone system

$$\tau_p(t) = J_p \ddot{\theta}(t) + D_p \dot{\theta}(t) + K_{sp} \theta(t), \quad (3)$$

$$\tau_y(t) = J_y \ddot{\phi}(t) + D_y \dot{\phi}(t), \quad (4)$$

where  $J_p = J_y = 0.0215$  are moments of inertia about pitch axis and yaw axis, respectively.  $D_p = 0.0071$  and  $D_y = 0.0220$  are viscous friction constants about pitch axis and yaw axis, respectively.  $K_{sp} = 0.0374$  is the stiffness about the pitch axis. The torques  $\tau_p(t)$  and  $\tau_y(t)$  which respectively acts on the pitch and the yaw axes are assumed to be proportional to the input DC voltages  $V_p(t)$  and  $V_y(t)$  of the rotors such that:

$$\tau_p(t) = K_{pp} V_p(t) + K_{py} V_y(t), \quad (5)$$

$$\tau_y(t) = K_{yp} V_p(t) + K_{yy} V_y(t), \quad (6)$$

where  $K_{pp} = 0.0011$  and  $K_{py} = 0.0021$  are thrust torque gain acting on pitch axis from pitch propeller and yaw propeller, respectively. Also,  $K_{yp} = -0.0027$  and  $K_{yy} = 0.0022$  are thrust-torque gains acting on yaw axis from pitch propeller and yaw propeller, respectively. By utilizing the equations (4)-(6) and the state variable  $x(t) = [\theta(t) \phi(t) \dot{\theta}(t) \dot{\phi}(t)]^T$ , this helicopter system can be represented as a sampled-data system (1) with the following matrices

$$A = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -K_{sp}/J_p & 0 & -D_p/J_p & 0 \\ 0 & 0 & 0 & -D_y/J_y \end{bmatrix}, \quad (7)$$

$$A_d = BK, \quad B = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ K_{pp}/J_p & K_{py}/J_p \\ K_{yp}/J_y & K_{yy}/J_y \end{bmatrix}, \quad (8)$$

$$K = \begin{bmatrix} 0.0432 & 0.0530 \\ 1.1617 & -0.6085 \\ -0.1687 & -0.2070 \\ -0.1789 & 0.0937 \end{bmatrix}^T. \quad (9)$$

Here, a matrix  $K$  is a gain of a sampled-state feedback controller, and  $B$  is a system matrix. In this example, i newly derive an analytic upper bound  $h = 12.0942$  of a periodic sampling interval. With a periodic sampling time, sampled-data system can be regarded as a linear discrete-time system. Integrating the differential equation (1) yields

$$x(t) = \Gamma(t - t_k), t \in [t_k, t_{k+1}], \\ \Gamma(\tau) = e^{A\tau} + \int_0^\tau e^{A(t-\tau)} dr A_d, \tau \geq 0.$$

Under the periodic sampling  $h = h_k$ , the system dynamics in (1) becomes

$$x(t_{k+1}) = \Gamma(h)x(t_k) \quad (10)$$

This system is asymptotically stable if and only if  $\Gamma(h)$  has all eigenvalues inside the unit circle. However, under asynchronous sampling, such method does not hold. Therefore, utilizing Theorem 2 of the paper [2], i verify that the system (1) with the matrices (7)-(9) is asymptotically stable under  $h_k \in [10^{-5}, 9.5042]$ . In Figure 2, all state responses with an initial condition  $x(0) = [10 \ 45 \ 0 \ 0]^T$  converge.

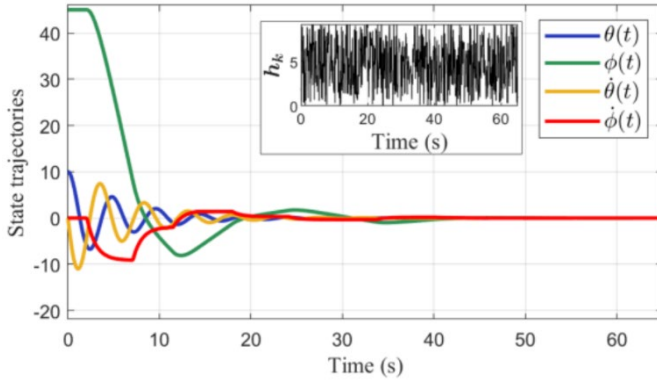


Fig 2 The state  $x(t) = [\theta(t) \ \phi(t) \ \dot{\theta}(t) \ \dot{\phi}(t)]^T$  trajectories of the helicopter system with an asynchronous sampling interval  $h_k \in [10^{-5}, 9.5042]$

Further, the system (1) with a state feedback controller  $u(t) = Kx(t)$  can be represented as follows.

$$\dot{x}(t) = Ax(t) + Bu(t) \quad (11)$$

Differently from the stability analysis problem, where the control gain  $K$  (9) is given, stabilization of the system (11) considers a free variable  $K$  in developing LMI conditions. This paper newly derives stabilization criteria and thus obtains following control gain, which guarantees that the system (1) with the matrices (7)-(8) is asymptotically stable under  $h_k \in$

$$K = \begin{bmatrix} 5.8700 & 0.2389 & 0.4564 & -0.7180 \\ 7.2041 & -0.1251 & 0.5601 & 0.3761 \end{bmatrix} \quad (12)$$

$[10^{-5}, 9.5042]$ . In Figure 3, all state responses with an initial condition  $x(0) = [10 \ 45 \ 0 \ 0]^T$  converge.

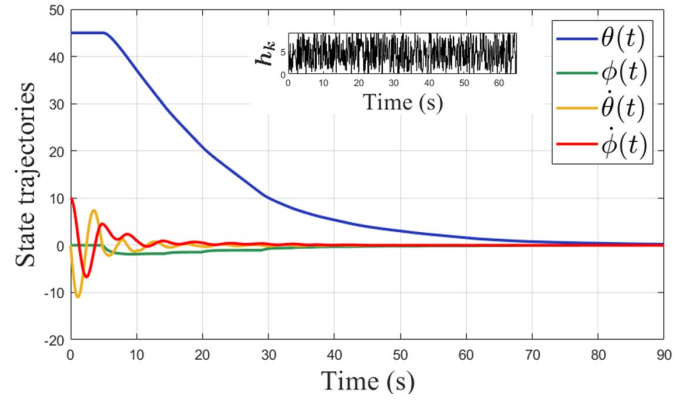


Fig 3 The state  $x(t) = [\theta(t) \ \phi(t) \ \dot{\theta}(t) \ \dot{\phi}(t)]^T$  trajectories of the helicopter system with an asynchronous sampling interval  $h_k \in [10^{-5}, 9.5042]$

### III. CONCLUSIONS

This paper has discussed stability analysis and stabilization of the sampled-data controlled drone system with an asynchronous sampling time. In the future works, the derived results also can be applied to problems concerned with sampled-data synchronization of various systems including fuzzy systems, switched systems, and delayed chaotic Lur'e systems.

### REFERENCES

- [1] N. K. Kwon and S. Y. Lee, "Novel equalities for stability analysis of asynchronous sampled-data systems," IEEE Access, vol. 8, pp. 177 195–177 205, 2020.
- [2] S. Y. Lee, "Improved stability criteria for sampled data systems via a novel looped-functional," in The 13th Asian Control Conference (ASCC), May 2022.
- [3] J.M. Park, "An improved stability criterion for networked control systems with a constant transmission delay," Journal of the Franklin Institute, vol. 359, no. 9, pp. 4346–4365, 2022.



# Quasi Convex Adaptive Sliding Mode Control Combined with Time-Delay Control for A Human-Assisting Manipulator

Dong Hee Seo<sup>1</sup>, Jin Woong Lee<sup>1</sup> and Seok Young Lee<sup>1,\*</sup>

<sup>1</sup>Department of ICT Convergence Engineering, Soonchunhyang University, Asan, South Korea

\*Contact: suk122@sch.ac.kr

**Abstract**— This paper proposes a novel adaptive sliding mode control (ASMC) algorithm based on quasi-convex functions for robot manipulators. The proposed ASMC utilizes quasi-convex functions to minimize overestimation of the control gain and aim to keep the sliding surface at arbitrarily small vicinity to zero. It also includes an algorithm to compensate for the time delay estimation (TDE) error induced by time delay control (TDC) method. This article applies the combination of TDC and a novel ASMC to guarantee asymptotic stability of a real robot manipulator with disturbance. The paper demonstrates the uniformly ultimately bounded (UUB) stability with arbitrarily small bound of sliding variable in the combined ASMC. The proposed ASMC is then applied to both simulation and real systems to compare the control performance with that of conventional ASMC methods.

## I. INTRODUCTION

Robot manipulators are utilized for a range of tasks that require precision. In particular, it is widely used in factories [1] and medical facilities [2], [3]. Robot manipulators are being applied for work and to assist or replace humans in everyday life. In both situations, manipulators require a high degree of motion accuracy. However, motion control of robot manipulators is difficult due to nonlinearities, time-varying parameters, unknown disturbances, and modeling uncertainties. Such undesirable factor can hinder the manipulators control performance and cause instability [4]. Nonlinear robust control algorithm called sliding mode control (SMC) have been utilized to deal with the above factors. [5]. The SMC is robust to unknown dynamics, and robust stabilization is ideally achieved when the switching gain is greater than an uncertain upper bound. However, in practice, the upper bound is unknown [4]. Thus, the switching gain is selected to be sufficiently large than the uncertainty. This may cause chattering, which can be eliminated by setting the appropriate switching gain [5]. This creates the contradiction of knowing the uncertainty of the unknown. Therefore, SMC is difficult in real robot manipulator applications. To address these problems, adaptive sliding mode control (ASMC) [6], [7] was proposed as a combined adaptive gain. In fact, as stated in [8], the high gain of a monotonically increasing adaptive law can lead to overestimation and cause chattering [9]. This increases chattering and reduces control performance.

To reduce overestimation, a function-based adaptive gain control, such as a barrier function adaptive sliding mode control (BFASMC) was proposed [10]. This approach ensures that the sliding surface converges to zero without significantly overestimating the gain. In a function-based adaptive gain, the function design significantly affects control performance. This paper proposes a new function-based adaptive sliding mode control with a quasi-convex function [11]. The quasi-convex adaptive sliding mode control (QCASMC) adjusts the adaptive gain according to the quasi-convex function when the sliding surface is near zero, which reduces overestimation. This paper compares the BFASMC and the QCASMC on a 2-DOF robot manipulator in simulation.

The controller is an algorithm based on a model that requires knowledge of the nominal dynamics of the robot manipulator. However, the nominal part of the manipulator dynamics is challenging to know or measure. TDC is commonly used in mechanical systems due to its performance in dealing with model uncertainties and disturbances. TDC estimates the robot manipulator information using the previous angular acceleration and input torque. The estimation error depends on the sampling period, which is called the operating cycle. However, it is challenging to provide a fast operating cycle, which results in a TDE error. TDE errors are an additional disturbance to the robot manipulator that degrade control performance. To reduce uncertainties such as TDE errors, various TDC methods have been studied. In paper [12], TDC using the TDE error of the previous sampling time is proposed. ASMC works with SMC in addition to the TDC technique [13]-[18]. This paper proposes a combined approach of QCASMC and TDC based on the previous discussion. Then, it can be applied to a real system to improve performance. The tracking error of the proposed ASMC is combined with the TDE technique to ensure the stability of the uniformly ultimately bounded (UUB) for sliding variables using the Lyapunov method. In Section II, TDC-based QCASMC for robot manipulators has been presented. In Section III, the simulation of the 2-Dof manipulator has been presented. In Section IV, the proposed ASMC with TDC is applied to a real robot manipulator. In Section V, we provide a brief conclusion of the paper.

## II. PROPOSED ASMC COMBINED WITH TIME-DELAY CONTROL

The Euler-Lagrange dynamic equation of an n-degree of freedom (DOF) robot manipulator can be consider as

$$M(q_t)\ddot{q}_t + C(q_t, \dot{q}_t)\dot{q}_t + G(q_t) + F(\dot{q}_t) = \tau_t, \quad (1)$$

where system parameters  $q, \dot{q}, \ddot{q} \in \mathbb{R}^n$  are the angular position, velocity, and acceleration vectors of each manipulator joint, respectively.  $M(q_t) \in \mathbb{R}^{n \times n}$  is the inertia matrix of the robot manipulator,  $C(q_t, \dot{q}_t) \in \mathbb{R}^{n \times 1}$  is the Coriolis matrix,  $G(q_t) \in \mathbb{R}^{n \times 1}$  is the gravity force vector,  $F(\dot{q}_t) \in \mathbb{R}^{n \times 1}$  is the friction force vector.  $\tau_t \in \mathbb{R}^{n \times 1}$  is the control input torque of each manipulator joint.

The dynamics (1) can be represented as follows:

$$\ddot{q}_t = \sigma_t + M^{-1}\tau_t, \quad (2)$$

where  $\sigma_t \triangleq \bar{M}^{-1}\{[M(q_t) - \bar{M}]\ddot{q}_t + C(q_t, \dot{q}_t)\dot{q}_t + G(q_t) + F(\dot{q}_t)\}$ , and  $\bar{M} = \text{diag}\{m_1, m_2, m_3, \dots, m_n\}$  is control gain matrix. Due to the time-varying and nonlinear nature of  $\sigma_t$ , it is difficult to determine its exact value. As a digital controller with a sampling period is utilized to control the robot manipulator, the TDC is utilized by the controller to estimate  $\sigma_t$  from  $\bar{\sigma}_t = \sigma_{t-L}$  due to its sampling period. where  $L$  is the sampling period. From (2), we can get  $\bar{\sigma}_t$

$$\bar{\sigma}_t = \ddot{q}_{t-L} - \bar{M}^{-1}\tau_{t-L}, \quad (3)$$

$$\tau_t + \bar{M}\bar{\sigma}_t = \bar{M}\ddot{q}_t. \quad (4)$$

The TDE error is defined by  $\mu_t = \sigma_t - \bar{\sigma}_t$  and,  $L$  small and close to zero and is bounded.

$$\|\mu_t\|_\infty \leq \mu^*, \quad (5)$$

where  $\mu^*$  is a positive value. The proof of (5) is taken from [12]. We define system error as  $e_t = q_{dt} - q_t$ , where  $q_{dt} \in \mathbb{R}^{n \times 1}$  is the desired angular position. The sliding surface is chosen as

$$s_t = \dot{e}_t + K_1 e_t, \quad (6)$$

where  $K_1 \in \mathbb{R}^{n \times n}$  denotes the nonzero positive diagonal matrix. Construction of the controller [19] using the sliding variable defined in the expression (6).

$$\hat{\tau}_t = -\bar{M}q_{dt-L} + \tau_{t-L} + \bar{M}(\ddot{q}_{dt} + K_d \dot{e}_t + K_p e_t). \quad (7)$$

Substituting (3) with (7), we obtain

$$\hat{\tau}_t = -\bar{M}\bar{\sigma}_t + \bar{M}(\ddot{q}_{dt} + K_d \dot{e}_t + K_p e_t). \quad (8)$$

Substituting into (2) with  $\hat{\tau}_t = \tau_t$ . Then rearrange the sliding constant  $s_t$ .

$$\begin{aligned} 0 &= \ddot{e}_t + K_d \dot{e}_t + K_p e_t + \sigma_t - \bar{\sigma}_t \\ &= \dot{s}_t + K_2 s_t = \ddot{e}_t + (K_1 + K_2)\dot{e}_t + K_1 K_2 e_t, \end{aligned} \quad (9)$$

where  $K_d = K_1 + K_2$  and  $K_p = K_1 K_2$ .  $K_1, K_2 \in \mathbb{R}^{n \times n}$  are positive diagonal gain matrices. Using TDC [12] to design the input torque.

$$\tau_t^p = \hat{\tau}_t - \gamma \bar{M}\mu_{t-L}, \quad (10)$$

where  $\gamma$  represents tunable parameter, while  $\mu_{t-L}$  denotes the previous TDE error.

The equation (10) is represented as follows:

$$\begin{aligned} \tau_t^p &= -(\gamma + 1)\bar{M}\ddot{q}_{t-L} + (\gamma + 1)\hat{\tau}_{t-L} + \gamma\bar{M}\ddot{q}_{t-2L} - \gamma\hat{\tau}_{t-2L} \\ &\quad + \bar{M}(\ddot{q}_{dt} + K_1 \dot{e}_t + K_2 s_t) \\ &= -\bar{M}\{\bar{\sigma}_t + \gamma\mu_{t-L}\} + \bar{M}(\ddot{q}_{dt} + K_1 \dot{e}_t + K_2 s_t). \end{aligned} \quad (11)$$

When an input torque (11) is applied, equation (2) follows desired dynamics.

$$\ddot{e}_t + K_d \dot{e}_t + K_p e_t + \mu - \gamma\mu_{t-L} = 0. \quad (12)$$

We propose novel ASMC and insert it to the control (11) as follow:

$$\tau_t^p = -\bar{M}\{\bar{\sigma}_t + \gamma\mu_{t-L}\} + \bar{M}(\ddot{q}_{dt} + K_1 \dot{e}_t + K_2 s_t) + \bar{M}\bar{K} \text{sgn}(s_t). \quad (13)$$

The switching gain  $\bar{K}$  is determined by a new adaptive law that contains a quasi-convex function, as follows:

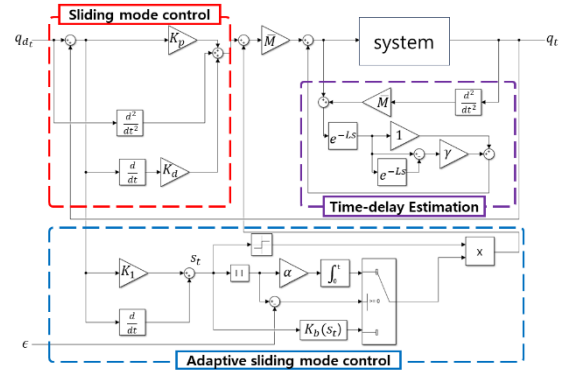


Fig. 2 Proposed ASMC diagram.

$$K_{i,t} = \begin{cases} \bar{K}_a = \alpha |s_t| & , \|s_t\|_\infty \geq \epsilon \\ K_b = 1 - e^{-\beta s_t^2} & , \|s_t\|_\infty < \epsilon. \end{cases} \quad (14)$$

$\epsilon$  is a small positive constant near zero,  $\alpha$  is a tunable positive gain for adaption speed, and the  $\beta$  is a tunable gain of slope of the convex part of the quasi-convex function. The proposed new adaptive law reduces the adaptive gain  $K_{i,t}$  to zero when  $s$  is near zero. where the control gain  $K_{i,t} \in \mathbb{R}^{n \times n}$  is a positive diagonal switching gain matrix. The  $\text{sgn}(s_t) \in \mathbb{R}^{n \times 1}$  is defined by

$$\text{sgn}(s_t) = \begin{cases} 1, & s_t \geq 0 \\ -1, & s_t < 0 \end{cases}. \quad (15)$$

**Theorem1.** For a manipulator (1) controlled by (13) with adaptive gain (14), the sliding variable enters of the sliding manifold,  $\|s_t\|_\infty < \epsilon$ , in a finite time  $t_\epsilon > 0$ .

After that, the sliding variable is guaranteed to be UUB for  $t \geq t_\epsilon$ :

$$\|s_t\|_\infty \leq \sqrt{n\epsilon^2 + \bar{K}^*}, \quad (16)$$

where  $\bar{K}^*$  is the maximum value of  $\sum_{i=1}^n \frac{1}{\alpha} (\bar{\mu}^* - K_{i,t})^2$ .

**Proof:** Let us define Lyapunov function as follows:

$$V_t = \frac{1}{2} s_t^T s_t + \frac{1}{2\alpha} \sum_{i=1}^n (\bar{\mu}^* - K_{i,t})^2, \quad (17)$$

where  $\bar{\mu}^*$  is the upper bound of  $\|\mu - \alpha\mu_{t-L}\|_\infty$ .

Based on the (6) and (9), the time derivate of Lyapunov function is given as follows:

$$\begin{aligned} \dot{V}_t &= s_t^T \dot{s}_t - \frac{\dot{K}_{i,t}}{\alpha} \sum_{i=1}^n (\bar{\mu}^* - K_{i,t}) \\ &= s_t^T (\ddot{q}_{dt} - \ddot{q} + K_1 \dot{e}_t) - \frac{1}{\alpha} \sum_{i=1}^n (\bar{\mu}^* - K_{i,t}) \dot{K}_{i,t} \\ &= s_t^T (-\mu + \gamma\mu_{t-L} - K_2 s_t + \bar{K} \text{sgn}(s_t)) - \frac{1}{\alpha} \sum_{i=1}^n (\bar{\mu}^* - K_{i,t}) \dot{K}_{i,t} \\ &\leq -K_2 s_t^T s_t + \sum_{i=1}^n \left\{ (\bar{\mu}^* - K_{i,t}) |s_{i,t}| - \frac{1}{\alpha} (\bar{\mu}^* - K_{i,t}) \dot{K}_{i,t} \right\} \\ &= -K_2 s_t^T s_t + \sum_{i=1}^n (\bar{\mu}^* - K_{i,t}) \left\{ |s_{i,t}| - \frac{\dot{K}_{i,t}}{\alpha} \right\}. \end{aligned} \quad (18)$$

As there is a boundary in  $\mu_t$ ,  $\mu - \alpha\mu_{t-L}$  is also bounded by  $\|\mu - \alpha\mu_{t-L}\|_\infty \leq \bar{\mu}^*$ . There are two conditions for the proposed adaptive law (14).

For the condition of  $\|s_t\|_\infty \geq \epsilon$ , the negativeness of  $\dot{V}_t$  is guaranteed as follows:

$$\dot{V}_t \leq -K_2 \|s_t\|_2^2, \quad (19)$$

$\|s_t\|_\infty < \epsilon$  can be reached by increasing the adaptive parameters  $\bar{K}_{i,t}$  for  $i = 1, 2, \dots, n$ .

For the condition of  $\|s_t\|_\infty < \epsilon$ , the negativity of  $\dot{V}_t$  is not guaranteed as follows:

$$\dot{V}_t \leq -K_2 \|s_t\|_2^2 + \sum_{i=1}^n (\bar{\mu}^* - K_{i,t}) \left\{ |s_{i,t}| - \frac{1 - e^{-\beta s_{i,t}^2}}{\alpha} \right\}, \quad (20)$$

$\bar{K}_{i,t}$  is adaptive parameters for  $i = 1, 2, \dots, n$ .

The upper bound of  $\|s_t\|_2$  is found for  $\|s_t\|_\infty < \epsilon$  based on  $V_t$ :

$$\frac{1}{2} \|s_t\|_2^2 \leq V_t \leq \frac{1}{2} \|s_t\|_2^2 + \sum_{i=1}^n \frac{1}{\alpha} (\bar{\mu}^* - K_{i,t})^2. \quad (21)$$

If  $\|s_t\|_\infty \geq \epsilon$ ,  $V_t$  decreases until  $\|s_t\|_\infty \leq \epsilon$ . Therefore, for  $\|s_t\|_\infty \leq \epsilon$ ,  $V_t$  has following upper bound:

$$V_t \leq \frac{1}{2} \sum_{i=1}^n \epsilon^2 + \frac{1}{2} \sum_{i=1}^n \frac{1}{\alpha} (\bar{\mu}^* - \bar{K}_{i,t})^2. \quad (22)$$

There exists a maximum value  $\bar{K}^*$  for  $\sum_{i=1}^n \frac{1}{\alpha} (\bar{\mu}^* - \bar{K}_{i,t})^2$  such that:

$$V_t \leq 0.5n\epsilon^2 + 0.5\bar{K}^*. \quad (23)$$

Obtain from (21) and (23) the upper bound of the sliding variable for  $\|s_t\|_\infty \leq \epsilon$

$$V_t \leq \sqrt{n\epsilon^2 + \bar{K}^*}. \quad (24)$$

In (24), the sliding variable of the robot manipulator (1) is UUB by the input torque.

The proof is complete.

### III. SIMULATION

#### A. Text Font of Entire Document

In the simulation, we consider the 2-DOF robot manipulators as shown in [12].

$$\begin{aligned} M(q) &= \begin{bmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{bmatrix}, C(q, \dot{q})\dot{q} = \begin{bmatrix} C_{11} \\ C_{21} \end{bmatrix}, \\ G(q) &= \begin{bmatrix} G_{11} \\ G_{21} \end{bmatrix}, F(\dot{q}) = \begin{bmatrix} F_{12} \\ F_{22} \end{bmatrix} \end{aligned} \quad (25)$$

with

$$\begin{aligned} M_{11} &= l_2^2 m_2 + 2l_1 l_2 m_2 \cos(q_2) + l_1^2 (m_1 + m_2), \\ M_{12} &= l_2^2 m_2 + l_1 l_2 m_2 \cos(q_2), \\ M_{21} &= M_{12}, M_{22} = l_2^2 m_2, \\ C_{11} &= -l_1 l_2 m_2 \sin(q_2) \dot{q}_2^2 - 2l_1 l_2 m_2 \sin(q_2) \dot{q}_1^2 \dot{q}_2^2, \\ C_{21} &= l_1 l_2 m_2 \sin(q_2) \dot{q}_2^2, \\ G_{11} &= m_2 l_2 g \cos(q_1 + q_2) + ((m_1 + m_2) l_1 g) \cos(q_1), \\ G_{22} &= m_2 l_2 g \cos(q_1 + q_2), \\ F_{11} &= f_{v1} \dot{q}_1 + f_{c1} \text{sgn}(\dot{q}_1), F_{21} = f_{v2} \dot{q}_2 + f_{c2} \text{sgn}(\dot{q}_2), \end{aligned}$$

where  $q_1$  and  $q_2$  are the angles of the robot manipulator joints 1 and 2, respectively. The joint parameters are set  $m_1 = 9[kg]$ ,  $m_2 = 6[kg]$ ,  $l_1 = 0.4[m]$ ,  $l_2 = 0.2[m]$ ,  $g = 9.8 \frac{m}{s^2}$ .

The friction coefficients are  $f_{v1} = 10[Nms]$ ,  $f_{c1} = 10[Nm]$ ,  $f_{v2} = 10[Nms]$  and  $f_{c2} = 10[Nm]$ . For the simulation, we set desired angular position  $q_d = [q_{1d}, q_{2d}]^T$ ,  $q_{1d} = \sin(0.5t)$  and  $q_{2d} = 0.6\sin(t)$ . The adjustable gains are set to  $\bar{M} = \{0.02, 0.02\}$ ,  $K_1 = \text{diag}\{10, 10\}$ ,  $K_2 = \text{diag}\{15, 15\}$ ,  $\alpha = 10$ ,  $\beta = 200$  and  $\epsilon = 0.5$ . The disturbance is given by

$$d_t = \begin{cases} [0.1 \cos(2t), 0.3 \sin(2t)]^T, & t < 5 \\ [0.2 \cos(2t), 0.4 \sin(3t)]^T, & t < 15 \\ [0.4 \cos(4t), 0.3 \sin(2t)]^T, & t \geq 15 \end{cases} \quad (26)$$

#### B. Title and Author Details

The simulation result is shown by comparing the sliding surface of BFASMC and QCASMC in Fig. 1 and Fig. 2 also shows the

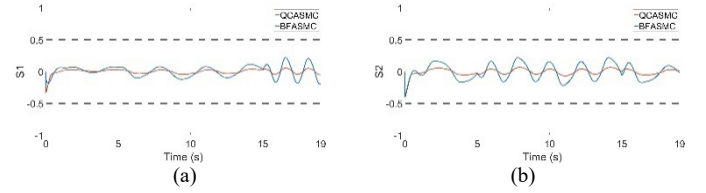


Fig. 2 Sliding surface of BFASMC and QCASMC. (a) joint 1. (b) joint 2.

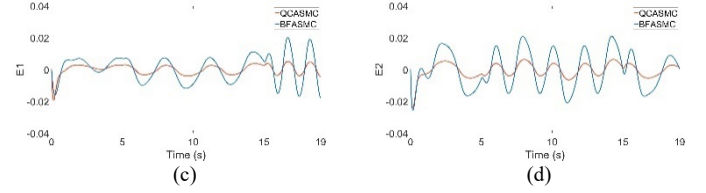


Fig. 3 Tracking error of BFASMC and QCASMC. (c) joint 3. (d) joint 4.

tracking errors of BFASMC and QCASMC. The QCASMC keeps the sliding surface and tracking error arbitrarily small in the vicinity of zero against the time-varying disturbances. The results of the simulation show that the QCASMC is more robust and stable than the BFASMC.

### IV. EXPERIMENT

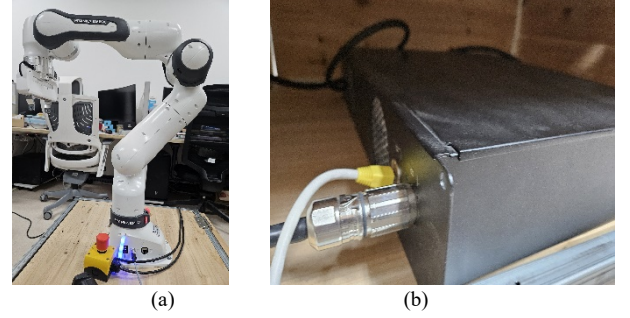


Fig. 4 Experiment setup based on Franka Emika Research 3. (a) manipulator. (b) control box

In this section, we present the experiment results of Franka Emika research 3, the 7-DOF robot manipulator in Fig. 3. The experiment utilized 4-DOF joints for position control.

#### A. Experiment Setup

The Reserch3 has following parameters: a mass of 17.8 kg, and a payload of 3 kg. The main frequency is 50~60Hz, and the sampling rate is 100Hz. Link-side torque sensors are present on the all 7 axes. For the experiments, we designed the following

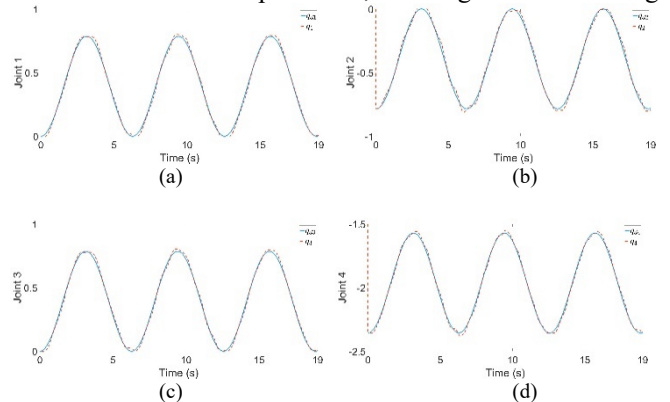


Fig. 5 Desired motion and trajectory of QCASMC with TDC. (a) joint 1. (b) joint 2. (c) joint 3. (d) joint 4.

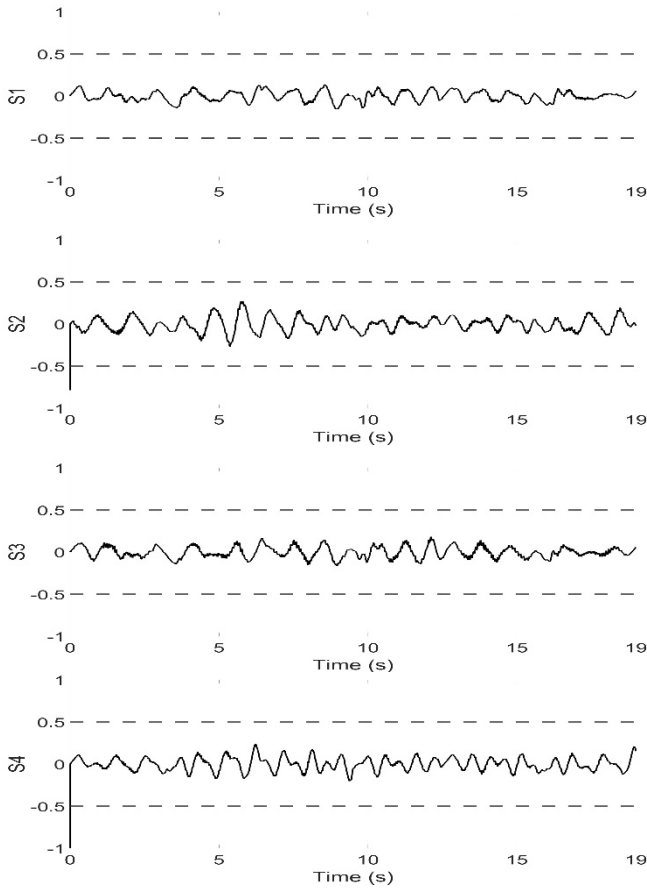


Fig. 6 Sliding surface of QCASMC with TDC. each joint of  $q_1, q_2, q_3, q_4$ .

parameters to  $q_d = [q_{1d}, q_{2d}, q_{3d}, q_{4d}]^T$ ,  $q_{1d} = \frac{\pi}{8} \sin(t)$ ,  $q_{2d} = \frac{\pi}{8} \sin(t)$ ,  $q_{3d} = \frac{\pi}{8} \sin(t)$ ,  $q_{4d} = \frac{\pi}{8} \sin(t)$ ,  $\alpha = 3$ ,  $\beta = 20$  and  $\gamma = 0.3$ ,  $\epsilon = 0.5$ ,  $K_1 = \text{diag}\{1, 1, 1, 1\}$ ,  $K_2 = \text{diag}\{0.03, 0.03, 0.03, 0.03\}$ ,  $\bar{M} = \{0.02, 0.02, 0.02, 0.02, 0.02\}$ .

### B. Experiment Results

The experimental results in Figures 4 and 5 show the sliding surface and tracking error for each joint of the QCASMC with TDC. Both results show convergence close to zero, and the desired motion and trajectory results in Fig. 6 demonstrate the stability and robustness of the proposed adaptive law and algorithm.

## V. CONCLUSIONS

This paper proposed a new ASMC and combines TDC techniques, applying them to a robot manipulator through simulation and experiment. The previous TDE error is used to compensate for the TDE error, and no modeling information is required. The proposed quasi-convex function-based ASMC reduces overestimation, and the sliding surface and tracking error converge to near zero, showing stability and robustness. For stability analysis, the tracking error of the proposed ASMC is combined with the TDE technique to ensure the stability of the UUB for sliding variables using the Lyapunov method. Simulations and experiments demonstrate that the proposed ASMC could be a viable alternative to traditional ASMC. In the future work, good estimation performance can be achieved by adjusting the  $\beta$  parameter of the quasi-convex function, adding

additional terms, and adjusting the parameters of the TDC algorithm.

## ACKNOWLEDGMENT

This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the ICAN(ICT Challenge and Advanced Network of HRD) support program(IITP-2024-2020-0-01832) supervised by the IITP(Institute for Information & Communications Technology Planning & Evaluation)

## REFERENCES

- [1] CHEBAB, Z. E., et al. Autonomous collaborative mobile manipulators: State of the art. In: *Symposium on Theory of Machines and Mechanisms/UMTS2015/TrISToMM*. 2015.
- [2] MITCHELL, Ben, et al. Development and application of a new steady-hand manipulator for retinal surgery. In: *Proceedings 2007 IEEE International Conference on Robotics and Automation*. IEEE, 2007. p. 623-629
- [3] WEINRIB, Harry P.; COOK, John Querin. Rotational technique and microsurgery. *Microsurgery*, 1984, 5.4: 207-212
- [4] BAEK, Jaemin; JIN, Maolin; HAN, Soohee. A new adaptive sliding-mode control scheme for application to robot manipulators. *IEEE Transactions on industrial electronics*, 2016, 63.6: 3628-3637
- [5] S. Islam and P. X. Liu, "Robust sliding mode control for robot manipulators," *IEEE Trans. Ind. Electron.*, vol. 58, no. 6, pp. 2444–2453, Jun. 2011.
- [6] Huang, Y.J, Kuo, T.C, and Chang, S.H, "Adaptive sliding mode control for nonlinear systems with uncertain parameter" *IEEE Transactions on System, Man, and Cybernetics – Part B: Cybernetics*, vol. 38, no.2, pp. 534-539, April. 2008.
- [7] F. Plestan, Y. Shtessel, V. Bré'geault and A. Poznyak, "New methodologies for adaptive sliding mode control" *International Journal of Control*, vol. 83, no. 9, pp. 1907-1919, Sep. 2010.
- [8] ROY, Spandan, et al. Overcoming the underestimation and overestimation problems in adaptive sliding mode control. *IEEE/ASME Transactions on Mechatronics*, 2019, 24.5: 2031-2039.
- [9] BANDYOPADHYAY, Bijan, et al. Advances in sliding mode control. *Lecture Notes in Control and Information Sciences*, 2013, 440.
- [10] Hussein Obeid, Leonid M. Fridman, Salah Laghrouche, and Mohamed Harmouche, "Barrier function-based adaptive sliding mode control" *Automatica*, vol. 93, pp. 540-544, July. 2018.
- [11] S. Dempe, N. Gadhi, and K. Hamdaoui, "Minimizing the difference of two quasiconvex functions" *Optimization Letters*, vol. 14, pp. 1765-1779, 2020.
- [12] JunMin Park, Woogyong Kwon and PooGyeon Park, "An Improved Adaptive Sliding Mode Control Based on Time-Delay Control for Robot Manipulators" *IEEE Transactions on Industrial Electronics*, vol. 70, no. 10, pp. 10363-10373, November 2022.
- [13] K. Youcef-Toumi and O. Ito, "A time delay controller for systems with unknown dynamics," *Trans. ASME, J. Dyn. Syst. Meas. Control*, vol. 112, no. 1, pp. 133–142, 1990.
- [14] T. S. Hsia et al., "Robust independent joint controller design for industrial robot manipulators," *IEEE Trans. Ind. Electron.*, vol. 38, no. 1, pp. 21–25, Feb. 1991.
- [15] M. Jin and P. H. Chang, "Simple robust technique using time delay estimation for the control and synchronization of Lorenz systems," *Chaos Solitons Fractals*, vol. 41, no. 5, pp. 2672–2680, 2009.
- [16] M. Jin, Y. Jin, P. H. Chang, and C. Choi, "High-accuracy tracking control of robot manipulators using time delay estimation and terminal sliding mode," *Int. J. Adv. Robot. Syst.*, vol. 8, no. 4, pp. 65–78, 2011.
- [17] Y.-X. Wang, D.-H. Yu, and Y.-B. Kim, "Robust time-delay control for the DC–DC boost converter," *IEEE Trans. Ind. Electron.*, vol. 61, no. 9, pp. 4829–4837, Sep. 2014.
- [18] M. Jin, J. Lee, and K. K. Ahn, "Continuous nonsingular terminal slidingmode control of shape memory alloy actuators using time delay estimation," *IEEE/ASME Trans. Mechatronics*, vol. 20, no. 2, pp. 899–909, Apr. 2015.
- [19] S.-j. Cho, M. Jin, T.-Y. Kuc, and J. S. Lee, "Control and synchronization of chaos systems using time-delay estimation and supervising switching control," *Nonlinear Dynam.*, vol. 75, no. 3, pp. 549–560, 2014.

# Channel Estimation Method Based on K-means Algorithm using optimal resource clustering.

Gayeon Kim<sup>1</sup>, Daegun Jang<sup>2</sup>, Yumin Kim<sup>3</sup>, and Byeong-Gwon Kang<sup>\*</sup>

<sup>1,2</sup>ICT Convergence, Soonchunhyang University, Asan, Korea

<sup>3</sup>, <sup>\*</sup>Information and Communication Engineering, Soonchunhyang University, Asan, Korea

<sup>\*</sup>Contact: First.gayeon17@sch.ac.kr, phone +82 10-3646-5917.

**Abstract**— The next-generation mobile communication service, 5G New Radio (NR), imposes high-level requirements to achieve high reliability, ultra-low latency, and high connection density. Additionally, the 3rd Generation Partnership Project (3GPP) is engaged in various standardization efforts to enhance 5G services further. Channel estimation stands as a core technology for improving signal quality, with demodulation reference signal (DM-RS) based channel estimation being adopted as the standard technique. While this approach effectively adapts to the dynamically changing channel environment by sharing wireless resources with data information, it may incur losses in terms of data transmission rates due to resource occupancy. To address this issue, a channel estimation method based on the K-means clustering algorithm has been introduced. However, this method had limitations in wireless resource unit channel estimation by restricting the minimum target of channel estimation to resource blocks (RBs). This paper proposes a blind channel estimation method based on the K-means algorithm, which is effective even in large delay spread and Doppler affected environments, by setting the minimum target of clustering to a resource element.

## I. INTRODUCTION

3GPP provides three services as use cases for the 5G new radio (NR) mobile communication service, 5G NR. First, enhanced mobile broadband (eMBB), massive machine-type communications (mMTC), and ultra-reliable and low latency communications (URLLC) [1]. These services have led to various standardization efforts due to the use of higher frequency bands compared to Long-Term Evolution (LTE), and currently, more enhanced 5G-Advanced (5G-A) standardization is underway.

Channel estimation in wireless communication systems is a key technology for improving signal quality, and currently, channel estimation based on demodulation reference signal (DM-RS) is adopted as a standard technique. While DM-RS effectively adapts to real-time channel variations allocated in wireless resources along with data information in the physical layer, it may incur a loss in data transmission rate depending on the number of DM-RS symbols occupying wireless resources.

To address the issue a blind channel estimation technique based on the K-means clustering algorithm has been proposed [2]. This method to perform channel estimation without occupying wireless resources with DM-RS and has the advantage of efficiently managing wireless resources. When performing channel estimation in [2], a single resource block (RB) is used as the minimum unit for clustering. However, this configuration has limitations in effectively addressing channel

variations caused by high levels of delay spread (DS) and Doppler effects, resulting in performance degradation.

This paper proposes a more effective optimal resource clustering based blind channel estimation algorithm that does not rely on DM-RS in wireless fading environments. The proposed method performs clustering with multiple orthogonal frequency division multiplexing (OFDM) symbols in the time domain and multiple subcarriers in the frequency domain as the minimum units. Each distinguished cluster undergoes channel estimation based on the K-means clustering algorithm, and the utility of the proposed method is examined through mean square error (MSE) performance analysis and data rate measurement.

## II. SYSTEM MODEL

In the 5G NR physical layer, various physical channels and signals are defined for different transmissions. These physical channels and signals are utilized for uplink and downlink transmissions, conveying control information, data packets, channel state information, synchronization information, and more. The aforementioned physical channels and signals are processed based on scheduling information and allocated to the resource grid to pass through the wireless fading channel. When digital modulated signals are assigned to the resource grid, they can be composed of single or multiple slots in the time domain and single or multiple RBs in the frequency domain [3]. The received signal  $Y_{i,j}$  passing through the wireless fading channel mentioned above can be represented as  $Y_{i,j} = H_{i,j}X_{i,j} + N_{i,j}$ , where  $H_{i,j}$  is tapped-delay line (TDL)-A channel,  $X_{i,j}$  is transmission signal, and  $N_{i,j}$  is additive white gaussian noise (AWGN) with mean 0 and variance 1.

## III. PROPOSED METHOD

The received signal  $Y_{i,j}$  passing through the TDL-A channel undergoes OFDM demodulation, followed by clustering into multiple groups considering DS and user equipment (UE) velocity [4]. Each group consists of multiple resource elements (REs), composed of single or multiple OFDM symbols in the time domain and single or multiple subcarriers in the frequency domain. All clusters need to effectively adapt to channel variations due to DS and UE velocity-induced Doppler effects. To achieve this, each cluster determines its clustering groups by considering combinations of factors that are divisors of the number of OFDM symbols associated with the signal within



slots on the time domain and the number of subcarriers on the frequency domain. For example, assuming clustering for 1 slot in the time domain and 1 RB in the frequency domain, the candidate groups for clustering on the time domain  $C_t = \{1, 2, 7, 14\}$  and on the frequency domain  $C_f = \{1, 2, 3, 4, 6, 12\}$ . The K-means algorithm is applied to all combinations of candidates on the time and frequency domain. However, considering QPSK as the assumed digital modulation in this study, clustering groups of fewer than four are excluded from the combination of  $C_t$  and  $C_f$ . The REs within each clustering group is distinguished as one of the four QPSK symbols, based on as

$$X_{i,j} = e^{\frac{j(2l-1)\pi}{4}}, \quad (1)$$

where  $l = (1, \dots, 4)$  is number of QPSK symbols. Based on (1), channel estimation is performed for each clustering, where the received signal  $Y_{i,j}$  is QPSK modulated, resulting in four centroids for each cluster. Through this process, signals within each cluster are discriminated as centroids, which are QPSK symbols, effectively canceling out noise. The estimated channel  $\hat{H}_{i,j} = Y_{i,j} / X_{i,j}$  is computed for each centroid, allowing for four channel estimations for one centroid. Then,  $\hat{X}_{i,j} = Y_{i,j} / \hat{H}_{i,j}$  is derived for each of the 16 estimated channels, which is used to calculate the bit error rate (BER) by comparing it with the actual transmitted signal. The estimated channel  $\hat{H}_{i,j}$  with the lowest BER in each cluster is considered as the estimated channel for that cluster. This process is repeated for all clusters, and the performance is analyzed using the MSE between  $\hat{H}_{i,j}$  and  $H_{i,j}$ . MSE is calculated as

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n \left( \frac{1}{m} \sum_{j=1}^m |H_{i,j} - \hat{H}_{i,j}|^2 \right), \quad (2)$$

where  $n$  is number of clustering groups and  $m$  is number of associations.

#### IV. SIMULATION RESULTS

In this chapter, we describe the MSE performance of the proposed method compared to the conventional method. It is assumed that the transmitted signal passed through a TDL-A channel, occupying 1 slot in the time domain and 10 RBs in the frequency domain. Fig. 2 illustrates the MSE performance of the conventional method represented by the red line and the proposed method represented by the blue line, with a DS of 50 ns and a UE speed of 60 km/h. Both methods demonstrate a decrease in MSE with an increase in the signal-to-noise ratio (SNR), with the proposed method exhibiting superior performance over the entire SNR range compared to the conventional method. Fig. 3 depicts the MSE performance under conditions of a UE speed of 120 km/h and a DS of 300 ns. In Fig. 3, significant channel variations can be observed in both the time and frequency domains compared to Fig. 2. The proposed method, by performing optimal resource clustering to overcome the limitations of the conventional method and conducting channel estimation at the RE level, demonstrates more effective performance in Fig. 3.

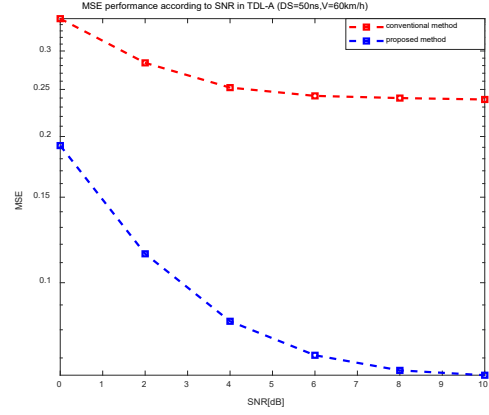


Fig. 2 MSE performance according to SNR (DS=50ns, V=60km/h).

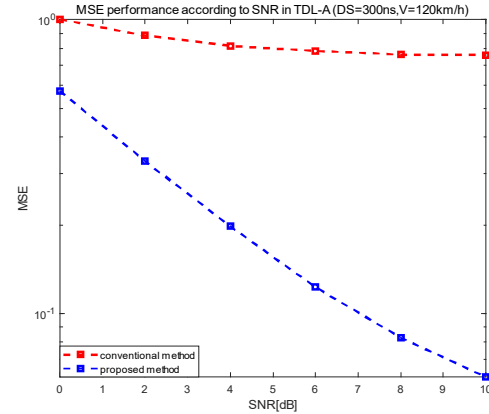


Fig. 3 MSE performance according to SNR (DS=300ns, V=120km/h).

#### V. CONCLUSION

In this paper, we introduce a channel estimation method based on the K-means algorithm with optimal resource clustering to overcome the limitations of conventional methods. Through simulations, we confirm the effectiveness of the proposed method in performing channel estimation even in scenarios with significant channel variations due to high DS and Doppler effects. Additionally, the analysis of simulation results highlights the necessity of optimal cluster configuration according to channel characteristics.

#### ACKNOWLEDGMENT

This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the ICAN(ICT Challenge and Advanced Network of HRD) support program(IITP-2024-2020-0-01832) supervised by the IITP(Institute for Information & Communications Technology Planning & Evaluation)

#### REFERENCES

- [1] Y.kim, Y.kim, J. Oh, H. Ji, J. Yeo, S. Choi, H. Ryu, H. Noh, T. Kim, F. Sun, Y. Wang, Y. Qi, and J. Lee, "New Radio (NR) and Its Evolution Toward 5G-Advanced," IEEE Wireless Communications, vol. 26, no. 3, pp. 2-7, Jun. 2019.
- [2] Hanho Wang, "Performance Evaluation of Channel Estimation Scheme Using Clustering Algorithm," KIIT, vol. 16, no. 2, pp. 61-66, 2018.
- [3] NR; Physical channels and modulation, V17.2.0, Release 17, 3GPP Standard TS 38.211, Jun. 2022.
- [4] TSG RAN; Study on channel model for frequencies from 0.5 to 100GHz, V17.0.0, Release 16, 3GPP TR 38.901, Dec. 2019

# Human Interaction Recognition Through Deep Learning Technique Based on Video from CCTV Cameras

Vesal Khean<sup>1,\*</sup>, Chomyong Kim<sup>2</sup>, and Yunyoung Nam<sup>3</sup>

<sup>2</sup>*ICT Convergence Research Centre, Soonchunhyang University, Asan, South Korea*

<sup>3</sup>*Department of Computer Science and Engineering, Soonchunhyang University, Asan, South Korea*

\*Contact: ynam@sch.ac.kr

**Abstract**— Human Interaction Recognition (HIR) is a field of study that involves the development of computer algorithms to detect and recognize human interactions in videos, images, or other multimedia content. The goal of HIR is to automatically identify and analyze the social interactions between people, their body language, and facial expressions. In this paper, we aim to address the problem of human interaction recognition in videos by exploring the long-term inter-related dynamics among multiple persons. Recently, Long Short-Term Memory (LSTM) has become a popular choice to model individual dynamic for single-person action recognition due to its ability of capturing the temporal motion information in a range. However, existing RNN models focus only on capturing the dynamics of human interaction by simply combining all dynamics of individuals or modeling them as a whole. Such models neglect the inter-related dynamics of how human interactions change over time. To this end, we propose a Long Short-Term Memory model to conducted the experiment with human interaction dataset. The aim is to recognize interactions between two individuals in a video using OpenPose framework to extract key points, and subsequently employ a LSTM model for interaction recognition.

## I. INTRODUCTION

Computer vision is an area of artificial intelligence that focuses on enabling computers to interpret and understand visual data from the world around them. It has become increasingly important in recent years as the amount of visual data generated by digital devices continues to grow faster. (CV) has numerous application scenarios, from facial recognition [1-2] and object detection to medical image analysis [3-5] and autonomous vehicles [6], [9], and recommendation systems [7-8].

The goal of (CV) is to enable machines to extract meaningful information from visual data and use that information to make decisions or take actions. This is accomplished through the use of algorithms and machine learning techniques, which enable computers to recognize patterns and features within images or videos. Some of the key techniques used in computer vision include interaction recognition [10-12], video classification [13-16], human action classification [17-19], and object detection [20-23]. As the field of (CV) has advanced, it has become possible to

develop more sophisticated and accurate algorithms for analyzing visual data. This has led to the development of human-to-human interaction, which is how humans perceive their environment through their five senses. Human-machine-human (HMH) interaction is when humans interact with each other through machines, such as robots or passive sensorized devices [24, 25]. In particular, machine learning (ML) techniques have been used to develop human-human interaction (HHI) systems for identifying human behavior [26], which provide powerful effects in human interaction recognition (HIR) during routine daily life.

Despite the advances in computer vision (CV), machine learning (ML), and deep learning (DL), there remain challenges in achieving high accuracy and reliability in the process of recognizing and classifying human actions or interactions with videos. These challenges include issues related to handling diverse datasets, refining interaction boundaries, and ensuring the robustness of classification models. One of the main challenges is the extraction of human key points within interaction videos, such as light-conditioning videos and camera angles that can overlap people in indoor and outdoor environments. Accurate extraction of key points is really important for the feature. Achieving a generalized approach to interaction classification remains paramount. The quality of this classification directly influences subsequent tasks such as target identification, feature extraction, and pattern recognition. By ensuring the broader applicability of interaction classification methods, we empower the extraction of relevant features for object-person and boundary classification within real-time videos. Ongoing research and development efforts strive to enhance precision and adaptability, enabling robust performance across diverse contexts.

## II. RELATED WORK

In recent years, significant advancements have been made in deep learning-based approaches for human interaction recognition. The study is to develop a multi-stream sequence learning framework for human interaction recognition that effectively combines skeleton key points and spatiotemporal visual representations to accurately classify various human interactions in surveillance videos. Utilizes two benchmark datasets, namely the UT-interaction (UT-I) dataset and the TV



human interaction (TV-HI) dataset, to evaluate the proposed multi-stream network. These datasets contain video samples of human interactions captured in different scenarios and environments. For method consists of a multi-stream network that includes two main blocks: a 1-D CNN with BD-LSTM and a 3-D CNN. The 1-D CNN and BD-LSTM stream learns human interactions based on key features extracted from pose estimation algorithms, while the 3-D CNN model captures temporal information. The outputs of these streams are concatenated to make the final prediction, enabling the model to effectively recognize interactions between multiple humans from video frames. The overall accuracy of the proposed model using the combined dataset was 96.0% [26]. The research work namely Human Interaction Recognition Based on Deep Learning and HMM, objective of the paper is to improve human interaction recognition accuracy by combining deep learning with traditional HMM methods. Dataset used in the experiments is the UT-interaction dataset, which includes six types of double interaction behaviours such as handshake, hug, kick, provocation, shoving, and punching. The method involves using an optimized ALEXNet convolutional neural network to extract behaviour features, training an LSTM network with Softmax method for feature extraction, and fusing classification results using the particle swarm optimization algorithm to establish a hybrid classification model. The results show that the hybrid model achieves higher recognition accuracy compared to other classical methods, with a recognition rate of 91.9% on the UT-interaction dataset [27]. Separately, this research namely Human Interaction Classification in Sliding Video Windows Using Skeleton Data Tracking and Feature Extraction. This research presents an approach for classifying human interactions in video frames using skeleton data. By combining knowledge-aware feature extraction, multi-stream neural network models, and sliding window processing, the method achieves high accuracy in interaction recognition on the NTU RGB+D dataset. The study demonstrates the effectiveness of the approach through cross-validation on the UT-Interaction dataset, showcasing robust performance and generalizability. The models developed offer a practical trade-off between accuracy and complexity, highlighting advancements in human interaction classification using deep learning techniques [28]. As for this last paper, the research introduces the task of Human-to-Human Interaction Detection (HID) in videos, aiming to detect subjects, recognize person-wise actions, and group people based on their interactive relations. The study creates the AVA-Interaction (AVA-I) dataset, a large-scale benchmark for HID, and develops the SaMFormer model, a Transformer-based framework for one-stage HID. SaMFormer achieves leading performance on existing benchmarks and the AVA-I dataset, demonstrating superior accuracy in detecting human interactions and interpreting scenes effectively [29].

### III. PROPOSED METHOD

In this section, we delve into the intricacies of the proposed model employed for executing the interaction recognition task.

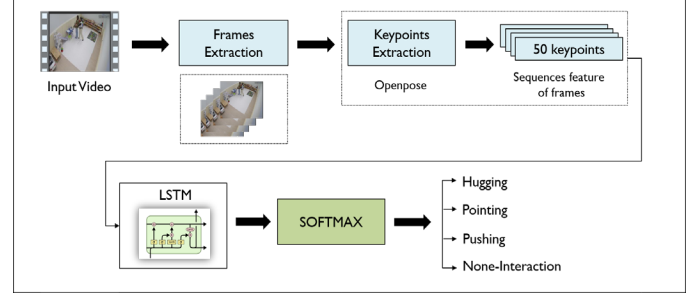


Fig. 1 The proposed whole flowchart of human interaction recognition task

#### A. Frame Extraction

The initial steps involve the ingestion of video data followed by frame extraction, which serves as the foundational preprocessing phase. This process enables the isolation of individual frames from the video, effectively breaking down the visual input into discrete units for subsequent analysis.

#### B. Key Points Feature Extraction

In this section, our proposed technique, interaction, like other actions, is viewed as sequential data unfolding over time. Instead of analyzing all frames from the videos, we selected 51 frames from each video. This approach aims to decrease computational load and processing time while still capturing the fundamental motion dynamics present in the video data. To further reduce computational demands during the pose estimation process, frames were resized to 800 pixels in width and 500 pixels in height. Pose estimation, executed using OpenPose, enabled the extraction of 2D joint locations of the skeleton. OpenPose identified 25 key points on the human body, each represented by 2D X and Y coordinates. Refer to Fig. 2 for an illustration of the 25 key points extracted by OpenPose from the human body.

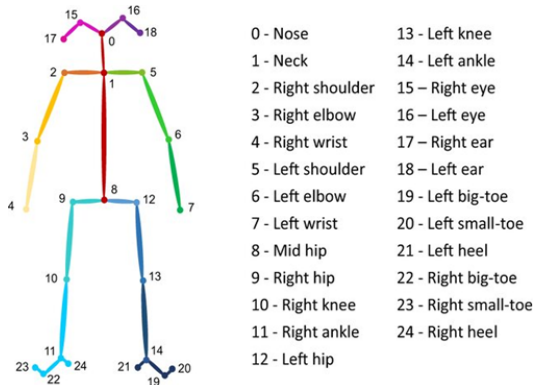


Fig. 2 Openpose detects human body 25 key points

The output from Openpose is an array with a shape of 50x3 for detected two objects in a frame; 50 represents the number of key points for detected two persons, and 3 represents the X-

coordinate, Y-coordinate, and Z-confidence score. We do not utilize the Z-confidence score; it is removed from the array. Therefore, the frames become an array with a shape of  $50 \times 2$ , where X values range from 0 to 800, corresponding to the width of the frame, and Y values range from 0 to 500, corresponding to the height of our frame. The extracted key points for each frame are stacked in order to construct sequential data. Each sequential data has a shape of  $51 \times 50 \times 2$ , where 51 denotes the number of frames, 50 signifies the number of key points detected by two people in each frame, and 2 represents the X and Y coordinates of each key point. We also described that, if there are less than 50 key points generated for each human, assign zero value to missing joints. It did not affect the accuracy of our proposed method because the movement and change in each skeleton key point are already described by the interaction patterns of both humans.

### C. Model Design Structure

In the model implementation phase, a Long Short-Term Memory (LSTM) architecture is employed to analyze the temporal dependencies within the extracted key points sequences. LSTM networks are well-suited for capturing long-range dependencies in sequential data, making them ideal for modeling complex temporal patterns in human interactions. By leveraging the memory cells and gating mechanisms inherent in LSTM units, the model can effectively learn and represent the intricate temporal dynamics of human interactions. Finally, the LSTM model is coupled with a softmax activation function, which serves as the output layer's classifier. This softmax classifier assigns probabilities to each interaction class, such as hugging, pointing, pushing, and none-interaction, based on the model's learned representations. By applying the softmax function, the model can make probabilistic predictions about the presence of different interaction types, enabling accurate classification of human interactions in real-world scenarios.

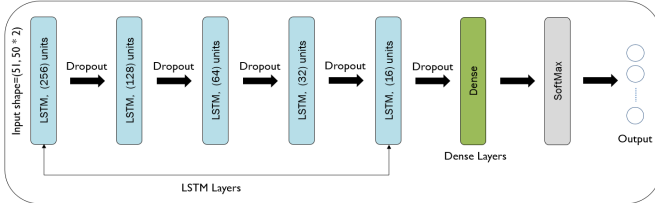


Fig. 3 The LSTM model designed for our proposed on the (HHI)

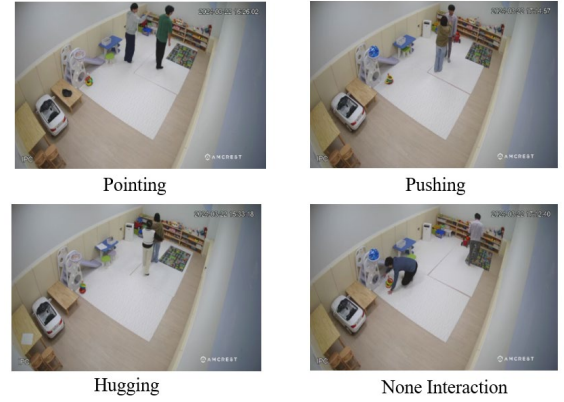
In our research, in the Fig. 3, the model architecture comprises an input layer implicitly defined through the input\_shape parameter of the initial LSTM layer, where input shape =  $(51, 50 * 2)$  specifies the data shape with 51 frames and  $50 * 2$  features per frame, each representing two coordinates (x, y) of key points. Stacked LSTM layers follow, with the first layer containing 256 units and subsequent layers with decreasing units (128, 64, 32, and 16). Setting return\_sequences=true ensures each LSTM layer provides the full sequence of outputs. Dropout layers are interspersed after each LSTM layer, with a dropout rate of 0.2 for the initial four layers and 0.1 for the final one, mitigating overfitting by randomly zeroing input units. The model culminates in a dense layer with a softmax activation function, featuring a

number of units equivalent to the classes in the dataset, facilitating human interaction recognition. The softmax function in the equation below is used to get the probability for the output layer.

## IV. RESULTS

### A. Dataset

In this study, thirteen healthy male and female adults participated in the interaction for durations exceeding three hours. The material for collecting data is four-channel cameras. During the data collection, we employed the participants, who took the interaction of four classes. For one interaction, the duration is 4 minutes, and we repeated it for all classes, so the total is 20 minutes for one interaction. In addition, each interaction has characteristic clues such as front hugging, back hugging, front pushing, back pushing, front pointing, and back pointing, and the last one is none interaction (two persons



focused on their respective objects)

Fig. 4 The human interaction dataset with four classes

The human interaction dataset is composed of 550 videos with ground truth labels, as shown in Table. 1, encompassing a diverse range of human interaction. The dataset provides four classes, namely Hugging, Pointing, Pushing, and None Interaction; each class has a total of 110 videos; each video duration is 1.7 seconds; 30 frames per second; frame resolution is  $1582 \times 1080$ ; and the combined four-channel cameras. This dataset is dedicated to human interaction analysis and has been instrumental in the development and evaluation of our proposed method, which is specially recognized for human interaction tasks.

Table. 1 The detail of dataset collecting information

Class Name	Video File	Duration	FPS	Frame Resolution	Camera Name
Hugging	110	1.7 seconds for each action	30	$1582 \times 1080$	CCTV 4 channel Cameras
Pushing	110				
Pointing	110				
None-Interaction	110				
Total	550				

### B. Data Division

Data division is a fundamental step in the preparation of datasets for machine learning and deep learning tasks. In this study, the dataset, comprised by combining four channel

cameras together, is split into two main subsets: 80% for training, which is divided into 385 videos, and 20% for testing, as shown in Table. 2. For the human interaction dataset, the testing set was divided into 165 videos, as shown in Table. 2. This division is performed to facilitate the development and evaluation of machine learning models.

Table. 2 The amount of training and testing set of our dataset

Dataset	Training set	Testing set
Human Interaction	385	165

### C. Experimental Setting

In setting up our classification model training, we employed the LSTM method, utilizing tools like Tensorflow, Keras, OpenCV, Numpy, and Scikit-Learn, all within Python programming. The operating system for human interaction recognition training was configured for training on a GPU on a PC server that has a GPU NVIDIA TITAN RTX), CPUs (Intel(R) Xeon(R) Silver 4114 CPU @ 2.20 GHz, 2 processors), and RAM of 192 GB.

### D. Experimental Result

The results of applying the proposed Long Short-Term Memory (LSTM) model to the human interaction dataset are presented below. The implementation of the LSTM method aimed to validate its performance in classifying human interactions. The obtained results demonstrate a maximum accuracy of 94.12%. Additionally, other performance measures such as recall rate, precision rate, and F1 score have a computational time of 2 sec and 11 ms, respectively. These results are summarized in Table. 3. Moreover, the confusion matrix depicting the classification outcomes of the human interaction dataset is illustrated in Fig. 5.

Table. 3 The result on the human interaction dataset with our model

Class Name	Precision	Recall	F1-score
Hugging	1.00	0.90	0.95
Pushing	0.83	1.00	0.90
Pointing	1.00	0.89	0.94
None interaction	0.95	1.00	0.98
Overall accuracy	94.12%		

In evaluating the performance of the LSTM model on the human interaction dataset as shown in Fig. 5, we examined the confusion matrix to gain insights into the model's classification outcomes. The confusion matrix provides a detailed breakdown of the model's predictions, showing the number of true positive, false positive, true negative, and false negative classifications for each class of human interaction. By analyzing the confusion matrix, we can assess the model's ability to correctly classify different types of human interactions and identify any areas where the model may be struggling.

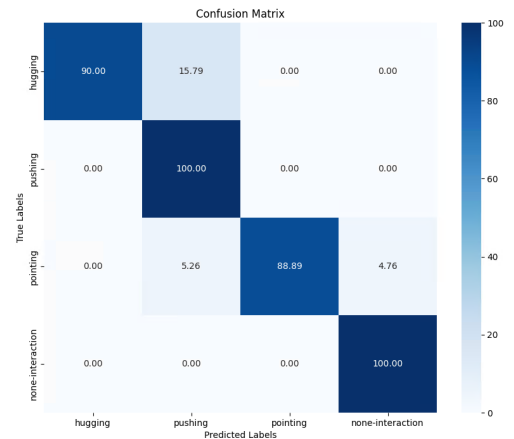


Fig. 5 The confusion matrix on the human interaction with our model

## V. CONCLUSIONS

In conclusion, our thesis represents a significant contribution to the part of human interaction analysis from the videos in the computer vision field. We have introduced a highly effective deep learning model proposed for multiple human interaction recognition tasks, specifically applied to the human interaction dataset. Our model has exhibited robustness in accuracy and precision, which demonstrates our proposed model's suitability for the classification. We effectively utilized deep learning architectures, including LSTM approach; furthermore, we have conducted a rigorous comparative analysis with existing works and consistently shown our model's superiority, setting a benchmark in the domain.

Our research has revealed important possibilities for using deep learning to recognize multiple types of human interaction. By showing how powerful it can be, we've created a strong foundation for accurately recognizing each type of interaction. Furthermore, in our examination of human interaction videos capturing real human interactions within many participants and recorded using 4 camera channels, we used all camera angles to make our model more accurate. Looking toward the future, we see promising opportunities for extending our research. We plan to broaden the scope of our model to perform with large datasets by combining many conditions of recording dataset environments, such as indoor and outdoor environments, to offer a more comprehensive and delicate approach to real-time human interaction detection and classification. This extension will involve accommodating multiple classes and distinguishing multiple people in the videos, ultimately facilitating advanced and detailed recognition in intelligent video surveillance and human-computer interaction fields.

## ACKNOWLEDGMENT

This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the ICAN(ICT Challenge and Advanced Network of HRD) support program(IITP-2024-2020-0-01832) supervised by the IITP(Institute for Information & Communications Technology Planning & Evaluation)

# REFERENCES

- [1] M. K. Rusia and D. K. Singh, "A comprehensive survey on techniques to handle face identity threats: challenges and opportunities," *Multimedia Tools and Applications*, vol. 82, no. 2, pp. 1669-1748, 2023. doi:10.1007/s11042-022-13248-6.
- [2] M. Sajjad, F. U. M. Ullah, M. Ullah, G. Christodoulou, F. A. Cheikh, M. Hiji, K. Muhammad, and J. J.P.C. Rodrigues, "A comprehensive survey on deep facial expression recognition: challenges, applications, and future guidelines," *Alexandria Engineering Journal*, vol. 68, pp. 817-840, 2023. doi: 10.1016/j.aej.2023.01.017.
- [3] D. Hu, Shuai Li, and Mengjun Wang, "Object detection in hospital facilities: A comprehensive dataset and performance evaluation," *Engineering Applications of Artificial Intelligence*, vol. 123, pp. 106223, 2023. doi: 10.1016/j.engappai.2023.106223.
- [4] R. Kaur and S. Singh, "A comprehensive review of object detection with deep learning," *Digital Signal Processing*, vol. 132, pp. 103812, 2023. doi: 10.1016/j.dsp.2022.103812.
- [5] Z. Lone and A. R. Pais, "Object detection in hyperspectral images," *Digit Signal Processing*, vol. 131, pp. 103752, 2022. doi: 10.1016/j.dsp.2022.103752.
- [6] F. Liu, F. Xue, W. Wang, W. Su, and Y. Liu, "Real-time comprehensive driving ability evaluation algorithm for intelligent assisted driving," *Green Energy and Intelligent Transportation*, vol. 2, pp. 100065, 2023. doi: 10.1016/j.geits.2023.100065.
- [7] D. Roy and M. Dutta, "A systematic review and research perspective on recommender systems," *Journal of Big Data*, vol. 9, no. 1, pp. 59, 2022. doi: 10.1186/s40537-022-00592-5.
- [8] K. Tiwari and D. Kumar Singh, "Machine learning-based recommendation system for disease-drug material and adverse drug reaction: Comparative review," *Materials Today: Proceedings*, vol. 51, pp. 304-313, 2022. doi: 10.1016/j.matpr.2021.05.404.
- [9] Z. Bao, S. Hossain, H. Lang, and X. Lin, "A review of high-definition map creation methods for autonomous driving," *Engineering Applications of Artificial Intelligence*, vol. 12, pp. 106125, 2022. doi: 10.1016/j.engappai.2023.106125.
- [10] Shafiqul, Islam Md, Mir Kanon Ara Jannat, Jin-Woo Kim, Soo-Wook Lee, and Sung-Hyun Yang, "Hhi-attentionnet: An enhanced human-human interaction recognition method based on a lightweight deep learning model with attention network from csi," *Sensors* 22, no. 16 (2022): 6018.
- [11] Ko, Woo-Ri, Minsu Jang, Jaeyeon Lee, and Jaehong Kim. "AIR-Act2Act: Human-human interaction dataset for teaching non-verbal social behaviors to robots." *The International Journal of Robotics Research* 40, no. 4-5 (2021): 691-697.
- [12] Ouyed, Ouiza, and Mohand Said Allili. "Group-of-features relevance in multinomial kernel logistic regression and application to human interaction recognition." *Expert systems with applications* 148 (2020): 113247.
- [13] Pareek, Preksha, and Ankit Thakkar. "A survey on video-based human action recognition: recent updates, datasets, challenges, and applications." *Artificial Intelligence Review* 54, no. 3 (2021): 2259-2322.
- [14] Jaouedi, Neziha, Nouredine Boujnah, and Med Salim Bouhlel. "A new hybrid deep learning model for human action recognition." *Journal of King Saud University-Computer and Information Sciences* 32, no. 4 (2020): 447-453.
- [15] Zhu, Linchao, Du Tran, Laura Sevilla-Lara, Yi Yang, Matt Feiszli, and Heng Wang. "Faster recurrent networks for efficient video classification." In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, pp. 13098-13105. 2020.
- [16] Islam, Md Shofiqul, Shanjida Sultana, Uttam Kumar Roy, and Jubayer Al Mahmud. "A review on video classification with methods, findings, performance, challenges, limitations and future work." *Jurnal Ilmiah Teknik Elektro Komputer dan Informatika* 6, no. 2 (2020): 47-57.
- [17] Sun, Zehua, Qihong Ke, Hossein Rahmani, Mohammed Bennamoun, Gang Wang, and Jun Liu. "Human action recognition from various data modalities: A review." *IEEE transactions on pattern analysis and machine intelligence* 45, no. 3 (2022): 3200-3225.
- [18] Kong, Yu, and Yun Fu. "Human action recognition and prediction: A survey." *International Journal of Computer Vision* 130, no. 5 (2022): 1366-1401.
- [19] Muhammad, Khan, Amin Ullah, Ali Shariq Imran, Muhammad Sajjad, Mustafa Servet Kiran, Giovanna Sannino, and Victor Hugo C. de Albuquerque. "Human action recognition using attention based LSTM network with dilated CNN features." *Future Generation Computer Systems* 125 (2021): 820-830.
- [20] Xiao, Youzi, Zhiqiang Tian, Jiachen Yu, Yinshu Zhang, Shuai Liu, Shaoyi Du, and Xuguang Lan. "A review of object detection based on deep learning." *Multimedia Tools and Applications* 79 (2020): 23729-23791.
- [21] Zou, Zhengxia, Keyan Chen, Zhenwei Shi, Yuhong Guo, and Jieping Ye. "Object detection in 20 years: A survey." *Proceedings of the IEEE* 111, no. 3 (2023): 257-276.
- [22] Wu, Xiongwei, Doyen Sahoo, and Steven CH Hoi. "Recent advances in deep learning for object detection." *Neurocomputing* 396 (2020): 39-64.
- [23] Oksuz, Kemal, Baris Can Cam, Sinan Kalkan, and Emre Akbas. "Imbalance problems in object detection: A review." *IEEE transactions on pattern analysis and machine intelligence* 43, no. 10 (2020): 3388-3415.
- [24] Küçüktabak, Emek Barış, Sangjoon J. Kim, Yue Wen, Kevin Lynch, and Jose L. Pons. "Human-machine-human interaction in motor control and rehabilitation: a review." *Journal of neuroengineering and rehabilitation* 18 (2021): 1-18.
- [25] Xu, Wei, Marvin J. Dainoff, Liezhong Ge, and Zaifeng Gao. "Transitioning to human interaction with AI systems: New challenges and opportunities for HCI professionals to enable human-centered AI." *International Journal of Human-Computer Interaction* 39, no. 3 (2023): 494-518.
- [26] Haroon, Umair, Amin Ullah, Tanveer Hussain, Waseem Ullah, Muhammad Sajjad, Khan Muhammad, Mi Young Lee, and Sung Wook Baik. "A multi-stream sequence learning framework for human interaction recognition." *IEEE Transactions on Human-Machine Systems* 52, no. 3 (2022): 435-444.
- [27] Gong, An, Chen Chen, and Mengtang Peng. "Human interaction recognition based on deep learning and HMM." *IEEE Access* 7 (2019): 161123-161130.
- [28] Puchała, Sebastian, Włodzimierz Kasprzak, and Paweł Piwowarski. "Human Interaction Classification in Sliding Video Windows Using Skeleton Data Tracking and Feature Extraction." *Sensors* 23, no. 14 (2023): 6279.
- [29] Wang, Zhenhua, Kaining Ying, Jiajun Meng, and Jifeng Ning. "Human-to-Human Interaction Detection." In *International Conference on Neural Information Processing*, pp. 120-132. Singapore: Springer Nature Singapore, 2023.

# Fall Event Direction Classification Based on Video Data from Multiple Cameras Using Deep Learning

Kuntha. Pin<sup>1,\*</sup>, Chomyong. Kim<sup>2</sup>, and Yunyoung. Nam<sup>3</sup>

<sup>1</sup>*Emotional and Intelligence Child Care System Convergence Research Center, Soonchunhyang University, Asan, South Korea*

<sup>2</sup>*ICT Convergence Research Center, Soonchunhyang University, Asan, South Korea*

<sup>3</sup>*Department of Computer Science and Engineering, Soonchunhyang University, Asan, South Korea*

\*Contact: pin.kuntha145@gmail.com

**Abstract**— Falls among the elderly and individuals with health conditions are a significant concern in healthcare, often leading to serious injuries and a reduced quality of life. Currently, most classification methods are binary, and there are many methods for fall direction classification using single or dual cameras. However, there are some limitations related to lack of specific directions of falls, the procedure for conducting datasets, and number of camera to capture multiple angles of views, which affect the method's generalizability. In this study, we propose a method for classifying three directions of fall events and non-fall using videos from eight cameras. We collected a dataset from 24 participants, with each event around 10 seconds and captured from various views using eight cameras at a high-quality frame rate of 60 fps. The dataset comprises 7736 forward falls, 5816 backward falls, 3440 sideward falls, and 5680 non-falls. The dataset is pre-processed to extract frames from video data and resize them. Our proposed method extracts skeleton features from each frame of video, which are then stacked continuously to create sequences of skeleton features. Convolutional Neural Networks (CNNs) extract spatial features from the skeleton sequence features, which are then input to Bidirectional Long Short-Term Memory (BiLSTM) networks. The BiLSTM networks extract temporal features from the sequence data and classify into three fall directions and non-falls. Dense layers and a final dense layer followed by SoftMax activation calculate class probabilities based on the extracted features. Our proposed method classifies three different directions of falls and non-falls using a large and generalized dataset. The result of our proposed method for fall classification achieved an accuracy of 93.18%.

## I. INTRODUCTION

Falls are a significant public health concern worldwide, particularly among the elderly population, where they can lead to severe injuries, reduced mobility, and even mortality [1]. Worldwide, the fall adults older than 65 years old leading cause unintentional injuries, 37.3 million of falls need medical attention and 646,000 resulting in deaths annually [2]. Therefore, the detection of fall events is crucial for safety in both solitary indoor scenes and crowded outdoor environments.

In recent years, various methods utilizing different devices have been proposed for fall classification. Wearable devices, such as tilt sensors, accelerometers, and gyroscopes, are widely used in previous studies [3]. However, using those sensors isn't convenient for patients; they can be uncomfortable, complex,

and hinder normal activity. Therefore, non-contact sensors like cameras have become popular for fall classification [4-5].

In this study, presents a novel approach to fall classification, focusing on the classification of backward-fall forward-fall, and sideward-falls, alongside non-fall activities. Traditional fall detection systems often adopt binary classification, categorizing activities simply as falls or non-falls. However, this oversimplified approach fails to capture the nuances of different fall directions, which are essential for tailored intervention strategies and medical decision-making. Our work addresses this gap by proposing a comprehensive fall classification system that discriminates between multiple fall directions and non-fall activities. In addition, the proposed method trained and evaluated on the diverse dataset which collect from multiple cameras with indoor and outdoor environments. The constructed model with lightweight needs small resource for running.

## II. RELATED WORKS

Many works study on methodology of fall classification or fall detection, with most of them are kind of binary classification, distinguishing between falls and non-falls. Additionally, some research focuses on multiclass detection, which encompasses both daily activities and fall classifications.

Sultana et al. [6] employed deep learning techniques for fall and non-fall classification. They utilized two indoor public datasets to implement and evaluate their methodology. These datasets encompassed various cameras operating at different frequencies (frames per second). Additionally, the durations of the video files varied, with only 10 frames extracted from each. Preprocessing steps, including frame extraction, and resizing, were applied before input into their two-dimensional convolutional neural network (2DCNN) combined with a gated recurrent unit (GRU). This model was then used to classify videos into two classes: fall and non-fall. This binary classification using indoor public datasets has limitations, particularly in terms of processing time because images are directly inputted into deep learning models without extracting other features instead, and suitability for indoor applications.

Yadav et al. [7] developed a method for detecting seven classes, including bending fall, lying, running, sitting, standing, and walking. The dataset was collected using a single camera,

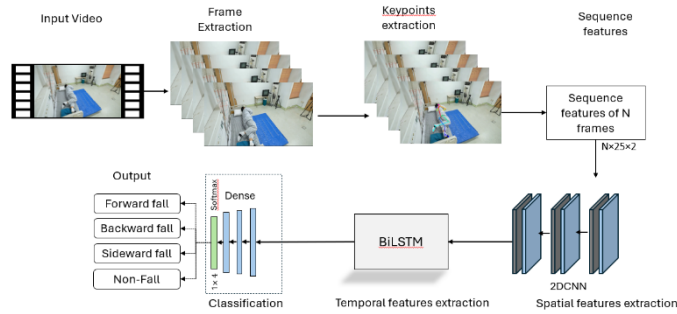


capturing 8 minutes of each activity. Preprocessing involved extracting 45 frames with a 30-frame overlap. Skeleton keypoint features were then extracted from the frames, followed by the utilization of a CNN and Long Short-Term Memory (LSTM) model to extract features, and classify normal activities and bending falls. The result achieved an accuracy of 89.64%. However, the classification does not focus on a specific side of the fall, addressing only one aspect of falls. In the real world, the possibility of falls extends to multiple directions. Identifying the specific direction of a fall aids doctors in focusing and assessing the need for surgery.

In this study, we propose a method to classify three directions of falls and non-falls using a combination of two deep learning models. We collected a dataset using 8 cameras to record videos simultaneously, encompassing indoor and outdoor environments with various scenarios to create a generalized dataset. To reduce computational time, we extract significant features for training the proposed model. The model extracts both spatial and temporal features from sequence data (videos) and classifies videos into different classes: forward fall, backward fall, sideward fall, and non-fall.

### III. PROPOSED METHOD

This section outlines the methodology of the proposed fall direction classification approach, which comprises three main steps. Firstly, a large video dataset is captured using multiple cameras and designed to depict various scenarios was assembled. Secondly, a pose estimation method was employed to extract key points of the human body. The key points from each video frame were then stacked to create sequential data containing 120 frames. Thirdly, a deep learning model was developed to extract spatial and temporal features from the sequential skeleton-features and classify it into one of four categories. The overview of the proposed method is presented in Fig.1.

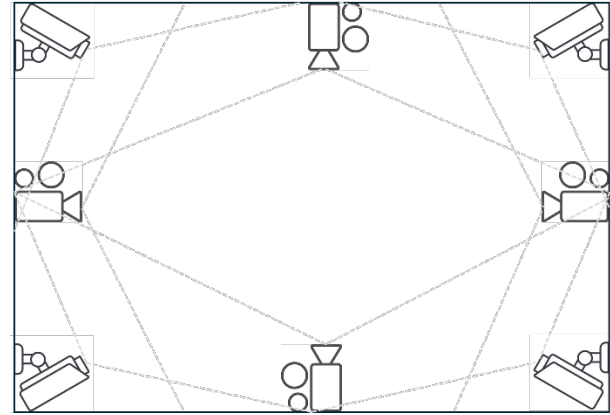


**Fig. 1** The schematic diagram of the proposed method for fall direction classification.

#### 1. Proposed Dataset

We propose a dataset comprising 24 participants (11 healthy adult men and 13 healthy adult women), with recordings captured by 8 cameras simultaneously capturing various activities, including falls in different directions as well as non-fall scenarios. Each camera was setup at different angles, distances and heights around the participant to capture each

event at different views. The installation camera positions are not fix at specific angles, but keep covering around participant, the camera randomly allocate as represented in Fig.2. The videos are of high resolution, with dimensions of 3840 pixels by 2160 pixels, and were recorded at a frame rate of 60 frames per second. Each participant was instructed to perform activities for approximately 10 seconds, encompassing pre-fall, fall, and post-fall phases for representing each direction of falls, while non-fall videos contain a range of activities such as daily routines, stumbling, sitting, wheelchair use, body movements, and more. These non-fall activities have the actions frequently undertaken by patients in various environments, including homes, hospitals, nursing therapy, streets, and other outdoor environments. The dataset size presents in Table. 1.



**Fig. 3** Eight Camera positions installation.

**Table 1.** Dataset Collection

Class	Video File	Duration	FPS	Frame Resolution	Number Camera
Forward fall	7,736	10 seconds for each video	60	3840×2160	8
Backward fall	5,816				
Sideward fall	3,440				
Non-Fall	5,680				
Total	22,672				

#### 2. Preprocessing and Keypoints Feature Extraction

Fall likes other actions, constitutes sequential data over a period of time. Instead of utilizing all frames from the videos, we extract 120 frames from a total of 600 frames for each video. This means the sampling interval is 5 frames, indicating that we select every 5th frame from the video sequence to form a sequence of 120 frames. By adopting this approach, we aim to reduce the computational load and processing time while still capturing the essential motion dynamics within the video data. This sampling strategy ensures that we maintain a

representative subset of frames for analysis, facilitating efficient processing and meaningful insights into the underlying motion patterns. The equation to calculate the sampling interval is given by Eq. (1). Each frame is downsized to 500 pixels in width and 300 pixels in height to reduce computational load in the pose estimation process.

$$\text{Sampling Interval} = \frac{\text{Total Number of Frames}}{\text{Number of Frames to Sample}} \quad (1)$$

Post-estimation is employed to extract the 2D joint locations of the skeleton. Skeleton key points detection is performed using OpenPose[8]. The pre-train of OpenPose extracts 25 key points from the human body, with each point represented by 2D X and Y coordinates. Fig.3 illustrates the 25 key points extracted by the OpenPose from the human body.

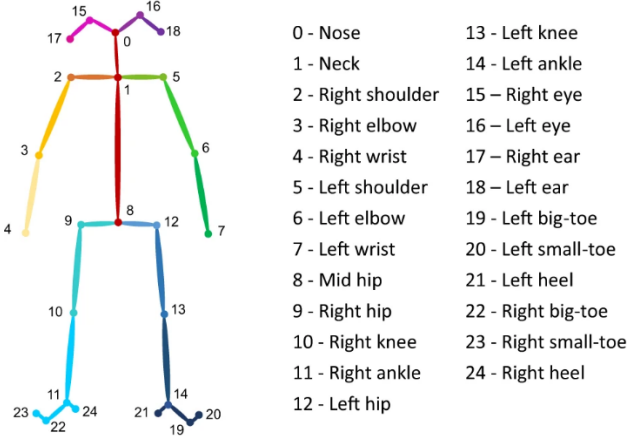


Fig. 4 OpenPose detects 25 keypoints of the human body [9].

The output from the OpenPose for single frame and a detected object is an array with a shape of  $25 \times 3$ ; 25 represents the number of key points, and 3 represents the X-coordinate, Y-coordinate, and z-confidence score. We do not utilize the z-confidence score, it is removed from the array. Therefore, the frame becomes an array with a shape of  $25 \times 2$ ; where X values range from 0 to 500, corresponding to the width of the frame, and Y values range from 0 to 300, corresponding to the height of our frame.

Extracted key points from frames of video are stacked in order to construct sequential data. Each sequential data has a shape of  $120 \times 25 \times 2$ , where 120 denotes the number of frames, 25 signifies the number of key points detected in each frame, and 2 represents the X and Y coordinates of each key point.

### 3. Proposed Model

The proposed model architecture is combination of two-dimensional convolutional neural network (2DCNN) model and bidirectional long short-term memory (BiLSTM) models. The 2DCNN model, responsible for extracting spatial features from the input data, is added as the initial layer. The subsequent BiLSTM model capture temporal dependencies within the sequential data. This combination enables the model to effectively learn and interpret both spatial and temporal

information, which is crucial for accurate fall direction classification.

The Convolutional Neural Network (CNN) model extracts spatial features from the input data. It consists of multiple Convolutional and MaxPooling layers followed by a Reshape layer to transform the spatial data into a two-dimensional array. The Convolutional layers apply filters to the input data to detect spatial patterns, while the MaxPooling layers downsample the feature maps to reduce computational complexity and extract dominant features. The Reshape layer converts the multidimensional feature maps into a two-dimensional array, preparing them for input into the subsequent BiLSTM model. The BiLSTM model processes the sequential data outputted by the 2DCNN model to capture temporal dependencies and perform classification. It consists of multiple BiLSTM layers, each with a specified number of units (30, 60, 60, and 60). The input BiLSTM layer receives the sequential data and passes it through BiLSTM units, which maintain memory over time and capture long-range dependencies. Subsequent BiLSTM layers further refine the temporal representations learned by the previous layers. Batch Normalization layers stabilize and accelerate training by normalizing the activations, while Dropout layers prevent overfitting by randomly deactivating a fraction of the BiLSTM units during training. Finally, Dense layers with ReLU activation functions extract high-level features from the BiLSTM output, followed by a Dropout layer to mitigate overfitting. The output layer with softmax activation computes class probabilities for classification. The proposed model architecture presents in Fig.4.

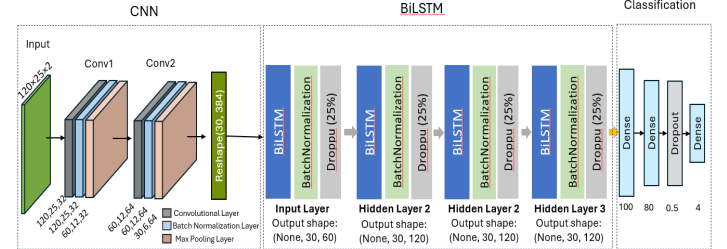


Fig. 5. The network architecture of proposed 2DCNN-BiLSTM model fall classification.

## IV. EXPERIMENTS AND RESULTS

The dataset is divided into training (70%), validation (15%), and testing (15%) sets. The input to the 2DCNN model is an array of shape  $120 \times 25 \times 2$ , while the output is shape of  $30 \times 384$ . The output of the 2DCNN serves as the input to the BiLSTM model. The BiLSTM model is tasked with learning to extract temporal features from sequences of spatial features, followed by Dense layers with vectors of 100, 80, and 4. Dropout layers are applied between each Dense layer to mitigate overfitting. For further details on the training parameters, please refer to Table 2.

Table 2. Training parameters

Training Parameters	Values
---------------------	--------



Epoch	100
Batch size	32
Optimization/ Learning rate	Adam/ 0.0001
Loss	sparse_categorical_crossentropy

The proposed model was trained for 100 epochs. For 4 classes, the total number of parameters is 547,832, with 1,512 being non-trainable and 546,320 being trainable. We experimented with two different numbers of generated frame  $F_g = 120$ , 180 to access the optimal number of generated frames that provide better accuracy.

For evaluation, the test set consists of sequential data total of 4,535 samples (15%) was utilized to assess the performance of the proposed model. Accuracy, the confusion matrix, and Receiver Operating Characteristic (ROC) curves were calculated to evaluate the classification performance. Our proposed model achieved an accuracy of 93.18% across four classes: forward-fall, backward-fall, side-fall, and non-fall. Result of applying the OpenPose for extracting key points from each frame of different class in Fig. 6. Classification performance of 120-generated frames and 180-generated frames are presented in Table 3 and 4. Also The confusion matrixes and ROC curves are shown in Fig. 7-10.



**Fig. 6.** Results of extracting key points from frame of four classes.

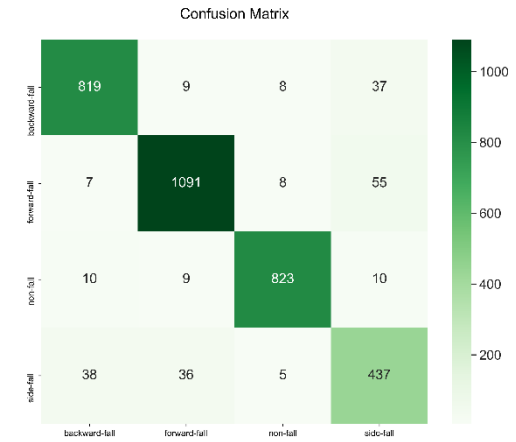
**Table 3.** Classification performance of collected dataset for 120-generated frames.

Class	Precision	Recall	F1-Score	Overall Accuracy
Backward-fall	93.71%	93.81%	93.76%	93.18%
Forward-fall	95.28%	93.97%	94.62%	
Side-fall	81.08%	84.69%	82.84%	

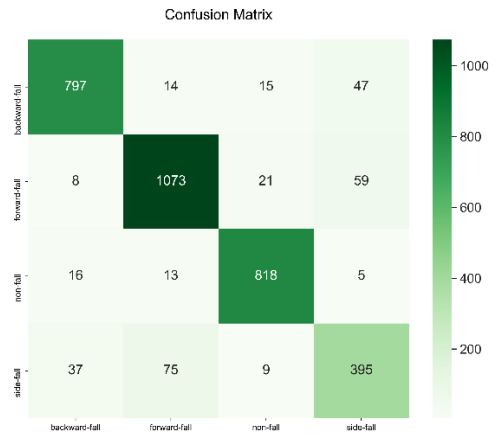
Non-fall 97.51% 96.60% 97.05%

**Table 4.** Classification performance of collected dataset for 180 generated frames.

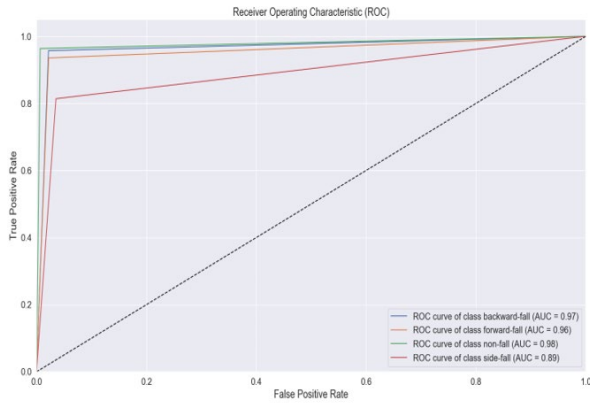
Class	Precision	Recall	F1-Score	Overall Accuracy
Backward-fall	92.89%	91.29%	92.09%	90.62%
Forward-fall	91.32%	92.42%	91.87%	
Side-fall	78.06%	76.55%	77.30%	
Non-fall	94.79%	96.01%	95.39%	



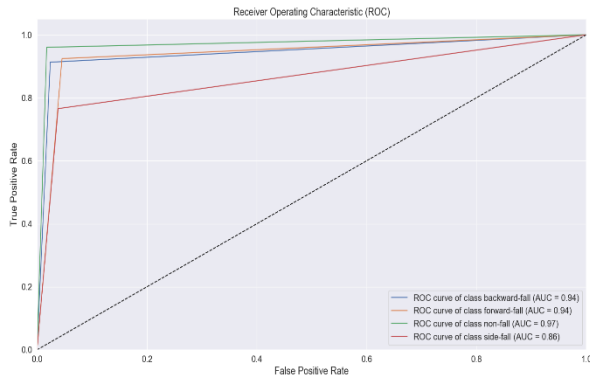
**Fig. 7.** Confusion matrix of experiment with 120-generated frames.



**Fig. 8.** Confusion matrix of experiment with 180-generated frames.



**Fig. 9** ROC of experimentation with 120-generated frames



**Fig. 10** ROC of experimentation with 180-generated frames

Examining the classification outcomes of the experiments, explicitly focusing on the 120-frame generation and 180-frame generation, provides useful insights into the efficacy of our suggested method for classifying falls. The experiment with 120-frame generation yielded excellent precision, recall, and F1-score results for all categories, demonstrating strong performance in accurately detecting backward-fall, forward-fall, side-fall, and non-fall events. The precision, recall, and F1-score of backward fall were 93.71%, 93.81%, and 93.76%, respectively, and were 95.28%, 93.97%, and 94.62% for forward fall, respectively. The performance metrics for both side-fall and non-fall events were remarkable, with precision, recall, and F1-score ranging from 81.08% to 97.51%. We observed a little lower performance for the classification performance of the experiment with the 180-frame generation compared to the 120-frame generation. While the precision, recall, and F1-score for most classes remained quite good, there was a noticeable decrease in these metrics, specifically for the side-fall class. The side-fall class has a precision of 78.06%, a recall of 76.55%, and an F1 score of 77.30%. These values indicated that the accuracy in detecting this particular class is somewhat lower compared to other forward-fall, backward-fall, and non-fall. Even though the total accuracy for the 180-frame generation still remained quite high accuracy of 90.62%, the efficiency of our fall directions detection approach was highlighted even with a higher number of frames.

The decrease in performance, from an accuracy of 93.18% with the experiment of 120-frame generation frames to 90.62%

with 180-frame generation, indicates a potential limit in data density. More frames can offer more data; they might add noise, burden the model, and boost computational demands without significant performance improvements. This shows the significance of carefully adjusting frame selection to balance the data information level and the model's efficiency. Exploring the optimal number of generated frames and improved frame selection techniques could boost classification accuracy.

Our fall direction classification method is designed for single individuals. The use of 8 cameras capturing various angles enhances the adaptability of our approach when applied in real-world settings. A substantial dataset is crucial for the effectiveness of our deep learning model. However, our study's performance requires further improvement. Moving forward, our goal is to enhance classification accuracy and validate our model with additional published fall direction datasets to evaluate its versatility across different datasets. Additionally, we intend to conduct real-world tests to assess its speed and practicality.

## V. CONCLUSIONS

In conclusion, our fall event direction classification, leveraging video data from multiple cameras and employing a developed 2DCNN and BiLSTM model, yielded an impressive classification accuracy of 93.18%. By extracting human key points from each video frame as features for training the deep learning model, we optimized computational efficiency and produced a more lightweight model compared to directly inputting frames. This innovative approach not only enhances performance but also streamlines processing, demonstrating the effectiveness of utilizing key points instead of raw image data for fall event classification. In the future, we aim to improve classification performance and optimize execution time.

## ACKNOWLEDGMENT

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2023-00218176).




## REFERENCES

- [1] X. Qingmei, X. Ou, and J. Li. "The risk of falls among the aging population: A systematic review and meta-analysis." *Frontiers in public health* 10, 2022.
- [2] S. Francy, and J. Shu. "An eight-camera fall detection system using human fall pattern recognition via machine learning by a low-cost android box." *Scientific reports* 11, no. 1, 2021.
- [3] Feng, Qi, Chenqiang Gao, Lan Wang, Yue Zhao, Tiecheng Song, and Qiang Li. "Spatio-temporal fall event detection in complex scenes using attention guided LSTM." *Pattern Recognition Letters* 130 (2020): 242-249.
- [4] Noor, Nadhira, and In Kyu Park. "A Lightweight Skeleton-Based 3D-CNN for Real-Time Fall Detection and Action Recognition." In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2179-2188. 2023.
- [5] Vishnu, Chalavadi, Rajeshreddy Datla, Debadiya Roy, Sobhan Babu, and C. Krishna Mohan. "Human fall detection in surveillance videos using fall motion vector modeling." *IEEE Sensors Journal* 21, no. 15 (2021): 17162-17170.
- [6] Sultana, A. et al. (2021) 'Classification of indoor human fall events using Deep Learning', *Entropy*, 23(3), p. 328. doi:10.3390/e23030328.
- [7] Yadav, Santosh Kumar, Achleshwar Luthra, Kamlesh Tiwari, Hari Mohan Pandey, and Shaik Ali Akbar. "ARFDNet: An efficient activity

- recognition & fall detection system using latent feature pooling." Knowledge-Based Systems 239 (2022): 107948.
- [8] Kim, Woojoo, Jaeho Sung, Daniel Saakes, Chunxi Huang, and Shuping Xiong. "Ergonomic postural assessment using a new open-source human pose estimation technology (OpenPose)." International Journal of Industrial Ergonomics 84 (2021): 103164.
- [9] Zhang, Mingming, Yanan Zhou, Xinye Xu, Ziwei Ren, Yihan Zhang, Shenglan Liu, and Wenbo Luo. "Multi-view emotional expressions dataset using 2D pose estimation." Scientific Data 10, no. 1 (2023): 649.





Gimje	 <p><b>&lt;Research-verification-certification system&gt;</b></p> <ul style="list-style-type: none"> <li>Gimje has established a &lt;research-verification-certification system&gt;.</li> <li>Establishment of &lt;seed development-commercialization model&gt; for crops.</li> </ul>
Goheung	 <p><b>&lt;Subtropical crop cultivation&gt;</b></p> <ul style="list-style-type: none"> <li>Subtropical crops are grown in Goheung.</li> <li>Construction of a large-scale complex using reclaimed land.</li> <li>Supports rental smart farm operation and settlement.</li> </ul>
Miryang	 <p><b>&lt;National Industrial Complex-Local University&gt;</b></p> <ul style="list-style-type: none"> <li>Miryang is focusing on exporting strawberries and mini paprika.</li> <li>Linked with &lt;National Industrial Complex-Local University&gt;.</li> <li>We present an energy model utilizing waste heat from a sewage treatment plant.</li> </ul>




([www.nabis.go.kr/termsDetailView.do?menucd=189&gbnCode=S51&eventNo=340](http://www.nabis.go.kr/termsDetailView.do?menucd=189&gbnCode=S51&eventNo=340))

## 2.2 DOMESTIC SMART FARM



Table 2 classifies smart farm cases.

There are four types: ①building type ②energy type ③material type and ④application type.

**Table 2.** Domestic cases type of Smart Farm

local government	Characteristics
	<b>TYPE 1 - Building type</b>
‘Daejeon on Farm’	 <p><b>&lt;Daeheung-dong building, Jung-gu, Daejeon&gt;</b></p> <ul style="list-style-type: none"> <li>A smart farm was created in an empty space on the 2nd basement and 8th floor of a building in Daeheung-dong, Jung-gu, Daejeon.</li> <li>A ‘Smart Farm for Technology Research’ has been installed here.</li> <li>Strawberries and medicinal crops were planted on the second basement level, and automated facilities were installed.</li> <li>Education, promotion, and community spaces were installed on the 8th floor of the building.</li> </ul>
‘Metro Farm’	 <p><b>&lt;4 Seoul subway station&gt;</b></p> <ul style="list-style-type: none"> <li>(Seoul Metropolitan City + Seoul Transportation Corporation + Farm8) installed smart farms in four Seoul subway stations.</li> <li>This is a place where vegetables are produced, grown, and sold.</li> <li>In addition, smart farms were installed in Gwangju and Busan subway stations.</li> </ul>
‘Zaram Smart Farm’	 <p><b>&lt;Gwangju Geumnam-ro 4-ga Station&gt;</b></p> <ul style="list-style-type: none"> <li>(Gwangju Urban Railroad Corporation + Barun Farm Co., Ltd.) cultivated strawberries, lettuce, and ginseng within Geumnam-ro 4-ga Station (1,089 m<sup>2</sup>) in 2022.</li> </ul>



Smart farm in the supermarket	 <p>&lt;Smart Farm – Launched in Hanaro Mart&gt; (<a href="http://www.hankyung.com/life/article/202107086022e">www.hankyung.com/life/article/202107086022e</a>)</p> <ul style="list-style-type: none"> <li>Smart Farm Center Co., Ltd. installs and sells smart farms within supermarkets.</li> </ul>	<p><b>TYPE 3 - Material type</b></p>  <p>‘Smart Farm Cube’</p> <p>&lt; ‘Smart Farm Cube’- Agricultural container&gt;</p> <ul style="list-style-type: none"> <li>This is a place where crops are grown by creating a ‘Smart Farm Cube’ using containers.</li> </ul>
	<p><b>TYPE 2 - Energy type</b></p>  <p>&lt;Sehan Energy Co., Ltd. and Yeongcheon City Smart Farm&gt;</p> <ul style="list-style-type: none"> <li>A (solar + geothermal) hybrid system was installed here. (<a href="http://www.kharn.kr/mobile/article.html?no=19217">www.kharn.kr/mobile/article.html?no=19217</a>)</li> <li>Sehan Energy Co., Ltd. and Yeongcheon City are cultivating subtropical crops using a (solar heat + geothermal) hybrid system.</li> </ul>	<p><b>TYPE 4 - Application type</b></p>  <p>Smart Farm Application</p> <p>&lt;Crop management platform ‘Ara Greenhouse’&gt; (<a href="http://www.news1.kr/articles/?5348768">www.news1.kr/articles/?5348768</a>)</p> <ul style="list-style-type: none"> <li>‘Ara Greenhouse’ is a system that utilizes smart agricultural machinery and cultivation technology.</li> <li>Agricultural productivity increases by 37.6% and income increases by 46.3%. Labor is expected to save 11.1%.</li> </ul>
Use of factory waste heat	 <p>&lt;Smart farm that recycles factory waste heat&gt; (<a href="http://www.economychosun.com/site/data/html_dir/2023/08/25/2023082500020.html">www.economychosun.com/site/data/html_dir/2023/08/25/2023082500020.html</a>)</p> <ul style="list-style-type: none"> <li>Here, vegetables and fruits (4,400 m<sup>2</sup>) are grown using the factory waste heat.</li> </ul>	<p><b>2.3 JAPAN’S SMART FARM</b></p> <p>In Japan, smart farms are called smart agriculture. In 2017, Japan classified smart agriculture into four types.</p> <ol style="list-style-type: none"> <li>①Robot Type: Robotized Low Energy Agriculture</li> <li>②AI Type: Agriculture that is easy for anyone to use</li> <li>③Big Data Type: Strategic Production Using Data</li> <li>④ IoT type: Linkage and efficiency of production, distribution, and sales. (<a href="http://www.maff.go.jp/j/kanbo/smart">www.maff.go.jp/j/kanbo/smart</a>)</li> </ol>



**Table 3.** Japan's cases type of Smart Farm

local government	Characteristics	
TYPE 1 - Building type		
<p>&lt;Tokyo&gt; Pasona O2 Underground Farm</p>	<div></div> <ul style="list-style-type: none"><li>· In 2010, more than 200 kinds of plants, vegetables, and fruits were grown with artificial lighting (halogen lamps, high-pressure sodium, LEDs, etc.) on an area of 4,000 m<sup>2</sup> under a building in Tokyo.</li><li>· Citizens are using open rest area cafes and restaurants.</li></ul>	<p>&lt;Tokyo Chiyoda Ward 区&gt; Otemachi (町)- Small ranch</p> <div></div> <ul style="list-style-type: none"><li>• On the 13th floor of a building in Otemachi, Tokyo, "Otemachi small ranch" is operated.</li><li>• This place provides a resting place for urban residents by raising cattle, goats, mini pigs, and alpacas.</li></ul>
TYPE 2 - Energy type		
<p>&lt;Yokosuka City, Kanagawa&gt; Toshiba's Abandoned factory</p>	<div></div> <ul style="list-style-type: none"><li>• Spinach and lettuce are grown in a factory in Toshiba.</li><li>• The factory has not been in use since the mid-1990s and has been growing vegetables in a clean room with an area of 1,969 m<sup>2</sup> in 2014.</li><li>• It also has a resting area for locals and an air purification effect.</li></ul>	<p>&lt;Kurokawa 郡 in Miyagi Prefecture&gt; A greenhouse that utilizes waste heat from a factory</p> <div></div> <ul style="list-style-type: none"><li>• Vegetables are grown using about 90 degrees of wastewater from the Toyota Motor Company plant.</li></ul>
<p>&lt;Fukushima Aizumawa Kamatsu City&gt; Fujitsu Semiconduct or Factory</p>	<div></div> <ul style="list-style-type: none"><li>• It grows vegetables in the clean room of the Fujitsu semiconductor plant.</li><li>• Low potassium lettuce is lowered to 1/5 potassium and is sold to hospitals for patient food.</li></ul>	<p>&lt;Yokohama city, Rikuzentakata City&gt; Domed plant factory</p> <div></div> <p>Domed plant factory (<a href="http://www.yokohama-sozokaiwai.jp/things/6652.html">www.yokohama-sozokaiwai.jp/things/6652.html</a>) a. a domed plant factory • In January 2014, a 20-meter-diameter dome-shaped plant factory was established in Yokohama City to grow vegetables.</p> <ul style="list-style-type: none"><li>• A circular water tank was installed inside.</li><li>• Vegetables can be harvested about one month after sowing using IT technology.</li><li>• Rikuzentakata City in Iwate Prefecture grows lettuce using renewable energy in a dome-type plant factory.</li></ul>
TYPE 3 - Material type		

Japan Dome House



- An agricultural space that replaces greenhouses.
- The dome house used a new material (polystyrene) with high functionality and durability.
- The agricultural dome has an area of 10 m<sup>2</sup> and can be used for various purposes such as housing, toilets, warehouses, and work spaces.
- In 2013, the Japanese prime minister visited Bahrain and Qatar to introduce and support the dome house.
- In countries with low food self-sufficiency rates, it can be a stable agricultural facility.

#### TYPE 4 - Application type

Application for Smart Farm



- Root Co., Ltd. has launched a smart experience farm system called 'Root Farm'.
- It actually connects farms and apps to transmit videos, photos, and crop data through smartphones and PCs.
- This app can grow crops remotely, just like games.
- And harvested crops can also be delivered to their homes.

### III. Conclusion

We looked at smart farms in Korea and Japan. The above is summarized as follows.

① Recently, the government plans to implement the 'Smart Agriculture Act' (2024.7). Smart farms seek to enhance agricultural technology by utilizing ICT technology and infrastructure.

② The 'Smart Farm Innovation Valley', promoted by the government since 2018, has been installed and operated in four places nationwide. Each of the four valleys has different major projects. It is carrying out youth agricultural support (Sangju), technological innovation (Gimje), subtropical crops (Goheung), and variety diversification (milyang).

③ In the future, in connection with universities, projects such as <Student + Young people> Fostering - Settling in the Region can be expected.

④ In both Korea and Japan, smart farms were installed inside subway stations, buildings, and factories.

The energy provided to smart farms and the materials that make smart farms are constantly being developed.

⑤ Japan's smart farms have made more efforts to develop energy and new materials.

Currently, smart cities are developing in cities and smart farms in rural areas. Smart farms are also possible in cities.

Inside city buildings, smart farms can become the new agriculture of the future. Architecture experts should work on building technologies for smart farms.

### REFERENCES

- 1 한국농촌경제연구원, 주간농업농촌 동향(도시형 식물공장), (www.krei.re.kr/selectBbsNttView.do?bbsNo=76&key=271&nttNo=45796)
- 2 디지털농업과 스마트농업의 차이?, 농수축산신문, 2022. 6. 7. (www.afnews.co.kr/news/articleView.html?idxno=227887)
- 3 스마트농업 육성 및 지원에 관한 법률[시행 2024. 7. 26] (www.law.go.kr/LSW/lsInfoP.do?lsiSeq=252869&viewCls=lsRvsDocInfoR#)
- 4 균형발전종합정보시스템 (www.nabis.go.kr/termsDetailView.do?menucd=189&gbnCode=S51&eventNo=340)
- 5 반도체 찍던 생산라인에서 채소가 자랍니다, 경향신문, 2014. 10. 6. (www.m.khan.co.kr/world/japan/article/201410062153515#c2b)
- 6 전국 첫 공실 건물 활용 스마트 '대전팜' 개장, 대전일보, 2024. 2. 6. (www.daejonilbo.com/news/articleView.html?idxno=2112519)
- 7 카타르에 한국형 스마트팜 혁신밸리 모델 심는다, 한국영농신문, 2024. 2. 27. (www.youngnong.co.kr/news/articleView.html?idxno=45806)
- 8 동양파이오뉴스, 2024. 11. 7. (www.dybionews.com/news/articleView.html?idxno=14315)
- 9 식품외식경영(www.foodnews.news/news/article.html?no=233433)
- 10 스마트팜코리아(www.smartfarmkorea.net)
- 11 ㈜바름팜 홈페이지(www.barunfarm.net)
- 12 애그테크(www.agtecher.com/ko/what-is-agtech-2)
- 13 이코노미조선(www.economychosun.com)
- 14 재팬돔하우스(www.dome-house.jp)
- 15 창조도시요코하마(www.yokohama-sozokaiwai.jp/things/6652.html)
- 16 골든플래닛(www.goldenplanet.co.kr)
- 17 일본 농림수산성(www.maff.go.jp)

# Fall Classification Using IMU Sensor Based on Deep Learning

Sokea TENG<sup>1</sup>, Jung-yeon KIM<sup>2</sup>, Yunyoung NAM<sup>3</sup>

<sup>1</sup>Department of ICT Convergence, Soonchunhyang University, Asan, 31538, Republic of Korea

<sup>2</sup>ICT Convergence Research Centre, Soonchunhyang University, Asan, 31538, Republic of Korea

<sup>3</sup>Department of Computer Science and Engineering, Soonchunhyang University, Asan, 31538, Republic of Korea

\*Contact: [ynam@sch.ac.kr](mailto:ynam@sch.ac.kr)

**Abstract**— Fall-Sense is a cutting-edge project that employs learning techniques along with wearable sensors to automatically detect and classify falling types. Falls can happen to anyone and are a crucial area of research for reliable classification and understanding of fall directions. In this study, we introduced a robust system using 12 wearable sensors on the body such as head, left and right shoulders, left and right upper arms, left and right forearms, pelvis, left and right upper legs, and left and right lower legs. By using built-in accelerometers, gyroscopes, and magnetometers, it accurately captures fall data, distinguishing between 'Non-fall' and 'Fall' events. Furthermore, it classified the direction of falls as 'Forward-fall,' 'Backward-fall,' or 'Lateral-fall.' In addition, we proposed a state-of-the-art supervised deep learning method that demonstrated the advantages of a deep architecture based on the combination of Convolutional and LSTM recurrent layers to perform fall classification from 12 wearable IMU sensors. The performance reached the highest accuracy of binary class and multi-classes 99.65 % and 97.89%, respectively. This innovative approach holds the promise of improving safety in various environments. Falls and fall direction classifications enable quick responses, reducing the risk of injuries, particularly for the elderly and patients who need constant monitoring, and assist other researchers in classifying fall types for the study of fall risk assessment with long-term data.

**Keyword:** Falls, Fall direction, classification, deep learning, feature extraction, Conv-LSTM models, IMU sensors, activity daily life.

## I. INTRODUCTION

A fall event is defined as sudden and unintentional collapses from an upright position when a person's legs can no longer support the body. These incidents can result in significant physical and emotional harm, disability, and a loss of independence. Additionally, falls may lead to post-fall syndrome, characterized by dependence, loss of autonomy, depression, and further limitations in daily activities, sometimes even resulting in premature death. It's worth noting that falls are not exclusive to the elderly or unhealthy individuals, but those can happen unexpectedly to anyone, anywhere.

Recent reports from the World Health Organization (WHO) highlighted falls as the second leading cause of unintentional

injury deaths worldwide. Each year, approximately 684,000 people globally lose their lives due to the falls, with over 80% occurring in low and middle-income countries. Fatal falls are most common among adults over 60 years of age. Furthermore, there are an estimated 37.3 million falls each year that require medical attention [1].

While fall events themselves are generally not life-threatening, they can result in serious health risks such as concussions and blood clots, sometimes leading to unfortunate fatalities, especially in cluttered environments. The lack of a timely response from emergency services, especially for those living independently, significantly increases these risks. Traditional surveillance systems, reliant on the constant presence of nurses and support staff, have been developed to address this challenge, but creating entirely fall-proof environments remains difficult.

Implementation of fall detection technologies and rescue services plays a crucial role in ensuring the safety of the elderly population and patients. These intelligent detection and prevention systems are essential to address the growing concern surrounding the well-being of individuals in such situations [2].

In response to these alarming statistics, wearable sensor devices such as accelerometers, gyroscopes, and pressure sensors have emerged. These devices can capture gait-related data and extract features that are used to predict fall risks and prevent falls [3]. Wearable sensors automatically record and analyse falling events.

Human Activity Recognition seeks to classify muscle activities and capture physiological data in a timely manner through pervasive computing. This not only contributes to medical diagnosis but also advances research in human activity [4]. As the social issue of aging continues to grow, interest in the activities of daily living (ADLs) among the elderly and related healthcare research is rapidly increasing.

The presented statistics underscore the significant impact of falls on older adults, underscoring the imperative for preventive measures and strategies to mitigate the incidence of



falls within this demographic. The literature offers various systems designed to identify falls, commonly utilizing two types of sensing devices: wearable [5] - [9] and non-wearable [10], to monitor or assess the user's motion. Examples of wearable technology include devices such as accelerometers, gyroscopes, and magnetometers [11]. Despite their affordability, wearable technologies exhibit several drawbacks, notably the necessity for constant user wear, which can be uncomfortable.

In contrast to the conventional manual feature extraction approach, deep learning minimizes the burden of feature design by employing end-to-end neural networks to autonomously learn and capture intricate high-level and meaningful features. Furthermore, the deep neural network structure proves to be well-suited for unsupervised learning and incremental learning, exhibiting superior scalability compared to traditional methods [12]. Khan et al. devised a hybrid deep learning (DL) model that leverages the strengths of two distinct DL architectures, namely Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM), to acquire spatial and temporal feature representations from input data. Additionally, the researchers assessed the performance of the developed CNN-LSTM model using a dataset consisting of 12 classes, gathered through the Kinect V2 sensor [13].

While Deep Learning (DL) has demonstrated remarkable success in fields such as computer vision, natural language processing, and speech recognition, it faces substantial challenges. The intricacies involved in configuring DL architectures and hyperparameters, coupled with the computational costs of model training, present significant obstacles. The reliance on large, labelled datasets for DL, while effective, can lead to performance issues when dealing with limited or imbalanced data. Additionally, the inherent 'black box' nature of DL models, the risk of overfitting, and the difficulty in result interpretation contribute to its limitations. The deployment of DL in real-time applications demands considerable resources, and ethical concerns regarding bias and accountability highlight the ongoing need for research and improvement in the field. In this study, we contribute to the field of biomedical knowledge discovery and engineering through the following scientific advancements:

- **Advanced Feature Extraction for Falls:** We carefully capture detailed information about various fall events, focusing on optimal feature extraction techniques. This reduces redundant data, ensuring low power consumption, fast computation, and high accuracy. Unlike traditional machine learning methods that struggle to differentiate similar activities through manual feature extraction, our approach effectively overcomes these challenges by using CNN for automated feature extraction.
- **Innovative Use of Conv-LSTM:** Our proposed solution leverages deep learning models, specifically the Conv-LSTM architecture, which integrates CNN for automated feature extraction and Long Short-Term Memory (LSTM) networks for precise classification of fall event directions.

- **Enhanced Accuracy Through Data Fusion:** By comparing raw data, norm (Euclidean magnitude), and combined raw with norm vectors, we demonstrate significant improvements in accuracy. For binary classification, the combined approach achieved an impressive accuracy of 99.65%, while for multi-class classification, it reached 97.89%. This fusion of data types enhances the overall performance of our wearable sensor technology.
- **Superior Performance in Model Comparison:** We validate the effectiveness of Conv-LSTM by comparing it against various traditional machine learning algorithms (RF, SVM, MLP, KNN) and other deep learning models (CNN, LSTM, GRU). Our results indicate that Conv-LSTM outperforms these methods in terms of accuracy, precision, recall, F1 score, and testing time for both binary and multi-class tasks.
- **Efficient Real-Time Fall Detection:** Despite the longer training time required by Conv-LSTM, its testing time remains under 1 second, making it suitable for real-time fall detection applications. This efficiency is crucial for practical deployment in healthcare settings, where timely intervention can prevent further injuries.

The paper is structured as follows: Section II provides an overview of the methodology employed in this study. Section III describes the experiments and results related to falls and fall direction and comparison with several ML, DL and proposed method. Section IV conclusion the results of falls and fall direction based on features vector from Norm, raw data and Conv-LSTM model.

## II. METHODOLOGY

Recently, there're many papers were studied and developed with various methods for fall risk assessment using wearable sensors. These techniques can be trained on labelled data to learn the patterns associated with falls and non-falls. In addition, another study was conducted on unsupervised learning algorithm to identify pattern of falls risk. Currently, there is no publicly available dataset containing physical fall activities. Therefore, we have contributed a new dataset containing records of four distinct physical activities performed by 20 participants. This section we will briefly explanation our proposed dataset, features extraction, and the internal our architecture were proposed.

### A. Material and Data collection

This section deals with the materials and methods applied in this study. The project aimed to implement a classification protocol utilizing trained data from an IMU (Inertial Measurement Unit) sensor, employing deep learning techniques to process the obtained dataset for classifying different types of fall events.

To record fall data, three sensors from the IMU, including the Accelerometer, Gyroscope, and Magnetic Field, were utilized. Each sensor consists of three axes (X, Y, Z) for measuring movement records. The data were collected from 20 participants, encompassing adults and youths of both genders,

in various settings including hospitals, homes, roads, and nursing homes.

Wearable sensors were attached to 12 specific locations on the body: the head, both shoulders, both upper arms, both forearms, the pelvis, both upper legs, and both lower legs. The obtained dataset was divided into four classes, with each class representing Non-Fall, Backward Fall, Forward Fall, and Lateral (side) falls. The details are presented in Table 1 and illustrated in Figure 1:

TABLE 1: THE NUMBERS OF FILE IN EACH CLASS, NUMBERS OF MAXIMUM ROW OF EACH FILE, AND TOTAL SAMPLE OF EACH CLASS.

Falls Type	Files each class	Maximum of row each file	Total samples of each class
None fall	720	600	432,000
Backward	736	600	441,600
Forward	977	600	586,200
Side	438	600	262,800
Total	2871	600	1,722,600

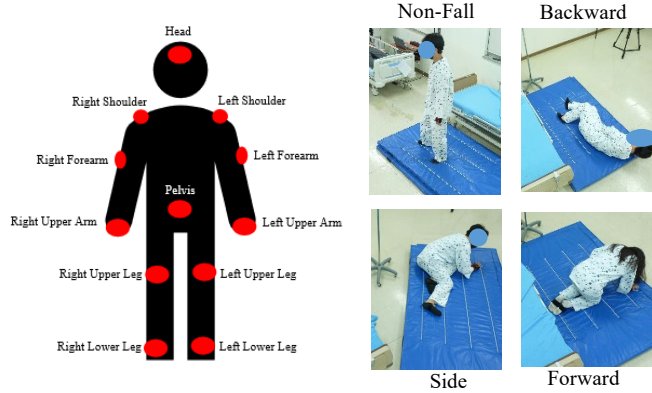


Figure 1: Describe the wearable place and show the type of falling.

### B. Data Pre-processing and labelling

Indeed, noise often occurs during data processing, which can interfere with the performance of the model. To address this issue, an additional preprocessing method involving noise reduction using a 1D Gaussian filter was proposed in this study. Additionally, data labelling for classification purposes was performed.

Considering that each file in the dataset contains a maximum of 600 rows (or 600 samples per file) with a sampling rate of 60 Hz, this effectively creates a window size where one sequence consists of 600 samples, equivalent to 10 seconds of data. The data was classified and labelled into four groups for supervised deep learning, with each group represented by a numerical label: "0" for "Non-fall," "1" for "Backward fall," "2" for "Lateral (side)," and "3" for "Forward fall."

In human activity recognition, the phase of feature extraction holds paramount importance. It enhances system performance by deriving feature vectors capable of discerning various activities. Particularly for continuous data like sensor readings,

the process of feature extraction or selection presents a significant challenge, as noted in [17]. Before inputting the data into the convolution layer for automatic feature extraction, augmentation features were applied by using 3 acceleration vectors by the calculation from eq (1), (2), and (3) (the accelerometer data, gyroscope data, and Magnetic field data). This approach aimed to reduce the complexity of the original features, which can be challenging to analyses or represent fall events.

$$Acc_{svm} = \sqrt{Acc_x^2 + Acc_y^2 + Acc_z^2} \quad (1)$$

$$Gyr_{svm} = \sqrt{Gyr_x^2 + Gyr_y^2 + Gyr_z^2} \quad (2)$$

$$Mag_{svm} = \sqrt{Mag_x^2 + Mag_y^2 + Mag_z^2} \quad (3)$$

Totally, we had 3 acceleration + 3 axis of Acc + 3 axis of Gyr + 3 axis of Mag = 12 vectors / channel (a wearable sensor)

Table 2 shown the summarizes all the features vectors were applied in the convolution layers.

TABLE 2: THE TABLE DESCRIPTION OF FEATURES EXTRACTED FOR REDUCING FEATURES AND IMPROVE PERFORMANCE FOR INPUT INTO CONVOLUTION LAYER.

Number	Feature Vectors	Features Vector Description
$f_1 - f_3$	Acc (x, y, z)	Raw 3 axis of accelerometer
$f_4$	$Acc_{svm}$	Acceleration vectors of 3 axis accelerometer
$f_5 - f_7$	Gyr (x, y, z)	Raw 3 axis of gyroscope
$f_8$	$Gyr_{svm}$	Acceleration vectors of 3 axis gyroscope
$f_9 - f_{11}$	Mag (x, y, z)	Raw 3 axis of magnetometer
$f_{12}$	$Mag_{svm}$	Acceleration vectors of 3 axis magnetometer

After pre-processing and calculated features vectors from three sensors including accelerometer, gyroscope, and magnetometer, we got the total 12 feature vectors per channel or an IMU sensor. And in this study, we have 12 IMU wearable on the body, so we have 144 features vectors. 144 features vector will be input to convolution layer.

### C. Feature extraction

For machine learning classification, feature extraction is a critical component and an indispensable part. The time domain and frequency domain were applied to extract the features for machine learning. The features will be extracted from the sliding window each slide of the window is 600 sampling (10s /window). In the time domain, we selected 8 important features including maximum, minimum, standard deviation, the sum of absolute, root mean square (RMS), mean, the difference between max and min values, and maximum difference between consecutive values along each row. For the frequency domain, we selected 10 important features such as maximum, minimum, standard deviation, sum of absolute, root mean square, kurtosis, skewness, mean, difference between max and min values, and maximum difference between consecutive values along each row of the FFT (Fast Fourier Transform). After extracting features, we got the features from each IMU

sensor as 12 vectors (from Table 2) x 18 features (from time and frequency domain) = 216 features/channel (a wearable placement). In this case, we take features from 12 IMU sensors that are wearable on body input to the machine learning part for comparison with the DL models that we proposed, so the total we have 2592 features from 12 IMU sensors. Table 3 shows the data shape for input to ML and DL for performance in this study.

TABLE 3: SHOW AND DESCRIBE THE DATA SHAPE FOR INPUT TO MACHINE LEARNING AND DEEP LEARNING.

Model	Shape of Data	Description
ML	(2871, 2592)	The input data for the machine learning model has 2871 samples, each with 2592 features
DL	(2871, 600, 144)	The input data for the deep learning model has a shape of (2871, 600, 144), indicating 2871 samples, each with 600 data points (sampled at 60 Hz over a 10-second window) and 144 features per data point.

#### D. Data Splitting

After preprocessing the data, it is partitioned into three distinct datasets: the training dataset, valuation dataset, and the testing dataset. In this study, the training dataset is designated to comprise 70% of the entire dataset, leaving the remaining 20% for valuation and 10% for testing purposes. The subsequent code snippet demonstrates the utilization of the 'train\_test\_split' function to randomly divide the data into these two subsets.

#### E. Data Scaler

Data scaling is essential in machine learning to ensure consistent evaluation and efficient operation of algorithms, particularly when dealing with diverse raw data values. Normalization and feature scaling enhance the performance and convergence speed of machine learning models by stabilizing the optimization process during training [18]. This step was applied only to the machine learning part.

#### F. Machine Learning for comparison

We used the machine learning for make the evidence to deep learning was proposed. Four machine learning was applied to comparison such as random forest (RF), support vector machine (SVM), multi-layer perceptron (MLP), and k-nearest neighbors (KNN).

Random Forest (RF), proposed by Breiman (2001), combines multiple decision trees for robust performance in high-dimensional spaces. It is widely used in fields like medicine and signal processing due to its high accuracy and stability.

Support Vector Machine (SVM) is a supervised learning algorithm that maximizes geometric margins to enhance generalization performance and is effective in signal classification through kernel functions like the Gaussian kernel.

k-Nearest Neighbor (kNN) classifies samples based on proximity to training data, making it suitable for real-time signal processing.

Multilayer Perceptron (MLP) leverages deep learning to model complex data relationships, useful in signal processing tasks.

The parameter of all ML that we used, it was applied hyper-parameter find turning such as RF used `n_estimators`: [min=10, max=1000], `max_features`: ["auto", "sqrt", "log2"], and `bootstrap`: [True, False]. For SVM used `C`: [1, 10, 100], `gamma`: ["scale", "auto"], and `kernel`: ["linear", "rbf", "poly"]. And for MLP used `MLPClassifier`(`max_iter`=1000), `hidden_layer_sizes`: [(100,), (100, 50), (150, 100, 50)], `activation`: ["tanh", "relu"], `solver`: ["sgd", "adam"], and `alpha`: [0.0001, 0.05]. and the last one is KNN used `n_neighbors`: [5, 10, 15], `weights`: ["uniform", "distance"], `metric`: ["euclidean", "manhattan", "minkowski"]. The below is show the best parameters we found from find turning hyper-parameters [table 4]. In these cases, the cross-validation using `StratifiedKFold` within `K` = [5, 10].

TABLE 4: THE TABLE SHOWN THE STRUCTURES OF MACHINE LEARNING MODEL AND PARAMETERS.

Models	Structure	Parameters
(RF)	Ensemble of decision trees	<code>n_estimators</code> =100, <code>bootstrap</code> =False, <code>max_features</code> ='sqrt'
(SVM)	Non-linear classification model	<code>kernel</code> ='rbf', <code>C</code> =10, <code>gamma</code> ='scale'
(MLP)	Feedforward neural network	<code>hidden_layer_sizes</code> = (100,), <code>max_iter</code> =1000, <code>activation</code> ='relu', <code>alpha</code> =0.0001, <code>solver</code> ='sgd'
(KNN)	Instance-based learning	<code>n_neighbors</code> =10, <code>metric</code> ='manhattan', <code>weights</code> ='uniform'

#### G. Proposed Deep Learning Architecture

This neural network architecture is designed for sequence classification tasks, particularly suited for multi-class classification. It begins with a Convolutional Neural Network (CNN) segment, which processes the sequential input data through two convolutional layers followed by max pooling to extract relevant features. The first convolutional layer applies 256 filters with a kernel size of 3, while subsequent layers use 64 filters each. The max pooling layers reduce the dimensionality of the feature maps to capture the most salient information within a pool size of 2 and batch normalization layers is applied to normalize the outputs also.

Following the CNN segment, a series of Long Short-Term Memory (LSTM) layers are employed to capture temporal dependencies within the data. Five LSTM layers with 128 units a dropout rate of 0.2 each are stacked to learn intricate patterns across the sequence. The first four LSTM layers return sequences, allowing the model to retain temporal information at each time step, while the final LSTM layer aggregates the



sequential information and outputs a single vector representing the entire sequence.

effectively capture temporal patterns and dependencies over extended periods.

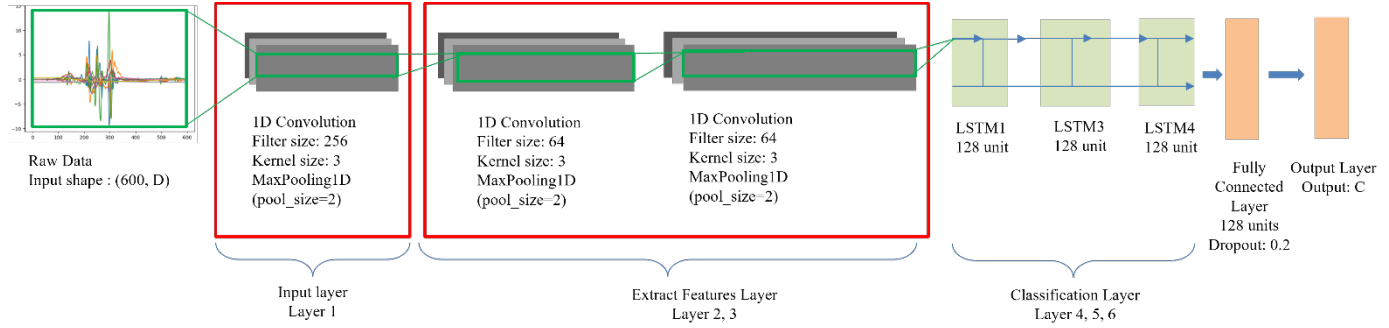


Figure 2: Show the model architecture for Conv-LSTM (D: input data channel and C: number of class activity).

After the LSTM layers, the model incorporates fully connected layers to further process the extracted features. A dense layer with 128 units and rectified linear unit (ReLU) activation function is employed to enhance non-linearity in the data representation. Additionally, a dropout layer with a dropout rate of 0.2 is introduced to mitigate overfitting by randomly deactivating a portion of neurons during training.

The output layer of the model utilizes the SoftMax activation function to generate class probabilities, making it suitable for multi-class classification tasks. The number of output units is determined by the number of classes in the classification problem. The model is trained using the Adam optimizer with a learning rate of 0.0001 and optimized for categorical cross-entropy loss, a common choice for multi-class classification problems.

Overall, this model architecture combines the strengths of CNNs in feature extraction from sequential data and the ability of LSTMs to capture long-term dependencies, making it well-suited for tasks involving sequential data with multiple classes. The model summary provides a concise overview of the network's structure and parameters, aiding in understanding its complexity and performance potential we show it in Figure 2.

#### H. Other DL models

As one of the most classic and representative deep learning algorithms, CNN (Convolutional Neural Network) is not only effective for image processing but also widely used for signal processing. In this context, CNNs are composed of convolutional layers and pooling layers, which work together to extract and learn hierarchical features from signal data, such as time-series signals [19]. This capability allows CNNs to capture important patterns and characteristics within the signal, making them highly effective for tasks such as classification, detection, and analysis of various types of signals.

LSTM (Long Short-Term Memory) is a particular form of RNN (Recurrent Neural Network) that excels at handling time series with long time intervals. It is specifically designed to manage and retain long-term dependencies through its unique gating mechanisms, which help in mitigating the vanishing gradient problem typically encountered in standard RNNs [20]. This makes LSTM especially suitable for tasks involving sequential data, such as signal processing, where it can

The GRU (Gated Recurrent Unit) model is another popular variant of RNN (Recurrent Neural Network), specifically designed to handle long-distance dependencies [21]. It effectively addresses the vanishing gradient problem through its gating mechanism, allowing it to maintain long-term information without the gradients vanishing during backpropagation.

To compare the results of the Conv-LSTM model with other deep learning architectures, we also designed several models, including CNN, LSTM, and GRU. We fine-tuned the hyperparameters of all models to achieve the best performance.

For the CNN model, we experimented with the following hyperparameters:

- Layer 1 and 2: conv1\_filters (32 to 128, step 16), conv1\_kernel (3 to 7, step 2), conv2\_filters (64 to 256, step 32), conv2\_kernel (3 to 7, step 2)
- Dense layer: dense\_units (64 to 256, step 32), dropout\_rate (0.0 to 0.5, step 0.1)

For the LSTM model with two layers, we used:

- Layer 1: lstm1\_units (32 to 128, step 16), lstm1\_dropout\_rate (0.0 to 0.5, step 0.1)
- Layer 2: lstm2\_units (32 to 128, step 16), lstm2\_dropout\_rate (0.0 to 0.5, step 0.1)

For the GRU model, we configured:

- Layer 1: gru1\_units (32 to 128, step 16), gru1\_dropout\_rate (0.0 to 0.5, step 0.1)
- Layer 2: gru2\_units (32 to 128, step 16), gru2\_dropout\_rate (0.0 to 0.5, step 0.1)

Table 5 presents the optimal hyperparameters for each of these models.

TABLE 5: THE BEST PARAMETERS IN CNN, LSTM AND GRU MODEL AFTER TURNING HYPER-PARAMETER.

	Layer Type	Parameter	Output Shape	Activation Function
CNN model	Input	-	(600, 144)	-
	conv1d (Conv1D)	Filter=96 Kernel=3	(None, 598, 96)	ReLU
	MaxPooling1D	Pool size=2	(None, 299, 96)	-
	conv1d (Conv1D)	Filter=256 Kernel=7	(None, 293, 256)	ReLU
	MaxPooling1D	Pool size=2	(None, 148, 256)	-

	flatten (Flatten)	-	(None, 37376)	-
	Full connected (Dense)	Unit=64	(None, 64)	ReLU
	dropout (Dropout)	Dropout rate=0.4	(None, 64)	-
	Output	-	(None, 4)	Softmax
	Total params: 2,606,244 Trainable params: 2,606,244 Non-trainable params: 0			
LSTM model	Input	-	(600, 144)	-
	lstm1 (LSTM)	Unit=128	(None, 600, 128)	Tanh
	dropout (Dropout)	Dropout rate=0.1	-	-
	lstm2 (LSTM)	Unit=112	(None, 112)	Tanh
	dropout (Dropout)	Dropout rate=0.2	-	-
	Output	-	(None, 4)	Softmax
GRU model	Total params: 248,196 Trainable params: 248,196 Non-trainable params: 0			
	Input	-	(600, 144)	-
	gru (GRU)	Unit=80	(None, 600, 80)	Tanh
	dropout (Dropout)	Dropout rate=0.1	-	-
	lstm2 (LSTM)	Unit=48	(None, 48)	Tanh
	dropout (Dropout)	Dropout rate=0.4	-	-
	Output	-	(None, 4)	Softmax
	Total params: 73,156 Trainable params: 73,156 Non-trainable params: 0			

### III. RESULTS AND DISCUSSION

#### A. Distributions of Data Sample

The term "imbalanced data" typically refers to a classification problem wherein the distribution of samples across different classes is unequal. In such scenarios, one class, known as the majority class, comprises a significantly larger number of samples compared to other classes, which are often referred to as minority classes. For instance, in Figure 3, in a dataset pertaining to fall events, the "forward" class might have a considerably higher number of samples compared to the "lateral" (side) class, which might have the fewest samples.

Training models on highly imbalanced datasets presents a challenge because traditional performance metrics assume a balanced distribution of classes [16]. Consequently, evaluating model performance becomes problematic as these metrics may be biased toward the majority class, leading to inaccurate assessments of the model's effectiveness.

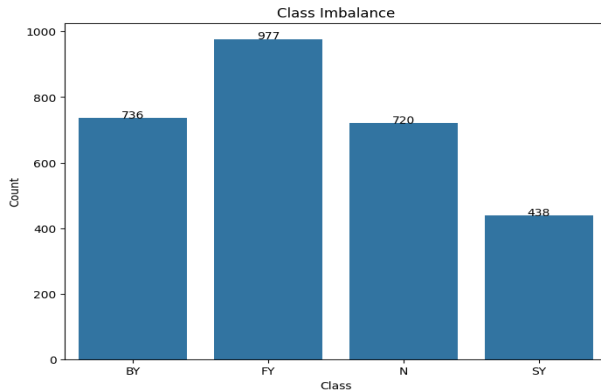


Figure 3: Imbalance data distributions between "N: Non-fall", "BY: Backward-fall", "FY: Forward-fall", and "SY: Lateral (side)"

#### B. Performance Evaluation

In this study, we conducted a comprehensive evaluation of our deep learning model using Conv-LSTM. Our evaluation criteria were based on metrics that include accuracy, recall, precision, and F-score [16]. All experiments in this study were carried out on a PC equipped with 16.0 GB (15.9 GB usable) and Intel(R) Core (TM) i5-6600 CPU @ 3.30GHz 3.30 GHz.

#### C. Results

The dataset has been segregated into two distinct components, each dedicated to specific classes. These components include data from three types of sensors the Accelerometer sensors, Gyroscope sensors, and Magnetic Field sensors. The data has been divided into two main classes Binary Class and Multiclass.

The Binary Class encompasses fall and non-fall events, while the Multiclass, which includes directional information, is further divided into Non-fall, Backward-fall, Forward-fall, and Lateral (Side) events. These classes are presented in Table 6 below.

TABLE 6: SHOW THE SEPARATE DATA INTO TWO DISTINCT CLASS AS BINARY CLASS AND MULTICLASS.

Name dataset	Binary Class	Multi Class
None fall	720	720
Backward	2151	736
Forward		977
Lateral (side)		438
Total	2871	2871

Based on the findings presented in Table 7, we identified several key features that significantly enhance accuracy when integrated into the convolutional layer. These features, derived from the Euclidean magnitude variable, contribute to constructing a robust framework for our wearable sensor technology.

For our performance, we used a traditional method to improve the accuracy of the Conv-LSTM model. Specifically, we compared three parts: the raw data, the norm (Euclidean magnitude), and combined raw with norm vectors. After selecting the norm vector and extracting features using the convolutional model, we achieved higher accuracy, as shown in Table 7.

TABLE 7: THE ACCURACY COMPARISON OF ALL FEATURES IN BINARY CLASS AND MULTI-CLASS.

Classes		Original	Norm	All
Binary Class	Accuracy	98.25	98.95	<b>99.65</b>
	Precision	98.25	98.96	99.65
	Recall	98.25	98.95	99.65
	F1 Score	98.23	98.94	<b>99.65</b>

Multi Classes	Accuracy	96.49	95.09	<b>97.89</b>
	Precision	96.62	95.07	97.91
	Recall	96.49	95.09	97.89
	F1 Score	96.51	95.05	<b>97.89</b>

Based on table 7, in the single class, it shows accuracy of raw data (original) is higher than the accuracy of norm data. So, it means we can use a norm vector for represented to fall and non-fall event. But for multi-classes, the accuracy of original data is higher than the accuracy of the norm, because multi-class has more fall events such as non-fall, fall-backward, fall-forward, and fall-lateral it is different information. So, it needs more information from different sensor axes such as three dimensions from accelerometers, 3 dimensions from gyroscopes, and 3 dimensions from magnetometers that provide the information depending on the fall event. To solve this problem, we combined the raw data with norm vectors and after that, it got high accuracy in both parts (binary class and multi-classes) 99.65 % and 97.89%, respectively.

#### D. The comparison to others method

After conducting Conv-LSTM analysis, we achieved impressive accuracy rates of 99.65% for binary classification and 97.89% for multi-class tasks. To validate the effectiveness of Conv-LSTM, we compared it against various methods including traditional machine learning algorithms like RF, SVM, MLP, and KNN, as well as deep learning models such as CNN, LSTM, and GRU. The results, detailed in Table 9 and 10, display accuracy, precision, recall, and F1 scores. So, let's looking on the table below, it's showed the accuracy, precision, recall, and F1 score. For machine learning evaluation, we employed cross-validation using StratifiedKFold with K values of 5 and 10, utilizing optimal hyperparameters. These results are summarized in Table 8.

TABLE 8: THE CROSS-VALIDATION OF MACHINE LEARNING. IT INCLUDES THE MEAN (STANDARD DEVIATION) ACCURACY AS %. NUMBERS IN BOLD REPRESENT THE BEST PERFORMANCE PER CLASSIFIER.

Cross-Validation		RF (%) Mean (SD)	SVM (%) Mean (SD)	MLP (%) Mean (SD)	KNN (%) Mean (SD)
Binary Class	K=5 Accuracy	94.77 (0.34)	<b>98.00 (0.22)</b>	97.06 (0.80)	93.98 (0.42)
	K=10 Accuracy	94.82 (1.49)	<b>98.01 (0.71)</b>	97.56 (0.93)	94.33 (0.84)
Multi Classes	K=5 Accuracy	85.12 (1.31)	<b>92.68 (0.55)</b>	91.49 (0.48)	83.67 (0.66)
	K=10 Accuracy	85.66 (1.82)	<b>93.13 (1.27)</b>	91.29 (1.98)	84.57 (2.17)

Following cross-validation, we obtained reliable parameters which were then utilized to re-evaluate accuracy, precision, recall, and F1 scores for comparison with Conv-LSTM. This comparison, depicted in Table 9, includes time consumption metrics.

TABLE 9: THE COMPARISON BETWEEN SEVERAL MACHINE LEARNING AND OUR PROPOSED METHOD SUCH AS RF, SMV, MLP, AND CONV-LSTM AND SHOW TIME CONSUMPTION WITHIN BINARY AND MULTI-CLASSES.

Classes		KNN	RF	MLP	SVM	Conv-LSTM
	Accuracy	93.39	94.55	97.45	<b>97.80</b>	<b>99.65</b>

Binary Class	Precision	93.31	94.50	97.47	97.79	99.65
	Recall	93.34	94.55	97.47	97.80	99.65
	F1 Score	93.26	94.52	97.46	<b>97.79</b>	<b>99.65</b>
	Train (s)	0.005	11.43	13.00	1.404	388.463
	Test (s)	0.361	0.016	0.006	<b>0.932</b>	<b>0.490</b>
Multi Classes	Accuracy	83.76	87.12	92.81	<b>94.78</b>	<b>97.89</b>
	Precision	83.50	87.64	92.71	94.74	97.91
	Recall	83.76	87.12	92.81	94.78	97.89
	F1 Score	83.35	86.32	92.74	<b>94.71</b>	<b>97.89</b>
	Train (s)	0.006	97.46	16.30	03.25	384.641
	Test (s)	1.355	0.185	0.006	<b>2.354</b>	<b>0.484</b>

In the machine learning part, we utilized Grid Search CV to fine-tune hyperparameters, as explained earlier. Notably, SVM exhibited the highest accuracy of 97.80%, and 94.78% for binary class and multi-classes, with a corresponding F1-score of 97.79% and 94.71%, respectively when compared with other several ML. But our proposed method got accuracy higher than SVM and its testing time was notably shorter. The results, as shown in Table 9, highlight Conv-LSTM's superiority across various metrics, including accuracy, precision, recall, F1 score, and testing time, in both binary and multi-class scenarios.

TABLE 10: THE COMPARISON BETWEEN SEVERAL DEEP LEARNING AND OUR PROPOSED METHOD SUCH AS CNN, LSTM, GRU, AND CONV-LSTM AND SHOW TIME CONSUMPTION WITHIN BINARY AND MULTI-CLASSES

Classes		LSTM	GRU	CNN	Conv-LSTM
Binary Class	Accuracy	96.49	95.79	96.86	<b>99.65</b>
	Precision	96.47	95.79	96.85	99.65
	Recall	96.49	95.79	96.84	99.65
	F1 Score	96.46	95.79	96.80	<b>99.65</b>
	Train (s)	802.75	550.09	95.049	388.463
	Test (s)	0.347	0.308	0.238	<b>0.490</b>
Multi Classes	Accuracy	82.81	91.23	93.68	<b>97.89</b>
	Precision	82.18	91.17	93.81	97.91
	Recall	82.81	91.23	93.68	97.89
	F1 Score	82.40	91.18	93.70	<b>97.89</b>
	Train (s)	1773.4	550.768	88.762	384.641
	Test (s)	0.339	0.285	0.225	<b>0.484</b>

In the deep learning part, we used Kernel Turn Random Search to fine-tune hyperparameters, as explained earlier. Conv-LSTM exhibited the highest accuracy of 99.65%, and 97.89% for binary and multi-classes, with a corresponding F1-score of 99.65%, and 97.89%, respectively. While Conv-LSTM required testing time was notably longer than other several deep learning. But in these cases, we tried to classify fall events, so we would not be concerned more about training time. Although it required a longer time for the training part. however, the testing time is less than 1s is suitable for use for performance in real-world situations.

#### IV. CONCLUSION

Our study makes significant contributions to biomedical knowledge discovery and engineering by advancing the field of

fall detection using wearable sensor technology. Through meticulous research and experimentation, we have achieved several scientific advancements. Advanced Feature Extraction for Falls, we developed optimal feature extraction techniques, leveraging Convolutional Neural Networks (CNN) to capture detailed information about fall events. This approach reduces redundant data, ensuring low power consumption, fast computation, and high accuracy. Unlike traditional methods, our approach utilizes automated feature extraction, overcoming challenges associated with manual feature extraction. Innovative Use of Conv-LSTM, our proposed solution integrates Convolutional Neural Networks (CNN) with Long Short-Term Memory (LSTM) networks in the Conv-LSTM architecture. This innovative model enables precise classification of fall event directions by capturing both spatial and temporal features. Enhanced Accuracy Through Data Fusion: By fusing raw data with norm vectors derived from the Euclidean magnitude variable, we achieved significant improvements in accuracy for both binary and multi-class classification tasks. This fusion of data types enhances the overall performance of our wearable sensor technology. Superior Performance in Model Comparison: We validated the effectiveness of Conv-LSTM by comparing it against various traditional machine learning algorithms and other deep learning models. Our results demonstrate that Conv-LSTM outperforms these methods in terms of accuracy, precision, recall, F1 score, and testing time for both binary and multi-class tasks. Efficient Real-Time Fall Detection: Despite requiring longer training time, the Conv-LSTM model maintains a testing time under 1 second, making it suitable for real-time fall detection applications. This efficiency is crucial for practical deployment in healthcare settings, where timely intervention can prevent further injuries.

Overall, this study provides valuable insights into the optimal selection and integration of features for improving the performance of Conv-LSTM models in wearable sensor technology applications, paving the way for more effective and reliable real-world implementation applications in healthcare.

#### ACKNOWLEDGMENT

This research was supported by the MSIT (Ministry of Science and ICT), Korea, under the ICAN (ICT Challenge and Advanced Network of HRD) support program (IITP-2024-2020-0-01832), supervised by the IITP (Institute for Information & Communications Technology Planning & Evaluation).

#### REFERENCES

- [1] World Health Organization, *Strategies for preventing and managing falls across the life-course*. Accessed:27-April-2021, Available [online].
- [2] M. M. Islam et al., "Deep Learning Based Systems Developed for Fall Detection: A Review," in *IEEE Access*, vol. 8, pp. 166117-166137, 2020, doi:10.1109/ACCESS.2020.3021943.
- [3] Velusamy A, Akilandeswari J, Prabhu R. (2023) "A Comprehensive Review on Machine Learning Models for Real Time Fall Prediction using Wearable Sensor-based Gait Analysis". doi:10.1109/ICIRCA57980.2023.10220663
- [4] S. Katz, A. B. Ford, R. W. Moskowitz, B. A. Jackson, and M. W. Jaffe, "Studies of illness in the aged: The index of ADL: A standardized measure of biological and psychosocial function," *JAMA*, vol. 185, pp. 914-919, Sep. 1963.
- [5] W. Saadeh, S. A. Butt and M. A. B. Altaf, "A Patient-Specific Single Sensor IoT-Based Wearable Fall Prediction and Detection System," in *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 27, no. 5, pp. 995-1003, May 2019, doi:10.1109/TNSRE.2019.2911602.
- [6] Pierleoni, P., Belli, A., Maurizi, L., Palma, L., Pernini, L., Panicia, M., & Valenti, S. (2016). A wearable fall detector for elderly people based on AHRS and barometric sensor. *IEEE sensors journal*, 16(17), 6733-6744. doi: 10.1109/JSEN.2016.2585667
- [7] Yves M. Galvão, Janderson Ferreira, Vinicius A. Albuquerque, Pablo Barros, Bruno J.T. Fernandes, A multimodal approach using deep learning for fall detection, *Expert Systems with Applications*, Volume 168, 2021, 114226, ISSN 0957-4174, doi: <https://doi.org/10.1016/j.eswa.2020.114226>.
- [8] Roberto Paoli, Francisco J. Fernández-Luque, Ginés Doménech, Félix Martínez, Juan Zapata, Ramón Ruiz, A system for ubiquitous fall monitoring at home via a wireless sensor network and a wearable mote, *Expert Systems with Applications*, Volume 39, Issue 5, 2012, Pages 5566-5575, ISSN 0957-4174, <https://doi.org/10.1016/j.eswa.2011.11.061>.
- [9] Shehroz S. Khan, Babak Taati, Detecting unseen falls from wearable devices using channel-wise ensemble of autoencoders, *Expert Systems with Applications*, Volume 87, 2017, Pages 280-290, ISSN 0957-4174, <https://doi.org/10.1016/j.eswa.2017.06.011>.
- [10] H. Sadreazami, M. Bolic and S. Rajan, "Contactless Fall Detection Using Time-Frequency Analysis and Convolutional Neural Networks," in *IEEE Transactions on Industrial Informatics*, vol. 17, no. 10, pp. 6842-6851, Oct. 2021, doi: 10.1109/TII.2021.3049342.
- [11] Tianhu Wang, Baoqiang Wang, Yunzhe Shen, Yang Zhao, Wenjie Li, Keming Yao, Xiaojie Liu, Yinsheng Luo, Accelerometer-based human fall detection using sparrow search algorithm and back propagation neural network, *Measurement*, Volume 204, 2022, 112104, ISSN 0263-2241, <https://doi.org/10.1016/j.measurement.2022.112104>.
- [12] Li, Z., Liu, Y., Guo, X., & Zhang, J. (2020, November). Multi-convLSTM neural network for sensor-based human activity recognition. In *Journal of Physics: Conference Series* (Vol. 1682, No. 1, p. 012062). IOP Publishing. doi: 10.1088/1742-6596/1682/1/012062.
- [13] Kraft, D.; Srinivasan, K.; Bieber, G. Deep Learning Based Fall Detection Algorithms for Embedded Systems, Smartwatches, and IoT Devices Using Accelerometers. *Technologies* 2020, 8, 72. <https://doi.org/10.3390/technologies8040072>
- [14] Y. Nam and J. W. Park, "Child Activity Recognition Based on Cooperative Fusion Model of a Triaxial Accelerometer and a Barometric Pressure Sensor," in *IEEE Journal of Biomedical and Health Informatics*, vol. 17, no. 2, pp. 420-426, March 2013, doi: 10.1109/JBHI.2012.2235075
- [15] Rosenfeld, Jonathan S. "Scaling laws for deep learning." arXiv preprint arXiv:2108.07686 (2021).
- [16] Singh, J.; Singh, N.; Fouda, M.M.; Saba, L.; Suri, J.S. Attention-Enabled Ensemble Deep Learning Models and Their Validation for Depression Detection: A Domain Adoption Paradigm. *Diagnostics* 2023, 13, 2092. <https://doi.org/10.3390/diagnostics13122092>
- [17] Ahmed, N.; Rafiq, J.I.; Islam, M.R. Enhanced Human Activity Recognition Based on Smartphone Sensor Data Using Hybrid Feature Selection Model. *Sensors* 2020, 20, 317. <https://doi.org/10.3390/s20010317>
- [18] Henry Friday Nweke, Ying Wah Teh, Mohammed Ali Al-garadi, Uzoma Rita Alo, Deep learning algorithms for human activity recognition using mobile and wearable sensor networks: State of the art and research challenges, <https://doi.org/10.1016/j.eswa.2018.03.056>.
- [19] Wu, Xiaodan & Cheng, Lingyu & Chu, Chao & Kim, Jungyoon. (2019). Using Deep Learning and Smartphone for Automatic Detection of Fall and Daily Activities. 10.1007/978-3-030-34482-5\_6.
- [20] Sepp Hochreiter, Jürgen Schmidhuber; Long Short-Term Memory. *Neural Comput* 1997; 9 (8): 1735–1780. doi: <https://doi.org/10.1162/neco.1997.9.8.1735>
- [21] Malhotra, Pankaj, et al. "Long Short Term Memory Networks for Anomaly Detection in Time Series." *Esann*. Vol. 2015. 2015.

# Golden Search Optimization with Deep Learning Enabled Computer-Aided Diagnosis for Bone Cancer Classification

Ashit Kumar Dutta<sup>1</sup>, Shtwai Alsubai<sup>2</sup>, Basit Qureshi<sup>3</sup>, Naved Ahmad<sup>4</sup>, Seifedine Kadry<sup>5,6,7</sup>, Yunyoung Nam<sup>8,\*</sup> and Jinseok Lee<sup>9,\*</sup>

<sup>1</sup>Department of Computer Science and Information Systems, College of Applied Sciences, AlMaarefa University, Ad Diriyah, Riyadh, Kingdom of Saudi Arabia; [adotta@mcst.edu.sa](mailto:adotta@mcst.edu.sa)

<sup>2</sup>Department of Computer Science, College of Computer Engineering and Sciences in Al-Kharj, Prince Sattam bin Abdulaziz University, Al-Kharj, Kingdom of Saudi Arabia; [Sa.alsubai@psau.edu.sa](mailto:Sa.alsubai@psau.edu.sa)

<sup>3</sup>Department of Computer Science, Prince Sultan University, Riyadh, Kingdom of Saudi Arabia; [Qureshi@psu.edu.sa](mailto:Qureshi@psu.edu.sa)

<sup>4</sup>Department of Computer Science and Information Systems, College of Applied Sciences, AlMaarefa University, Ad Diriyah, Riyadh, Kingdom of Saudi Arabia; [nahmad@mcst.edu.sa](mailto:nahmad@mcst.edu.sa)

<sup>5</sup>Department of Applied Data Science, Noroff University College, Kristiansand, Norway; [skadry@gmail.com](mailto:skadry@gmail.com)

<sup>6</sup>Artificial Intelligence Research Center (AIRC), Ajman University, Ajman, 346, United Arab Emirates

<sup>7</sup>Department of Electrical and Computer Engineering, Lebanese American University, Byblos, Lebanon

<sup>8</sup>Department of ICT Convergence, Soonchunhyang University, Asan, 31538, Korea; [ynam@sch.ac.kr](mailto:ynam@sch.ac.kr)

<sup>9</sup>Department of Biomedical Engineering, Kyung Hee University, Yongin, Republic of Korea; [gonasago@khu.ac.kr](mailto:gonasago@khu.ac.kr)

**Abstract**— Bone cancer is considered a life-threatening health problem, and, in most cases, it leads to death. Physicians use CT-scan, X-rays, or MRI images to recognize bone cancer. The manual process requires expertise and is a tedious task. Hence, it is crucial to establish an automatic system to identify and classify healthy bone and cancerous bones. The earlier diagnosis is the only factor that increases the probability of survival of cancer-affected patients. Artificial intelligence (AI), particularly deep learning (DL) with convolutional neural network (CNN) has shown great promise in categorizing two-dimensional images of some common diseases and depending on databases of millions of annotated or unannotated images. The study presents a new golden search optimization with deep learning enabled computer-aided diagnosis for bone cancer classification (GSODL-CADBCC) on X-ray images. The GSODL-CADBCC technique accurately distinguishes the input X-ray images into healthy and cancerous ones. To accomplish this, the presented GSODL-CADBCC technique uses the bilateral filtering (BF) technique to remove the noise. The presented GSODL-CADBCC technique exploits the SqueezeNet model to produce feature vectors, and the golden search optimization (GSO) algorithm proficiently chooses the hyperparameters. Finally, the extracted features can be classified by improved cuckoo search (ICS) with long short-term memory (LSTM) model. Compared with recent DL models, the experimental outcomes demonstrated that the

**GSODL-CADBCC technique achieves promising performance on bone cancer classification.**

**Keywords:** Bone cancer; Computer-aided diagnosis; Medical imaging; Deep learning; Golden search optimization

## 1. INTRODUCTION

Generally, bones are attached to the body's muscles and support the movements [1]. Bone ligaments were filled with spongy bone marrow, a fibrous tissue. Bone cancer originated from healthy cells and formed as cancer [2]. Cancer grows slowly and spreads to other body parts. It might destroy bone tissues, and bone will be weaker. Initial identification is the only factor that increases the possibility of living for cancer-affected patients. Differential diagnoses of bone cancer are based on the assessment of the age of the patient and traditional radiographs [3]. The plain radiograph is the most useful analysis for distinguishing such cases. Bone scans, computed tomography (CT), magnetic resonance imaging (MRI), and positron emission tomography (PET) were more delicate in identifying bone cancers. Still, the commonly used advanced imaging modality is time taking and costly. In addition to demographic data like the matrix type, patient's age, location, periosteal reaction radiographic appearance of the tumours involving size, cortical destruction, and margin are other important clues assisting the radiologist in distinguishing indolent from aggressive bone cancers [4]. Since bone tumours



are relatively uncommon and have various appearances, some radiologists have developed sufficient expertise to make definitive diagnoses. Among radiologists, precision in the analysis of bone lesions is low, resulting in misdiagnoses which are harmful to patient outcomes [5].

Early classification and detection are done to ensure the effective treatment of bone cancers [6]. For primary malignancy, radical surgical resection can be possible only if they are identified in the initial levels. Radiotherapy, protected weight-bearing or surgical augmentation is required to prevent fractures around osteolytic cancers [7]. Though plain radiographs were commonly utilized for routine screening for bone cancers, a higher rate of misdiagnoses upon visual inspection was stated, as bone cancers show common ambiguous features and various morphologies. Artificial intelligence (AI) technology developments present novelties in medical data analysis [8]. Deep learning (DL), a higher-level NN resembling the human brain, overcomes complicated issues that low-level AI cannot. Convolutional neural networks (CNNs) methods have demonstrated superior outcomes in analyzing healthcare images with complicated paradigms. The capability of DL in interpreting 2D medical images is similar to an average human specialist in the domain. Several research works have reported outstanding outcomes in classifying or diagnosing a disease utilizing microscopy, plain radiography, MRI, ultrasound, endoscopy and CT [9, 10]. In this context, AI technology is utilized for detecting bone cancer on plain radiographs. When the AI-related classification system executes well in medical practice, human errors, time, and cost are drastically minimized.

This study presents a new golden search optimization with the deep learning-enabled computer-aided diagnosis for bone cancer classification (GSODL-CADBCC) on X-ray images. The presented GSODL-CADBCC technique uses bilateral filtering (BF) to remove the noise. The presented GSODL-CADBCC technique exploits the SqueezeNet model to produce feature vectors, and the hyperparameters are proficiently chosen by the golden search optimization (GSO) algorithm. Finally, the extracted features can be classified by improved cuckoo search (ICS) with long short-term memory (LSTM) module. A wide range of experiments was performed to study the performance of the GSODL-CADBCC technique on medical imaging datasets. In short, the key contributions are listed as follows.

- An intelligent GSODL-CADBCC technique encompasses BF based preprocessing, SqueezeNet feature extraction, GSO based hyperparameter tuning, LSTM classification, and ICS based parameter optimization is presented. To the best of our knowledge, the GSODL-CADBCC model has been never presented in the literature.
- Hyperparameter optimization of the SqueezeNet and LSTM models using GSO and ICS algorithms using cross-validation helps to boost the predictive outcome of the proposed model for unseen data.

## 2. RELATED WORKS

Georgeanu et al. [11] present a methodology for predicting bone cancer malignancy based on MRI. The datasets contain MRI scans with diagnoses confirmed by the histopathological investigation. Two pre-trained residual CNN will categorize the image extracted from the MRI datasets for T1 and T2. With the aid of pretrained CNN, the algorithm converges faster during training and trains on smaller data. Liu et al. [12] validate and build machine learning and deep learning fusion mechanism for classifying intermediate, benign, and malignant bone cancers based on patient clinical features and standard radiographs of the tumor. The machine learning (ML) technique fuses data from radiographs and clinical characteristics and could enhance the classification ability for bone tumors.

Wang et al. [13] used a DL technique for automatically counting cells in colour bone marrow microscopic images. The presented algorithm makes use of a Feature Pyramid Network and a Fast region based convolutional neural network (RCNN) to create a model that considers different illumination levels and is accountable for the stability of the colour component. The empirical result shows that the presented model is similar to some state-of-art systems. A user interface enables pathologists to operate the system easily. Calin et al. [14] introduced a fast methodology of object-based classification, which facilitates easy interpretation of bone scintigraphy images. Firstly, the full lambda-schedule technique, along with an edge-based segmentation technique, has been applied for identifying the objects in the bone scintigraphy, and the spatial and textural features of the object were evaluated. Then, a group of objects were chosen as training datasets based on visual inspection of an image and were allocated to different stages of radionuclide accumulation beforehand, implementing the data classifications using support vector machine (SVM) and k-nearest neighbor (KNN) classifiers.

Georgeanu et al. [15] present a CNN architecture in addition to an image processing technique for detecting and classifying bone MRI scans into benign or malignant tumors. With the help of the transfer learning technique, the performance of both pre-trained CNN architectures, namely ResNet-50 and VGG-16 models, were compared. Han et al. [16] assessed the performance of the DL classifier for bone scans of prostate cancer patients. Two distinct 2D CNN frameworks have been introduced: (1) tandem architecture incorporating entire body and local patches, termed as “global–local unified emphasis” (GLUE) and (2) whole body–based (WB). Eweje et al. [17] designed a DL approach which could discriminate benign and malignant bone cancers with patient demographics and MRI. The image-based model was produced through the EfficientNet-B0 framework, and logistic regression (LR) was trained through lesion position, patient age, and sex. At last, a voting ensemble model was constructed.

Sushmitha and Jagadeesh [18] drive of this work is to connect the efficiency of CNN and KNN techniques in a new detection of bone cancers. CNN and K- KNN techniques can be utilized for recognising a 20 instance bone MRI collection. The authors [19] presents the recognition of bone tumor in the database obtained in the medical database. A primary step was extraction feature of segmentation bone image utilizing GLCM approach was executed for extracting features with



respect to statistical texture-based and the secondary step was classifier of extraction feature utilizing K-NN with DT system. Rajagopal et al. [20] examines several approaches of medicinal image processing and DL and executes them for classifying and detecting cancers as benign/malignant. Afterward, the segmentation of cancer, classifier of benign and cancerous cells was complete with utilize of DL approach based CNN technique. Ranjitha et al. [21] presented a bone malignant growth detection exploiting k-means segmentation and KNN technique for recognizing the bone disease exploiting image processing system to ultra sound images of bones.

Several CAD models are existed in the literature to perform the classification process. Though several ML and DL models for liver cancer classification are available in the literature, it is still needed to enhance the classification performance. Owing to the continual deepening of the model, the number of parameters of DL models also increases quickly, which results in model overfitting. At the same time, different hyperparameters significantly impact the efficiency of the CNN model. The hyperparameters such as epoch count, batch size, and learning rate selection are essential to attain effectual outcomes. Since the trial and error method for hyperparameter tuning is tedious and erroneous, metaheuristic algorithms can be applied. Therefore, in this work, we employ GSO and ICS algorithms for the parameter selection of the SqueezeNet and LSTM models, respectively.

### 3. THE PROPOSED MODEL

In this study, we have introduced an effective GSODL-CADBCC technique for automated bone cancer classification on X-ray images. The presented GSODL-CADBCC technique encompasses BF-based noise removal, GSO with SqueezeNet-based feature extraction, LSTM classification, and ICS-based hyperparameter tuning. Fig. 1 exhibits the working principle of the GSODL-CADBCC technique. The figure states that the input X-ray image is primarily preprocessed using the BF filter. The feature vectors are produced using the SqueezeNet model with a GSO-based hyperparameter tuning strategy. Finally, the ICS with LSTM model determines the cancerous and normal image.

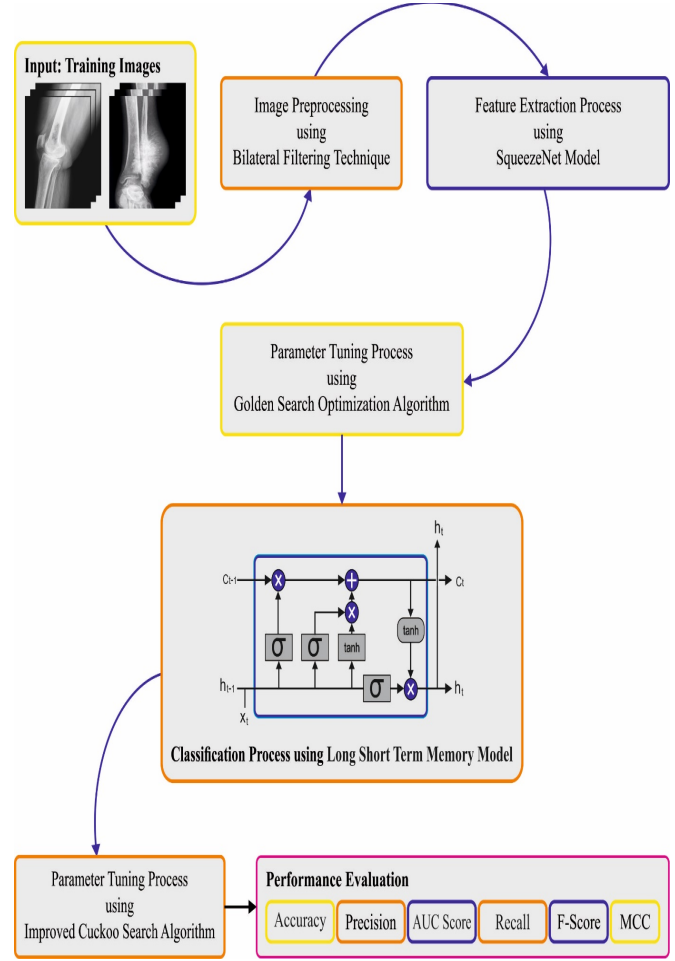


Figure 1. Workflow of GSODL-CADBCC model.

#### 3.1. Image Preprocessing using BF Technique

Firstly, using the BF technique, the GSODL-CADBCC technique eradicates the noise in the X-ray images. BF algorithm is a nonlinear filtering technique that completely considers spatial and pixel value information and forms a compromise processing on the image [22]. BF algorithm is used to decrease the noise effects, process images and preserve the edges. In the local region, they are narrow and always long which is not similar to the noise. The following equation evaluates the results of the BF algorithm.

$$i_{\text{new}}(x, y) = \frac{\sum_{a=x-c}^{x+c} \sum_{b=y-c}^{y+c} i(a, b) \omega(a, b)}{\sum_{a=x-c}^{x+c} \sum_{b=y-c}^{y+c} \omega(a, b)} \quad (1)$$

where  $i_{\text{new}}(x, y)$  denotes the pixel value attained afterwards applying the BF algorithm.  $c$  indicates the window size that affects the computation of  $\omega(a, b)$ ,  $i(a, b)$  shows the pixel value of a specific point in the window, and  $\omega(a, b)$  embodies the value computed by two Gaussian functions. The initial Gaussian function is evaluated using Eq. (2),

$$= \exp\left\{-\frac{(a-x)^2 + (b-y)^2 - \omega_s(a,b)}{2\sigma_s^2}\right\} \quad (2)$$

where  $\omega_s(a,b)$  signifies variable  $\sigma_s$  control the initial Gaussian function and  $\omega_s(a,b)$  the value. The second Gaussian function is evaluated using Eq. (3).

$$\omega_v(a,b) = \exp\left\{-\frac{(i(a,b) - i(x,y))^2}{2\sigma_v^2}\right\} \quad (3)$$

where  $\omega_v(a,b)$  represents the second Gaussian function.

### 3.2. Feature Extraction using Optimal SqueezeNet Model

In this phase, the preprocessed images are fed as input to the SqueezeNet model to generate feature vectors. Generally, CNN involves fully connected (FC), convolutional, and pooling layers [23]. Firstly, the feature was extracted with the help of convolutional and pooling layers. Then, the feature map in the convolution layer is transformed into 1D vector. Next, the resulting layer classifies the input images. The network reduces the square divergence between the classifier and predictive outcomes and changes the weighted parameter through BP. The neuron in each layer is orderly in 3D height, width, and depth where depth determines the channel count of input images or amount of input feature map, and width and height characterize the size of neuron. The convolution layer encompasses multiple convolutional filters, which extract features in the image with convolution model. The convolution filter of the present layer convoluted the input feature map to realize the resulting feature map and remove local features. Next, the nonlinear feature map was realized by the activation function. The subsampling or pooling layers are last convolution layers. It employs downsampling method is a particular value as outcomes in specific area. Fig. 2 shows the architecture of SqueezeNet model.

The parameter count for VGGNet and AlexNet optimize, the SqueezeNet architecture is developed that is minimal parameter but maintained accuracy. The fire element develops a significant component in SqueezeNet. This component was detached to Squeeze and Expand architecture. The  $1 \times 1$  convolution layer attained substantial consideration. Consequently, accomplishing linear combination of many feature maps and data integration on the channel. When the output and input channel count is greater, then convolution kernels develop greater. Next, add  $1 \times 1$  convolution to every inception system that reduces the input channel count, and the convolution kernel variable and complex function were decreased. Finally, add  $1 \times 1$  convolution layer to improve the count of channels and the feature extraction. a superior activation graph was presented to convolution layer once the sampling reduction technique is delayed, hence the highest activation graph reserve further data and provides improved efficacy of the classifier.

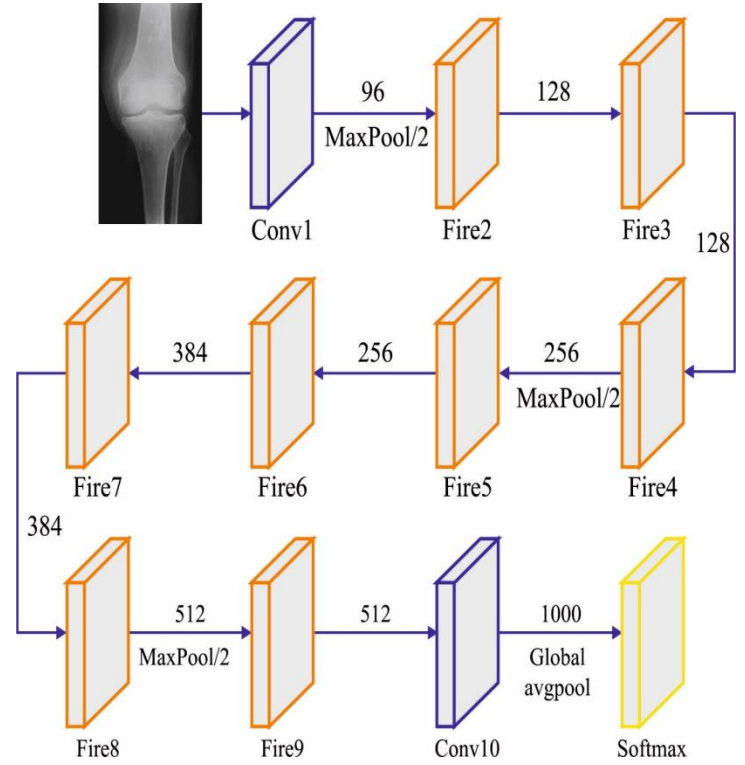


Figure 2. Structure of SqueezeNet.

To improve the efficiency of the SqueezeNet module, the GSO is used. GSO includes exploitation and exploration stages and provides a balance between two conflicting capabilities. The model comprises three major phases namely initialization, evaluation, and updating the existing population. The proposed GSO steps are given in the following [24]:

$$O_i = lb_i + \text{rand} \times (ub_i - lb_i) ; i = 1, 2, \dots, N \quad (4)$$

In Eq. (4),  $O_i$  present the position of  $i$ -th objects in the search range. Furthermore,  $ub_i$  and  $lb_i$  denotes the lower and upper boundaries of the object, correspondingly. During evaluation, the initial population is assessed by using objective function and the objects with better fitness values are chosen as  $O_{gbest_i}$ . In the golden change phase, the object is sorted based on the fitness and the object with worst fitness is changed by the random solution. In every iteration, the object was moved towards the better solution with the help of step size operators ( $St_i$ ).  $St$  equation comprises three phases. The initial part characterizes the preceding value of step size that is multiplied by the (T) transform operator that is iteratively reduced for balancing global and local search space. The next part presents the distance between the present location of  $i$ -th objects and their personal better location attained by the cosine of random number within  $[0, 1]$ .

The last part signifies the distance between  $i$ -th object's present location and the better location attained, multiplied by the sine of randomly generated numbers amongst  $[0, 1]$ . In the initial iteration,  $St_i$  would be randomly generated and upgraded through subsequent formula as follows [24]:

$$St_i(t+1) = T \cdot St_i(t) + C_1 \cdot \cos(r_1) \cdot (Obest_i - x_i(t)) + C_2 \cdot \sin(r_2) \cdot (Ogbest_i - x_i(t)) \quad (5)$$

In Eq. (5),  $C_1$  and  $C_2$  indicates a random integer within  $[0,1]$ ,  $r_1$  and  $r_2$  denotes random number within  $[0,1]$ ,  $Obest_i$  denotes the best prior location attained by the  $i$ -th objects and  $T$  indicates transfer operator that transform from exploration to exploitation search for improving the performance and controlling the balance between global and local search in eaalir and later iterations.  $T$  is a reducing function and is estimated as

$$T = 100 \times \exp(-20 \times \frac{t}{t_{\max}}) \quad (6)$$

where  $t_{\max}$  indicates the maximal amount of iterations. At every iteration, the approach proceeds by regulating the distance that every object moves in all the dimensions of the problem space. Eq. (5) demonstrates that the step size is a stochastic parameter and might enable the object to follow wider cycle in the problem space. A reasonable interval is presented for controlling this oscillation and to prevent divergence and explosion, to clamp the object movement based on the following expression:

$$-St_{\max} \leq St_i \leq St_{\max} \quad (7)$$

In Eq. (7),  $St_{\max}$  indicates a selected maximal movement allowed, which determines the maximal change one object might undergo in its positional coordinate during the iteration as follows:

$$St_{\max} = 0.1 \times (ub_i - lb_i) \quad (8)$$

Finally, the object moves to the global optimal in the search range. The fitness selection is a crucial factor in the GSO approach for the hyperparameter tuning process. Solution encoding is exploited for assessing the aptitude (goodness) of candidate solution. Now, the accuracy value is the main condition utilized for designing a fitness function.

$$\text{Fitness} = \frac{TP}{TP + FP} \quad (9)$$

From the expression,  $TP$  represent the true positive and  $FP$  denotes the false positive value.

### 3.3. Bone Cancer Classification

For automated recognition of bone cancer, the LSTM model is used. LSTM is an innovative structure in recurrent neural network (RNN) that is frequently employed to handle time series tasks. In comparison to conventional feed forward neural network (FFNN), LSTM has memory ability and allows the data stream for continuing to flow inside network [25]. It is capable of linking the prior data to the existing problem. However, the past status data is useful for deciding the present state of equipment. LSTM has four major components, viz., output gate, input gate, forget gate and cell state. The data of cell state is updated by using three gates that automatically make LSTM add or remove data to neural network. Forget gate decides how much data is forgotten from the prior cell state  $C_{t-1}$  and it is mathematically modelled as:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \quad (10)$$

$h_{t-1}$  and  $x_t$  characterize hidden state at  $t-1$  time and input feature at  $t$  time, correspondingly. In the study, the input feature is the output of CNN. Every component in feature map is regarded as the input of LSTM in one moment.  $W_f$  and  $b_f$  denotes weight parameter and bias term of forgetting gate layer.  $\sigma$  represents the sigmoid function. The output of  $f_t$  value lies within 0 to 1. 1 shows the data of the cell state is fully retained, whereas 0 indicates the data is dropped out completely. Input gate determines how much of newly learned data is added to the present cell state  $C_t$ :

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \quad (11)$$

It is evaluated by nonlinearly mapping prior hidden states and existing input features. Following, the input gate selects partial data from  $\tilde{C}_t$  as follows:

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \quad (12)$$

The meaning of parameter in Eq. (12) is same as in Eq. (10). Combining Eq. (10)-(12), the present cell state  $C_t$  is evaluated by:

$$C_t = f_t \odot C_{t-1} + i_t \odot \tilde{C}_t \quad (13)$$

The present cell state is made up of two terms, viz.,  $f_t \odot C_{t-1}$  and  $i_t \odot \tilde{C}_t$ . The initial one signifies the filtered data afterwards being rejected, and the last one represents the recently generated feature data. The final gate, viz., output gate, defines which part of the cell state is going to output. The last output of hidden layer  $h_t$  is evaluated by:

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \quad (14)$$

$$h_t = o_t \odot \tanh(C_t) \quad (15)$$

From above equations, it is clearly understood that the abovementioned three nonlinear gates regulated the data flow in and out of the LSTM network. Furthermore, the hidden state  $h_t$  has every historical state data from time 0 to  $t$ . The time series data is useful for constructing precise health indicators.

Finally, the ICS algorithm fine tune the hyperparameter of the LSTM module to achieve maximum accuracy. CS is derived from the lifestyle of bird family named cuckoo. The special lifestyle of this bird and its characteristics in breeding and egg laying are the primary motivation for the improvement of new evolutionary optimization techniques [26]. CS begins with early population that has matured cuckoo and their eggs. Every matured cuckoo has its egg laying radius (ELR) and lays eggs in its ELR in host bird habitat. Next, some eggs that are less like host bird eggs will be demised.

Survived cuckoo population will be migrated to the best location. The first cuckoo based on the location in environments is separated into some clusters and lastly, every cuckoo will be migrated towards the better cuckoo in the better cluster with  $\lambda$  percentage and with the deviation of  $\alpha$  radians. Once each position of cuckoo became closer, the process ends. Being closer means that this environment has maximum food source and fewer eggs will demise. But the original CS can able to resolve a large number of problems, by shifting some of its operations, and based on the problems, we could enhance its overall performance. In the proposed model, local search technique is added termed "Simulated Annealing" and tried few operations of CS are to enhance it. The outcomes

demonstrated that the modification employed on the CS enhanced the performance. The changes are given below:

(1) Adaptive ELR: At the highest iteration, the estimated solution should be altered lesser than the prior iteration, and the size of ELR should be decreased. It shows that eggs aren't laid farther away from the parents. In the presented method, the size of ELR is decreased while the amount of iterations is improved.

(2) Adaptive number of eggs: the number of eggs of every cuckoo is randomly defined, between the minimum and the maximum number of eggs. In the presented method, it is potential that a weaker cuckoo might lay more eggs when compared to the stronger one. This might decrease the convergence speed. In the presented technique, the number of eggs of every cuckoo is set based on the cost value:

$$\text{eggNum}_i = \text{minEggNum} + \left[ (\text{fitness}_i - \text{fitness}_{\min}) \times \frac{\text{maxEggNum} - \text{minEggNum}}{\text{fitness}_{\max} - \text{fitness}_{\min}} \right], (16)$$

where,  $\text{fitness}_{\min}$  = The minimal fitness in the existing population,  $\text{fitness}_{\max}$  = The maximum fitness in the existing population,  $\text{fitness}_i$  = the fitness value of the existing cuckoo. Result shows that this development increased the convergence speed for this type of problem.

(3) Repair the malformed cuckoo: in the original CS, few components of the habitation vector of cuckoo might surpass the search range of the problem. In such cases, these elements are returned to the search range of the problem. For instance, when the search range is  $[-M, M]$ , a component with value of  $M + \alpha$  would be  $M - \alpha$  after the repair step.

(4) Apply a local search for better solution: In all the iterations, to enhance the better solution, we applied the AE algorithm on the better cuckoo of all the clusters. Once the better cuckoo at every cluster gets enhanced, remaining cuckoos are affected by these improvements because of the Migration step.

The ICS algorithm derives a fitness function to accomplish better performance of the classification. It describes a positive integer to symbolize the superior performance of the candidate solution. The reduction of classification error rate can be regarded as the fitness function.

$$\text{fitness}(x_i) = \frac{\text{number of misclassified samples}}{\text{Total number of samples}} \times 100 \quad (17)$$

#### Algorithm 1: Pseudocode of the ICS

- (1) Begin.
- (2) Initialization
- (3) Evaluate the cost of every cuckoo.
- (4) Allocate few eggs between  $\text{max}_{\text{num}}$  and  $\text{min}_{\text{num}}$  to all the cuckoos.

- (5) Evaluate the ELR for all the cuckoos and define the location of every egg of cuckoo based on the parent cuckoo.
- (6) If some components of the habituated vector of cuckoo exceeded the predetermined range then repair these components.
- (7) If number of present cuckoos in the population is better than  $\text{max}_{\text{num}}$ , then demise few weak cuckoos.
- (8) Define cuckoo society by k-means clustering and find the better cuckoo in all the clusters.
- (9) Employ a local search algorithm on the better cuckoo.
- (10) Move every cuckoo of every cluster towards the better cuckoo in that cluster.
- (11) Define egg laying radius for all the cuckoos.
- (12) When the ending condition is not fulfilled then go to 3
- (13) The better cuckoo is the better solution.
- (14) End.

## 4. RESULTS AND DISCUSSION

The proposed model is simulated using Python 3.6.5 tool on PC i5-8600k, GeForce 1050Ti 4GB, 16GB RAM, 250GB SSD, and 1TB HDD. The parameter settings are given as follows: learning rate: 0.01, dropout: 0.5, batch size: 5, epoch count: 50, and activation: ReLU.

In this study, the bone cancer classification outcomes of the GSODL-CADBCC model are tested using a bone cancer X-ray dataset. The dataset holds 200 images with two classes, as provided in Table 1. The publicly available data sets for research on the bone X-ray image are collected from different sources such as the Indian Institute of Engineering Science and Technology, Shibpur (IEST) and The TCIA (Cancer Imaging Archive). Few sample images are shown in Fig. 3.



Figure 3. Sample X-ray images.

Table 1. Dataset used.

Class	No. of Images
Cancerous bone	100
Healthy bone	100
Total No. of Images	200

The confusion matrices generated by the GSODL-CADBCC model on bone cancer classification process with distinct sizes of training set (TRS) and testing set (TSS) are given in Fig. 4. The results illustrated that the GSODL-CADBCC model has accurately categorized the cancerous and healthy bone images.

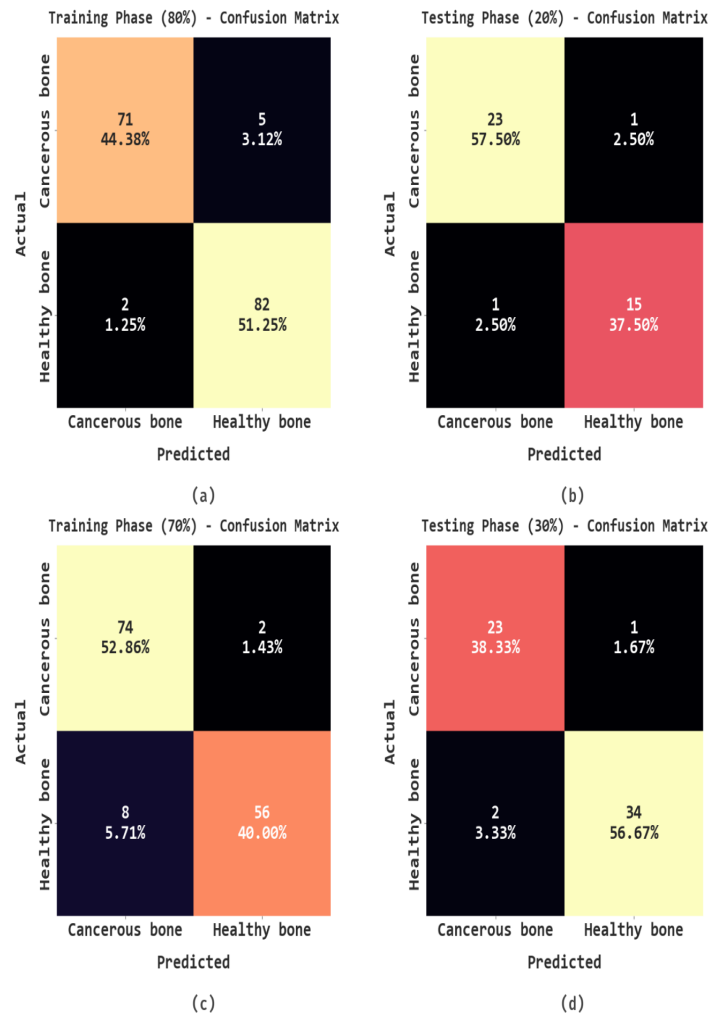


Figure 4. Confusion matrices of GSODL-CADBCC model (a) 80% of TRS, (b) 20% of TSS, (c) 70% of TRS, (d) 30% of TRSS.

In Table 2 and Fig. 5, the bone cancer classifier results of the GSODL-CADBCC model on 80:20 of TRS and TSS are exhibited. It is noticed that the GSODL-CADBCC model has identified the cancerous and healthy bone images effectively. For instance, on 80% of TRS, the GSODL-CADBCC model has reached average  $accu_{bal}$  of 95.52%,  $prec_n$  of 95.76%,  $reca_l$  of 95.52%,  $F_{score}$  of 95.60%,  $AUC_{score}$  of 95.52%, and MCC of 91.28%. On the other hand, on 20% of TSS, the GSODL-CADBCC model has resulted in average  $accu_{bal}$  of 94.79%,  $prec_n$  of 94.79%,  $reca_l$  of 94.79%,  $F_{score}$  of 94.79%,  $AUC_{score}$  of 94.79%, and MCC of 89.58%.

Table 2. Bone cancer classification results of GSODL-CADBCC model on 80:20 of TRS/TSS.

Class Labels	$Accu_{bal}$	$prec_n$	$reca_l$	$F_{score}$	$AUC_{score}$	$MC_C$
Training Phase (80%)						

Cancerous bone	93.42	97.26	93.42	95.30	95.52	91.28
Healthy bone	97.62	94.25	97.62	95.91	95.52	91.28
Average	95.52	95.76	95.52	95.60	95.52	91.28
Testing Phase (20%)						
Cancerous bone	95.83	95.83	95.83	95.83	94.79	89.58
Healthy bone	93.75	93.75	93.75	93.75	94.79	89.58
Average	94.79	94.79	94.79	94.79	94.79	89.58

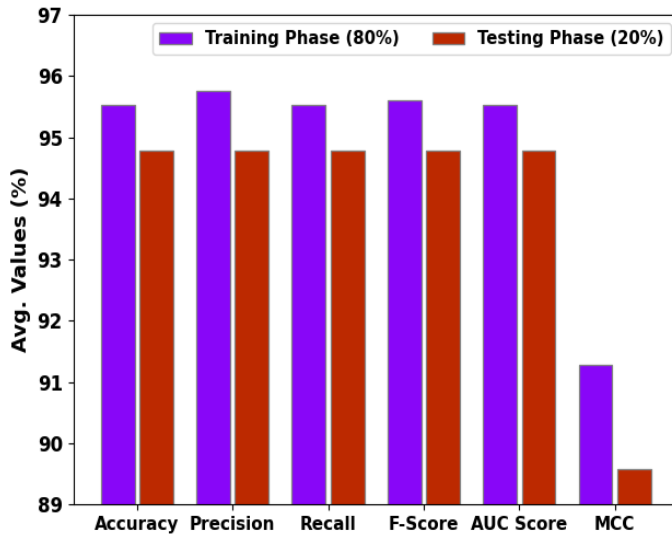


Figure 5. Average results of GSODL-CADBCC model on 80:20 of TRS/TSS.

In Table 3 and Fig. 6, the bone cancer classifier results of the GSODL-CADBCC model on 70:30 of TRS and TSS are exhibited. It is noticed that the GSODL-CADBCC model has identified the cancerous and healthy bone images effectually. For instance, on 70% of TRS, the GSODL-CADBCC model has reached average  $accu_{bal}$  of 92.43%,  $prec_n$  of 93.40%,  $reca_1$  of 92.43%,  $F_{score}$  of 92.74%,  $AUC_{score}$  of 92.43%, and MCC of 85.83%. On the other hand, on 30% of TSS, the GSODL-CADBCC model has resulted in average  $accu_{bal}$  of 95.14%,  $prec_n$  of 94.57%,  $reca_1$  of 95.14%,  $F_{score}$  of 94.83%,  $AUC_{score}$  of 95.14%, and MCC of 89.71%.

Table 3. Bone cancer classification results of GSODL-CADBCC model on 70:30 of TRS/TSS.

Class Labels	$Accu_{bal}$	$prec_n$	$reca_1$	$F_{score}$	MCC
Cancerous bone	97.37	90.24	97.37	93.67	85.83
Healthy bone	87.50	96.55	87.50	91.80	85.83
Average	92.43	93.40	92.43	92.74	85.83
Cancerous bone	95.83	92.00	95.83	93.88	89.71
Healthy bone	94.44	97.14	94.44	95.77	89.71
Average	95.14	94.57	95.14	94.83	89.71

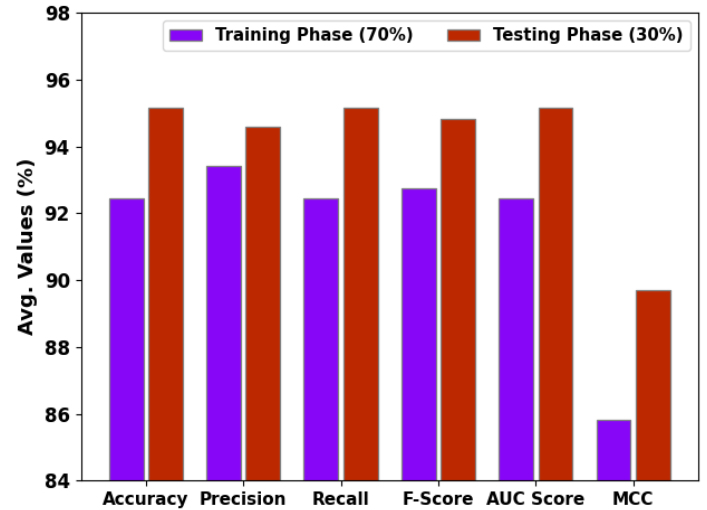


Figure 6. Average results of GSODL-CADBCC model on 70:30 of TRS/TSS.

The training accuracy (TAC) and validation accuracy (VAC) study of the GSODL-CADBCC technique is depicted in Fig. 7. It can be easily noticeable that the GSODL-CADBCC model accomplishes improved values of TAC and VAC. It is also observed that the TAC and VAC reach maximum values at 200 epochs.



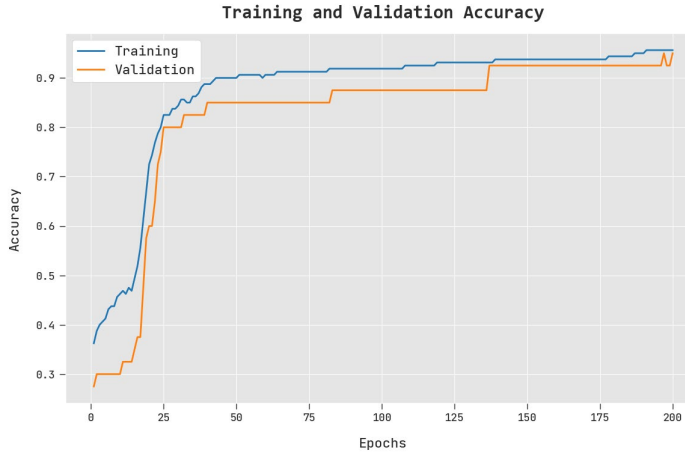
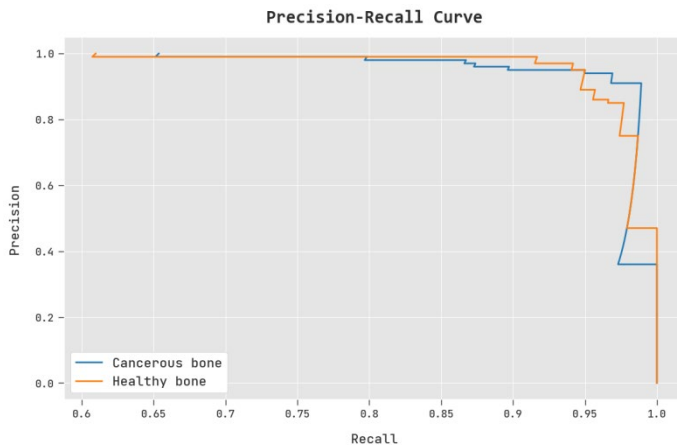


Figure 7. TAC and VAC study of GSODL-CADBCC model.



Figure 8. TLS and VLS study of GSODL-CADBCC model.

The training loss (TLS) and validation loss (VLS) study of the GSODL-CADBCC model is exhibited in Fig. 8. It is noticed that the GSODL-CADBCC model gains reduced TLS and VLS values. In addition, it is assured that the TLS and VLS reach least values at 200 epochs.

Figure 9.  $\text{Prec}_n - \text{reca}_1$  study of GSODL-CADBCC model.

In Fig. 9, the  $\text{prec}_n - \text{reca}_1$  study of the GSODL-CADBCC model on bone cancer classification is well studied. By looking at the results, it is guaranteed that the GSODL-CADBCC model reaches improved  $\text{prec}_n - \text{reca}_1$  values under both cancerous and healthy bone classes.

The receiver operating characteristic (ROC) curve of the GSODL-CADBCC approach is given in Fig. 10. ROC is a graph displaying the performance of a classification model at every classification threshold. The results demonstrated that the GSODL-CADBCC model has been found to be proficient with increased ROC values under both cancerous and healthy bone classes.

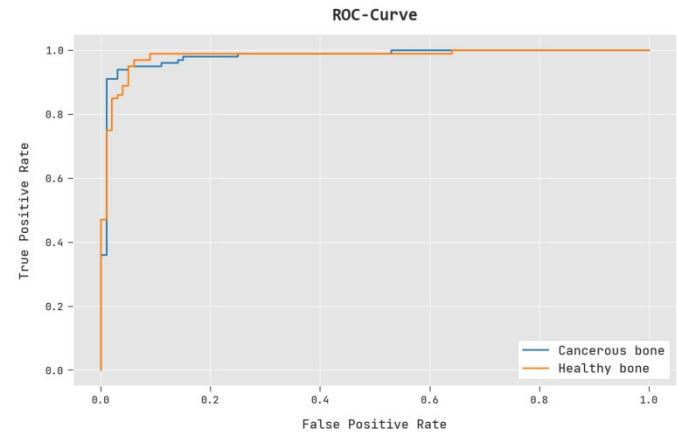


Figure 10. ROC study of GSODL-CADBCC model.

To exhibit the betterment of the GSODL-CADBCC model, a widespread comparison study with recent models [27, 28] is given in Table 4. The result indicates that RF model has shown poor outcomes compared to other models. Next, the learning-based intelligent optimization (LIO) model has shown somewhat improvised performance with  $\text{accu}_y$  of 85.37%. Followed by, the SVM, FE-ML, and MRI-DL models have accomplished moderately closer  $\text{accu}_y$  values of 93.59%, 92.81%, and 95.04% respectively. But the GSODL-CADBCC model showed promising results with maximum  $\text{accu}_y$  of 95.52%.

Table 4. Comparative bone cancer classification results with recent models [27, 28].

Measure	Accu-racy	Preci-sion	Re-call	F1 score
GSODL-CADBCC	95.52	95.76	95.52	95.60
RF Algorithm	70.39	76.95	80.06	78.55
SVM Model	93.59	91.45	96.13	94.08
LIO Algorithm	85.37	87.84	81.43	89.06

FE-ML Model	92.81	94.27	90.31	95.03
MRI-DL Algorithm	95.04	93.72	94.16	94.85

A comparative computation time (CT) inspection of the GSODL-CADBCC model is provided in Table 5 and Fig. 11. The experimental values represented that FE-ML model has shown poor outcomes compared to other models with CT of 231s. Next, the MRI-DL and LIO models have shown somewhat improvised performance with  $\text{acc}_y$  of 226s and 204s respectively. Followed by, the RF and SVM models have accomplished moderately closer reasonable CT of 112s and 156s respectively. But the GSODL-CADBCC model showed promising results with minimal CT of 54s.

Table 5. Comparative CT analysis on bone cancer classification.

Methods	Computational time (sec)
GSODL-CADBCC	54
RF Algorithm	112
SVM Model	156
LIO Algorithm	204
FE-ML Model	231
MRI-DL Algorithm	226

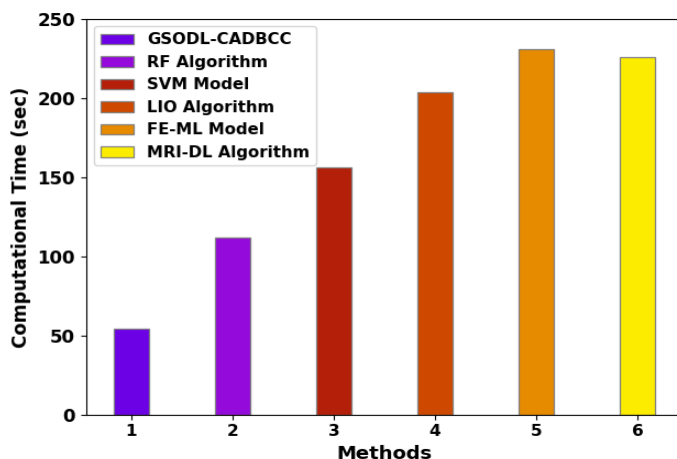


Figure 11. Comparative CT assessment of GSODL-CADBCC model.

These results demonstrated that the GSODL-CADBCC model can attain effective bone cancer classification performance. The enhanced outcomes of the GSODL-CADBCC technique is due to the application of GSO and ICS algorithms for the parameter selection of the SqueezeNet and LSTM models respectively.

## 5. CONCLUSION

In this study, we have introduced an effective GSODL-CADBCC technique for automated bone cancer classification

on X-ray images. The presented GSODL-CADBCC technique encompasses BF based noise removal, GSO with SqueezeNet based feature extraction, LSTM classification, and ICS based hyperparameter tuning. The design of GSO and ICS algorithms helps in optimal selection of the hyperparameters of the SqueezeNet and LSTM models. A wide range of experiments was performed to study the performance of the GSODL-CADBCC technique on medical imaging datasets. Compared with recent DL models, the experimental outcomes demonstrated that the GSODL-CADBCC technique achieves promising performance on bone cancer classification. Therefore, the GSODL-CADBCC technique can be exploited for automated and accurate bone cancer classification. In future, deep instance segmentation techniques can be included in the GSODL-CADBCC technique for improving its classification efficacy.

**Funding details:** This study was supported by a Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT & Future Planning NRF-2020R1A2C1014829 and the Soon-chunhyang University Research Fund.

**Acknowledgments:** The authors would like to acknowledge the support provided by AlMaarefa University while conducting this research work.

**Conflicts of Interest:** The authors declare that they have no conflict of interest. The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript.

**Data Availability Statement:** Data sharing not applicable to this article as no datasets were generated during the current study.

**Ethics approval:** This article does not contain any studies with human participants performed by any of the authors.

**Consent to Participate:** Not applicable.

**Informed Consent:** Not applicable.

## REFERENCES

- [1] He, Y.; Pan, I.; Bao, B.; Halsey, K.; Chang, M.; Liu, H.; Peng, S.; Sebro, R.A.; Guan, J.; Yi, T.; et al. Deep learning-based classification of primary bone tumors on radiographs: A preliminary study. *EBioMedicine* 2020, 62, 103121.
- [2] Eweje, F.R.; Bao, B.; Wu, J.; Dalal, D.; Liao, W.-H.; He, Y.; Luo, Y.; Lu, S.; Zhang, P.; Peng, X.; et al. Deep Learning for Classification of Bone Lesions on Routine MRI. *EBioMedicine* 2021, 68, 103402.
- [3] Saraiva, M.M.; Ribeiro, T.; Afonso, J.; Andrade, P.; Cardoso, P.; Ferreira, J.; Cardoso, H.; Macedo, G. Deep Learning and Device-Assisted Enteroscopy: Automatic Detection of Gastrointestinal Angiectasia. *Medicina* 2021, 57, 1378.
- [4] Gitto, S.; Cuocolo, R.; van Langevelde, K.; van de Sande, M.A.; Parafioriti, A.; Luzzati, A.; Imbriaco, M.; Sconfienza, L.M.; Bloem, J.L. MRI radiomics-based machine learning classification of atypical cartilaginous tumour and grade II chondrosarcoma of long bones. *EBioMedicine* 2022, 75, 103757.

- [5] O. Bandyopadhyay, A. Biswas, and B. B. Bhattacharya, "Bonecancer assessment and destruction pattern analysis in longbone X-ray image," *Journal of Digital Imaging*, vol. 32, no. 2, pp. 300–313, 2019.
- [6] D. Shrivastava, S. Sanyal, A. K. Maji, and D. Kandar, "Bone cancer detection using machine learning techniques," in *Smart Healthcare for Disease Diagnosis and Prevention*, vol. 20, pp. 175–183, Academic Press, 2020.
- [7] B. S. Vandana, P. J. Antony, and R. A. Sathyavathi, "Analysis of malignancy using enhanced graphcut-based clustering for diagnosis of bone cancer," in *Information and Communication Technology for Sustainable Development*, pp. 453–462, Springer, 2020.
- [8] A. Torki, "Fuzzy rank correlation-based segmentation method and deep neural network for bone cancer identification," *Neural Computing and Applications*, vol. 32, no. 3, pp. 805–815, 2020.
- [9] von Schacky, C.E., Wilhelm, N.J., Schäfer, V.S., Leonhardt, Y., Gassert, F.G., Foreman, S.C., Gassert, F.T., Jung, M., Jungmann, P.M., Russe, M.F. and Mogler, C., 2021. Multitask deep learning for segmentation and classification of primary bone tumors on radiographs. *Radiology*, 301(2), pp.398-406.
- [10] W. Li, G. G. Wang, and A. H. Gandomi, "A survey of learningbased intelligent optimization algorithms," *Archives of Computational Methods in Engineering*, vol. 28, no. 5, pp. 3781– 3799, 2021
- [11] Georgeanu, V.A., Mămuleanu, M., Ghiea, S. and Selișteanu, D., 2022. Malignant Bone Tumors Diagnosis Using Magnetic Resonance Imaging Based on Deep Learning Algorithms. *Medicina*, 58(5), p.636.
- [12] Liu, R., Pan, D., Xu, Y., Zeng, H., He, Z., Lin, J., Zeng, W., Wu, Z., Luo, Z., Qin, G. and Chen, W., 2022. A deep learning–machine learning fusion approach for the classification of benign, malignant, and intermediate bone tumors. *European Radiology*, 32(2), pp.1371-1383.
- [13] Wang, D., Hwang, M., Jiang, W.C., Ding, K., Chang, H.C. and Hwang, K.S., 2021. A deep learning method for counting white blood cells in bone marrow images. *BMC bioinformatics*, 22(5), pp.1-14.
- [14] Calin, M.A., Elfarra, F.G. and Parasca, S.V., 2021. Object-oriented classification approach for bone metastasis mapping from whole-body bone scintigraphy. *Physica Medica*, 84, pp.141-148.
- [15] Georgeanu, V., Mamuleanu, M.L. and Selișteanu, D., 2021, June. Convolutional neural networks for automated detection and classification of bone tumors in magnetic resonance imaging. In *2021 IEEE International Conference on Artificial Intelligence, Robotics, and Communication (ICAIRC)* (pp. 5-7). IEEE.
- [16] Han, S., Oh, J.S. and Lee, J.J., 2022. Diagnostic performance of deep learning models for detecting bone metastasis on whole-body bone scan in prostate cancer. *European Journal of Nuclear Medicine and Molecular Imaging*, 49(2), pp.585-595.
- [17] Eweje, F.R., Bao, B., Wu, J., Dalal, D., Liao, W.H., He, Y., Luo, Y., Lu, S., Zhang, P., Peng, X. and Sebro, R., 2021. Deep learning for classification of bone lesions on routine MRI. *EBioMedicine*, 68, p.103402.
- [18] Sushmitha, K. and Jagadeesh, P., 2022, February. Feature Extraction and Classification of Bone Tumor using CNN Classifier with KNN Classifier. In *2022 International Conference on Business Analytics for Technology and Security (ICBATS)* (pp. 1-5). IEEE.
- [19] Satheesh Kumar, B. and Bb, S., 2021. Bone Cancer Detection Using Feature Extraction with Classification Using K-Nearest Neighbor and Decision Tree Algorithm.
- [20] Rajagopal, S., Kanimozhi, S., Chakrabarti, A. and Velez, D.G., 2021. Convolution Neural Network Based Bone Cancer Detection. *SPAST Abstracts*, 1(01).
- [21] Ranjitha, M.M., Taranath, N.L., Arpitha, C.N. and Subbaraya, C.K., 2019, July. Bone cancer detection using K-means segmentation and Knn classification. In *2019 1st International Conference on Advances in Information Technology (ICAIT)* (pp. 76-80). IEEE.
- [22] Naveen, P. and Sivakumar, P., 2021. Adaptive morphological and bilateral filtering with ensemble convolutional neural network for pose-invariant face recognition. *Journal of Ambient Intelligence and Humanized Computing*, 12(11), pp.10023-10033.
- [23] Naveen, P. and Sivakumar, P., 2021. Adaptive morphological and bilateral filtering with ensemble convolutional neural network for pose-invariant face recognition. *Journal of Ambient Intelligence and Humanized Computing*, 12(11), pp.10023-10033.
- [24] Noroozi, M., Mohammadi, H., Efatinasab, E., Lashgari, A., Eslami, M. and Khan, B., 2022. Golden Search Optimization Algorithm. *IEEE Access*, 10, pp.37515-37532.
- [25] Belagoune, S., Bali, N., Bakdi, A., Baadji, B. and Atif, K., 2021. Deep learning through LSTM classification and regression for transmission line fault detection, diagnosis and location in large-scale multi-machine power systems. *Measurement*, 177, p.109330.
- [26] Dana Mazraeh, H., Kalantari, M., Tabasi, S.H., Afzal Aghaei, A., Kalantari, Z. and Fahimi, F., 2022. Solving Fredholm Integral Equations of the Second Kind Using an Improved Cuckoo Optimization Algorithm. *Global Analysis and Discrete Mathematics*.
- [27] Georgeanu, V.A.; Mămuleanu, M.; Ghiea, S.; Selișteanu, D. Malignant Bone Tumors Diagnosis Using Magnetic Resonance Imaging Based on Deep Learning Algorithms. *Medicina* 2022, 58, 636. <https://doi.org/10.3390/medicina58050636>
- [28] Sharma, A., Yadav, D.P., Garg, H., Kumar, M., Sharma, B. and Koundal, D., 2021. Bone cancer detection using feature extraction based machine learning model. *Computational and Mathematical Methods in Medicine*, 2021.

# Home Mobility With LLM

Jin Woong Lee<sup>1</sup>, Dong Hee Seo<sup>1</sup>, Hyuk Mo An<sup>1</sup>, and Seok Young Lee<sup>1,\*</sup>

<sup>1</sup>Department of ICT Convergence Engineering, Soonchunhyang University, Asan, South Korea

\*Contact: suk122@sch.ac.kr

**Abstract**— This paper explores integrating large language models (LLMs) with home mobility robots to facilitate user-friendly control through natural language commands. The project aims to convert natural language instructions into executable robot commands using ChatGPT-4o and ROS2. We developed and tested a system that translates user commands into ROS2 code, automatically compiles and uploads this code to a WeGO LIMO robot, and executes the instructions. Our results demonstrate improved accessibility and efficiency in robot control, highlighting the potential for broader applications in home mobility and other fields.

## I. INTRODUCTION

Rapid advances in natural language processing models in artificial intelligence have led to the development of large language models (LLMs) that are revolutionizing a wide range of applications. LLMs like GPT-4o, LLaMA, BARD, and others have shown remarkable results in various tasks, including translation, text generation, image creation, code modification, and more. These models excel in text generation, machine translation, and code synthesis tasks, prompting exploration of their potential in robotics. Robotics, however, presents unique challenges, requiring an understanding of real-world physics, context, and the ability to perform physical actions based on textual commands.

The core objective of this project is to integrate ChatGPT-4o, a state-of-the-art conversational AI, with a home mobility robot to convert natural language commands into executable robot instructions. Specifically, we aim to use ChatGPT-4o with the Robot Operating System 2 (ROS2) to create a system that allows users to control a WeGO LIMO robot through simple, intuitive natural language commands. Traditional robot control systems often require extensive programming knowledge, creating a barrier for general users. By leveraging ChatGPT-4o's natural language understanding capabilities, we seek to eliminate this barrier, making robot control accessible to a wider audience. This project explores how natural language commands can be effectively translated into ROS2 commands, compiled, and executed by the WeGO LIMO robot, ensuring both accuracy and efficiency in task execution.

In this paper, we present the methodology for integrating ChatGPT-4o with ROS2, detail the experimental setup and results, and discuss the implications of our findings for the future of user-friendly robotics. Our approach aims to enhance the accessibility and usability of home mobility robots, potentially transforming their role in everyday life.

## II. METHODOLOGY

Our project involves several key steps to integrate ChatGPT-4o with a home mobility robot using ROS2. The methodology is designed to systematically convert natural language commands into executable robot instructions, ensuring both accuracy and efficiency in task execution. The following subsections outline the main components of our methodology:

### A. Natural Language Processing

The first step involves utilizing ChatGPT-4o to interpret user commands. ChatGPT-4o is fine-tuned to understand and extract the intent behind natural language instructions. This involves training the model to recognize various commands related to home mobility tasks, such as navigation, object manipulation, and status reporting.

- **Input:** User inputs a natural language command.
- **Command Interpretation:** ChatGPT-4o understands the natural language command with its pre-training information.
- **Output:** ChatGPT-4o writes a Python code for the node that perform the natural language command.

### B. Code Scraping

Once the natural language command is interpreted, the next step is to translate it into ROS2 code. This involves developing algorithms that map the structured command representation to specific ROS2 commands.

- **Clipping:** The API fetches the output from ChatGPT-4o.
- **Extraction:** The API extracts the Python code from the clipped output.
- **Storage:** The extracted Python code is stored by the API



Fig. 1 WeGO LIMO robot

### C. Automated Packaging

The translated ROS2 code needs to be compiled and uploaded to the WeGo LIMO robot. This step focuses on automating the deployment process to minimize user intervention and ensure smooth operation.

- **Node:** The stored Python code is made into a node in ROS2 by the API.
- **Packaging:** The node are packaged to meet the required dependencies.
- **Build:** The generated package is built.

### D. Execution

The WeGo LIMO robot executes the uploaded commands, and its performance is monitored in real-time. This step involves collecting data on the robot's actions and providing feedback to refine the system.

- **Command Execution:** The built package is executed automatically by the API.
- **Next Command:** The API prepares to receive the next command.

## III. EXPERIMENT AND RESULT

The experiments were conducted to evaluate the effectiveness of our system in interpreting and executing natural language commands using ChatGPT-4 integrated with ROS2 on a WeGo LIMO robot. The following sections outline the experimental setup, results, and analysis.

### A. Experiment setup

The experiments were conducted in a controlled indoor environment. We utilized ROS2 for its robust networking capabilities and enhanced real-time performance. The WeGO LIMO robot was used as the primary robotic platform for executing commands. For our experiment, we train ChatGPT-4 with WeGO LIMO and ROS2, as shown in Fig. 2.

### B. Experiment result

A command is given to ChatGPT-4o in natural language, as shown in Fig. 3, and ChatGPT-4o writes the main Python code to perform that command.

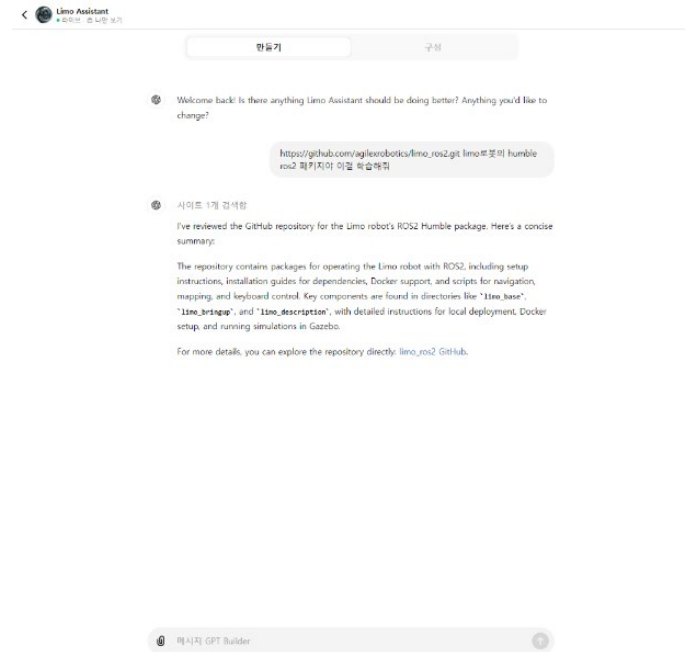


Fig. 2 ChatGPT-4o learning WeGo LIMO and ROS2



Fig. 3 ChatGPT-4o writing code by natural language command

The written code is packaged by the API developed in this paper, as shown in Fig. 4, and is automatically built and executed.



Through the aforementioned process, we have created and

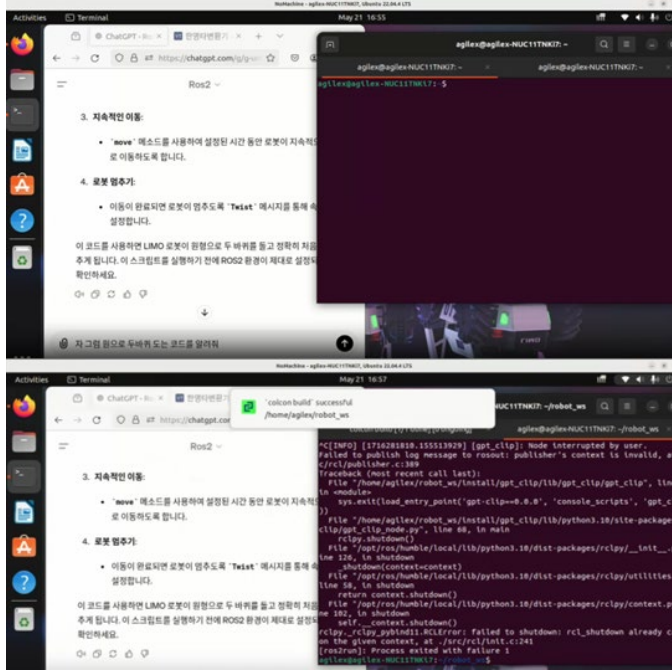


Fig. 4 API



Fig. 5 LIMO performing the command

executed a package that performs the command and verified that WeGO limo performs the command.

Fig. 5 shows the package created through the API developed in this paper running and WeGO RIMO moving.

#### IV. CONCLUSION

The experimental results validate the effectiveness of integrating ChatGPT-4 with ROS2 for controlling home mobility robots. By leveraging ChatGPT-4o's advanced natural language processing capabilities, we have developed a system that translates user-friendly natural language commands into executable ROS2 code, facilitating seamless robot control. The integration with the WeGO LIMO robot demonstrated high accuracy and efficiency in command execution. Our experiments showed that the system accurately performed both simple and complex commands, proving the effectiveness of the command interpretation and translation algorithms. These results indicate that further improvements can enhance the

system's performance even more.

This approach significantly enhances the accessibility and functionality of home mobility robots, making sophisticated robotic control systems accessible to users without extensive programming knowledge. Future work will focus on refining command parsing algorithms to improve the accuracy of complex command execution, optimizing response times for real-time interactions, and exploring broader applications and potential enhancements in other domains. These advancements demonstrate the practical applicability of combining LLMs like ChatGPT-4o with robotics, paving the way for more intuitive and effective human-robot interactions.

#### ACKNOWLEDGMENT

This research was supported by the MSIT (Ministry of Science and ICT), Korea, under the ICAN (ICT Challenge and Advanced Network of HRD) support program (IITP-2024-2020-0-01832), supervised by the IITP (Institute for Information & Communications Technology Planning & Evaluation).

#### REFERENCES

- [1] T. B. Brown et al., "Language Models are Few-Shot Learners," in *Advances in Neural Information Processing Systems*, vol. 33, pp. 1877-1901, 2020.
- [2] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proc. 2019 Conf. North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pp. 4171-4186, 2019.
- [3] A. Vaswani et al., "Attention is All You Need," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, pp. 5998-6008, 2017.
- [4] R. Bommasani et al., "On the Opportunities and Risks of Foundation Models," in *IEEE Trans. Neural Networks and Learning Systems*, vol. 32, no. 2, pp. 1111-1123, 2021.
- [5] R. Thoppilan et al., "LaMDA: Language Models for Dialog Applications," in *Proc. 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 2327-2340, 2022.
- [6] Y. Zhang et al., "Dialogpt: Large-Scale Generative Pre-Training for Conversational Response Generation," in *Proc. 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 270-281, 2019.
- [7] H. Zhang, Y. Zhang, and M. Sun, "Semantics-Aware BERT for Language Understanding," in *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 42, no. 3, pp. 684-698, 2020.
- [8] C. Raffel et al., "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer," in *J. Machine Learning Research*, vol. 21, pp. 1-67, 2020.
- [9] P. Henderson et al., "Ethical Challenges in Data-Driven Dialogue Systems," in *Proc. IEEE Int. Conf. Robotics and Automation (ICRA)*, pp. 4118-4125, 2017.
- [10] Y. Liu et al., "RoBERTa: A Robustly Optimized BERT Pretraining Approach," in *Proc. 2019 Conf. Empirical Methods in Natural Language Processing and the 9th Int. Joint Conf. Natural Language Processing (EMNLP-IJCNLP)*, pp. 2452-2461, 2019.
- [11] Microsoft, "Using GPT-3 for Robotics: Design Principles and Model Capabilities," in *IEEE Robotics and Automation Letters*, vol. 7, no. 3, pp. 712-719, 2022.



# PUSCH DM-RS Pattern Optimization in 5G

Daegun Jang<sup>1</sup>, Gayeon Kim<sup>2</sup>, and Byeong-Gwon Kang<sup>\*</sup>.

<sup>1,2</sup> *ICT Convergence, Soonchunhyang University, Asan, Korea*

<sup>\*</sup> *Information and Communication Engineering, Soonchunhyang University, Asan, Korea*

<sup>\*</sup>Contact: First.wowhensum@sch.ac.kr, phone +82 10-7210-4419

**Abstract**—Channel estimation is an essential technology used in all wireless communication systems to improve link performance. DM-RS-based channel estimation, which has been adopted as a standard technology because it can effectively respond to channel changes, requires efficient operation of resources due to the characteristics of DM-RS that occupies time-frequency resources. In this paper, we propose a ResNet-based PUSCH DM-RS pattern optimization technique and verify its performance by achieving 94.37% accuracy and a data rate of more than 10% of the learned network.

## I. INTRODUCTION

demodulation Reference Signal (DM-RS), used for channel estimation, is transmitted along with the data information to be demodulated and allocated to time-frequency resources along with the data information. This enables effective response to real-time channel changes, leading to the adoption of DM-RS-based channel estimation as a standard technology [1]. To address real-time channel variations, various DM-RS patterns are defined. However, due to the characteristic of DM-RS occupying time-frequency resources along with data information, it is necessary to select the optimal DM-RS pattern for transmission.

In this paper, we propose a residual neural network (ResNet)-based physical uplink shared channel (PUSCH) pattern optimization technique for efficient DM-RS utilization in 5G NR systems. The proposed technique pre-investigates the pattern that achieves the highest data transmission rate for a specific wireless fading channel based on the DM-RS patterns suggested by 3GPP. This pattern is used as the correct value for ResNet learning, which divides the received signal elements into real and imaginary groups. The performance of the trained network is then evaluated. The proposed technique verifies that it is possible to minimize the overhead of time-frequency resources and efficiently utilize these resources by optimizing PUSCH DM-RS patterns compared to standard technologies.

## II. SYSTEM MODEL

### A. NR Frame Structure

In the 5G NR system, various types of numerology are supported, and based on this, various slot lengths within a 1ms subframe on the time domain and subcarrier intervals on the frequency domain are supported. There are always 14 OFDM symbols in a single slot, and 12 consecutive subcarriers constitute a physical resource block (PRB), and a plurality of

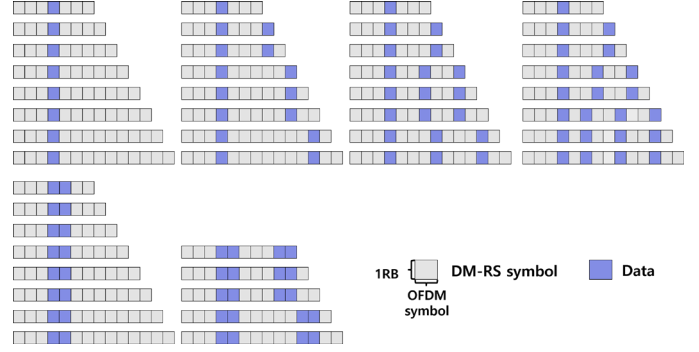


Fig. 1 Example of PUSCH DM-RS

consecutive PRBs may constitute a sub-channel. A plurality of slots on the time domain and a plurality of sub-channels on the frequency domain can form a resource grid (RG). Each time-frequency resource in RG is divided into a resource element (RE). RE consists of one OFDM symbol on the time domain and one subcarrier on the frequency domain [1].

### B. NR PUSCH and DM-RS

In the 5G NR system, PUSCH is used to transmit uplink data payload. PUSCH is assigned to a plurality of OFDM symbols on the time domain and a plurality of sub-channels on the frequency domain and undergoes transport processing as in [xx] before being transmitted. When PUSCH is allocated to RG, the allocation position and length on the time domain are determined based on the PUSCH mapping type. There are two PUSCH mapping types, A and B. They are largely divided into slot-based scheduling and mini slot-based scheduling and are used as parameters to determine the DM-RS pattern [2].

The PUSCH DM-RS, which occupies time-frequency resources alongside the PUSCH, is a reference signal used for decoding the received PUSCH. It is generated as  $r(n) = \frac{1}{\sqrt{2}}(1 - 2c(2n) + j\frac{1}{\sqrt{2}}(1 - 2c(2n+1)))$ , where  $c(n)$  is pseudo-random sequence. The generated DM-RS symbols are allocated to the resource grid (RG) based on the patterns specified by 3GPP [1]. Tables 6.4.1.1.3-3 and 6.4.1.1.3-4 in [1] illustrate the patterns for single symbol and double symbol DM-RS transmission when  $l_0 = 3$ , and PUSCH mapping type A, as shown in Fig 1.

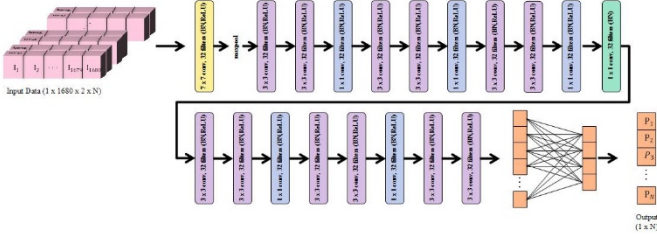


Fig. 2 Proposed ResNet structure.

### III. PROPOSED DM-RS PATTERN OPTIMIZATION METHOD

Each convolutional layer has a size of  $1 \times 1, 3 \times 3$ , and uses 32 filters. BN is batch normalization and normalizes the input to each layer with mean and variance. The rectified linear unit (ReLU) function is used as the activation function, and the maximum pooling layer uses MaxPool. After passing through 18 convolutional layers, the learned PUSCH DM-RS pattern is output through a fully connected layer and a SoftMax layer.

#### A. RESNET-BAESD PUSCH DM-RS PATTERN OPTIMIZATION

The data used for training consists of wireless resources composed of 1 slot in the time domain and 10 RBs in the frequency domain, and the DM-RS patterns considered in this paper are applied. All symbols are QPSK-modulated and pass through a tapped-delay line (TDL) - A channel. It is assumed that all patterns pass through the same TDL-A channel if they have the same delay spread, user equipment (UE) velocity, and SNR. Among all DM-RS patterns that pass through this channel, the pattern that achieves the highest normalized data rate  $R_n$  based on the following equation is used as the correct value [3].

$$R_n = \frac{\alpha - \tau}{\alpha} \times R, \left( R = \log_2 \left( 1 + \rho \frac{|\hat{H}|^2}{|H - \hat{H}|^2} \right) \right), \text{where} \quad (2)$$

$\alpha$  represents the total number of symbols,  $\tau$  represents the number of used DM-RS symbols,  $\rho$  represents the SNR power, and  $R$  represents the data rate before normalization.

In this paper, the mean square error (MSE) is calculated for a total of 300 scenarios to select the optimal pattern among various PUSCH DM-RS patterns, considering delay spread, UE velocity, and SNR.

$$MSE = \frac{1}{n} \sum_{i=1}^n |H_i - \hat{H}_i|^2, \text{where} \quad (3)$$

$H_i$  is  $i$ th estimated channel value and  $\hat{H}_i$  represents predicted value of  $i$ th trained channel.

### IV. SIMULATION RESULT

Fig.3 shows the accuracy and loss rate of the proposed system model. The blue curve represents the accuracy, and the red curve represents the loss rate. The proposed system model has an accuracy of 94.37% and a loss rate of 0%. Fig. 4 shows the data rate comparison between the fixed pattern and the

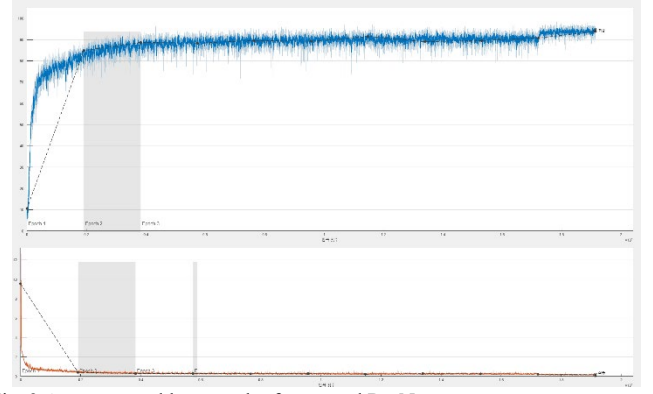


Fig. 3 Accuracy and loss graph of proposed ResNet

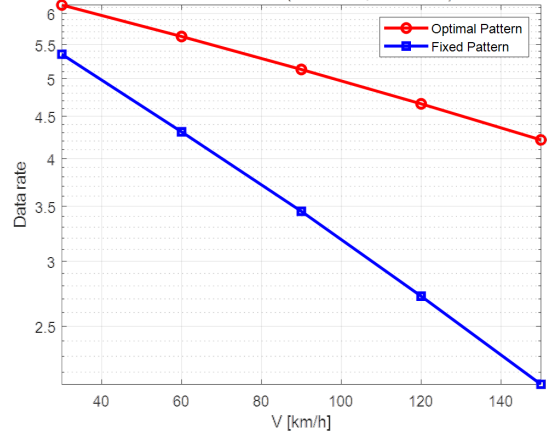


Fig. 4 MSE performance according to the speed of optimal and existing patterns.

optimal pattern selected by the system model, with double symbol transmission and DM-RS configuration type 1. Both patterns show a decrease in the data rate as the UE velocity increases. The optimal pattern outperforms the existing technology by at least 10% and shows an even more significant improvement in data rate as the UE velocity increases.

### V. CONCLUSIONS

In this paper, we proposed a channel estimation technique that selects the optimal DM-RS pattern from various DM-RS patterns considering delay spread, UE velocity, SNR, etc. in 300 scenarios. The proposed ResNet learning model showed an accuracy of 94.37% and achieved a data rate improvement of at least 10% compared to existing channel estimation techniques.

### ACKNOWLEDGMENT

This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the ICAN(ICT Challenge and Advanced Network of HRD) support program(IITP-2024-2020-0-01832) supervised by the IITP(Institute for Information & Communications Technology Planning & Evaluation)

### REFERENCES

- [1] TSG RAN; NR; Physical Channels and Modulation, V17.2.0, Release 17, 3GPP Standard TS 38.211, Jun. 2022.
- [2] TSG RAN; NR; Multiplexing and channel coding (Release 17), document TS 38.212 V17.0.0, 3GPP, Dec. 2021
- [3] TSG RAN; NR; Physical channel and modulation (Release 17), TR 38.901 V17.0.0 3GPP Dec.2021.

# Development of Disorder Early Diagnosis platform using XR headset

Sejin Gown<sup>1\*</sup>, Yunyoung Nam<sup>2</sup>

<sup>1</sup>Department of Computer Science and Engineering, Soonchunhyang University, Asan 31538, Korea

<sup>2</sup>Department of ICT Convergence, Soonchunhyang University, Asan 31538, Korea

\*Contact: [lovecein4858@naver.com](mailto:lovecein4858@naver.com)

**Abstract** – Developmental disorders can be treated more effectively if early intervention is provided at a young age. However, it is difficult for parents to identify normal language and behavior before the child reaches the appropriate age. And not everyone can go to the hospital and get diagnosed by a specialist. For this reason, many children miss the opportunity for treatment until they show obvious symptoms of developmental disorders. This paper discusses meta-bus content for the early diagnosis of developmental disorders. VR devices can collect data such as eye tracking, facial tracking, hand tracking, etc. Content that can elicit children's responses should be created to collect data for early diagnosis in the future. Currently, we have created content such as block stacking, ball throwing, bubble popping, etc. using Unity. We tested this content on actual children. Ball-throwing was found to be inappropriate due to the limitations of VR hand tracking. Bubble popping successfully elicited children's responses and collected the necessary data. Block stacking was not tested on children. We plan to create a deep learning model for early diagnosis of developmental disorders by collecting and analyzing more data in the future. And the current VR content is not optimized. This causes stuttering when running and needs to be resolved.

## I. INTRODUCTION

South Korea is facing challenges in nurturing future talents due to the low birth rate issue [1]. Therefore, it is crucial to identify and foster outstanding talents even amidst the low birth rates. Developmental disorders are more effectively treated when diagnosed early [2]. However, they are often difficult for parents or caregivers to detect, and obtaining accurate diagnoses from hospitals can be challenging. Consequently, many toddlers with developmental disorders miss the appropriate treatment window, as highlighted in the announcement of survey results for people with developmental disabilities in 2021 [3]. To address these issues, this paper proposes a method for early diagnosis of developmental disorders using metaverse content. Metaverse content enables various activities in virtual spaces through VR devices. VR devices are equipped with sensors capable of detecting eye tracking, facial tracking, hand tracking, etc., allowing for the tracking of user expressions, eye movements, hand gestures, etc., which can be stored as data. To utilize this data effectively, content that can elicit responses from toddlers is necessary. In this study, VR content such as stacking blocks, throwing balls, blowing bubbles, and breaking dishes was developed using the Unity engine. Additionally, data storage and visualization modules were developed to store information on gaze position, hand position, head position, and the

positions of objects. Furthermore, the three-dimensional appearance of each object was projected into two dimensions for visualization. Among the developed content, stacking blocks and throwing balls serve as assessments for toddlers' sociability, evaluating their ability to understand and follow rules. Popping bubbles is intended to increase toddlers' interest. However, in tests conducted with actual toddlers, throwing balls failed to sufficiently capture their interest and did not function properly due to the limited hand tracking range of the VR device. Stacking blocks have not yet been tested with toddlers, so validation is necessary. Popping bubbles effectively engaged toddlers' interest and successfully recorded data. In the future, we plan to analyze the extracted data to develop a deep learning model capable of suggesting the types of developmental disorders and treatment methods. Additionally, we will enhance the visual and performance aspects of VR content through optimization.

## II. SYSTEM IMPLEMENTATION

### A. Data extraction

We developed a data extraction module in Unity to extract data for analysis. This module records the 3D positions of eyes, heads, hands, and objects and converts them into 2D format for storage as video. Additionally, the module stores the 3D data to allow for analysis from various angles. Fig. 1 illustrates the transformation of 3D objects into 2D format.

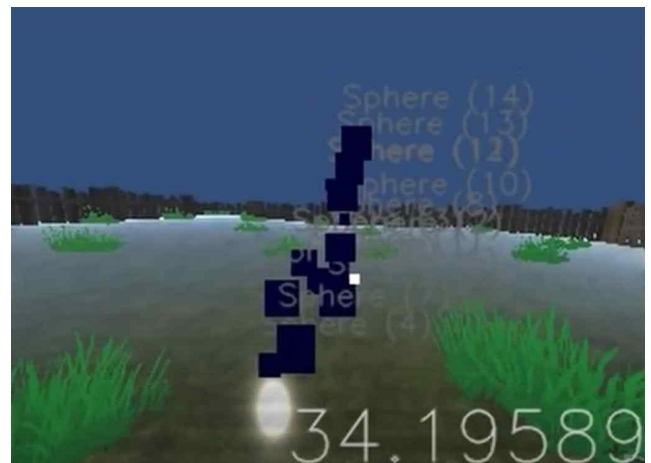


Fig.1 Visualization

Table 2 contains the contents of the data stored for data analysis.

TABLE.2 Eye Concentration

Data type	Data content
Timer	Timer that starts after project starts
eye_pnt	Where the eyes are looking
gaz_obj	Name of the object you are looking at
eye_cls	0 if eyes are open, 1 if eyes are closed
obj_pst	Coordinate values for drawing the object's bounding box
head_roll	Radius of rotation around the X axis of the head
head_pitch	Radius of rotation around the Y axis of the head
head_yaw	Radius of rotation around the Z axis of the head
hand_pst	Position of hand on screen
hand_hld	Name of object held by each hand
hand_gest	A value that indicates what gesture you are using when holding an object.

Three contents have been developed in Unity: block stacking, ball throwing, and bubble popping. These serve as social interaction tests, with bubble popping focusing on engaging the child's interest. Fig.3 depicts the block stacking content utilizing hand tracking, where players stack blocks upwards. To facilitate smooth gameplay, markers indicating the intended block movement upon releasing the hand have been incorporated.



Fig.3 Block stacking

Fig.3 Block stacking

Fig. 4 showcases the ball-throwing content, utilizing hand tracking. This content involves throwing balls to score as high as possible on a target board. To address issues with hand tracking recognition extending beyond its detection range, adjustments have been made so that when the hand re-enters the detection range from outside, the ball will accurately target the board.



Fig.4 Ball throwing

Fig. 5 depicts the Bubble Pop content, utilizing hand tracking. This content involves popping bubbles generated by hand movements using fingers.

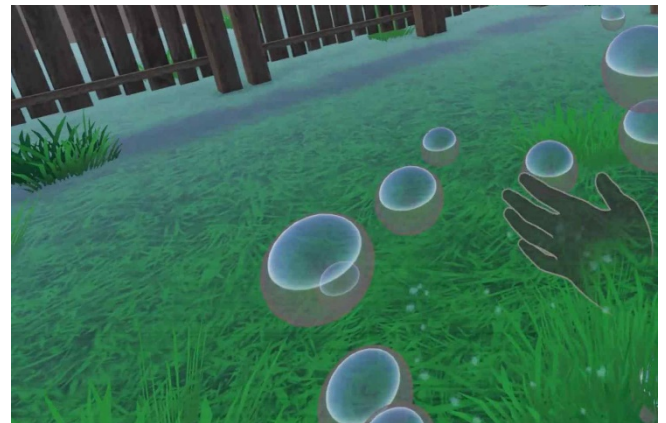


Fig.5 Popping bubbles



### III. EXPERIMENTAL RESULTS

#### A. Check eye concentration

Upon reviewing the data, it was found that there was a data error in the analysis of concentration levels during ball throwing, preventing further analysis. However, it was observed that the child's concentration appeared to be lower, as they showed more interest in surrounding objects than in the ball itself. Referring to Table 2, when examining concentration based on whether the gaze was directed towards the object or not during bubble blowing, it was observed that

for 3-year-old children, the concentration level was 75.1%, and for 7-year-old children, it was 65.4%.

Age	Eye concentration
3	75.1%
7	65.4%

TABLE.2 Eye Concentration

### IV. CONCLUSIONS

This paper focuses on developing content for the early diagnosis of developmental disorders. Three social interaction contents were created using Unity, aiming to collect data from children's responses to enable diagnosis and treatment. In the bubble popping content, when tested on a non-target 7-year-old child, a concentration rate of 65.4% was observed, while a concentration rate of 75.1% was achieved with the target 3-year-old children. Ball-throwing was deemed unsuitable. Future work involves testing the block stacking content, obtaining additional data for analysis, and using the collected data to develop a diagnostic and treatment deep learning model. Challenges with optimizing VR devices currently result in lag issues in the content, which also need to be addressed.

### Acknowledgmen

This research was supported by Korea Institute for Advancement of Technology(KIAT) grant funded by the Korea Government(MOTIE) (P0012724, HRD Program for industrial Innovation

#### REFERENCES

- [1] <https://www.index.go.kr/unify/idx-info.do?pop=1&idxCd=5061>  
<https://doi.org/10.5626/JCSE.2019.13.3.124>
- [2] 이소현. "자폐 범주성 장애의 조기발견 및 조기개입의 역할 및 과제" 유아특수교육연구 9, no.1 (2009) : 103-133.
- [3] [https://www.mohw.go.kr/board.es?mid=a10503000000&bid=0027&tag=&act=view&list\\_no=372831&cg\\_code=](https://www.mohw.go.kr/board.es?mid=a10503000000&bid=0027&tag=&act=view&list_no=372831&cg_code=)

# Computer-aided hepatocellular carcinoma detection on the hepatobiliary phase of gadoxetic acid-enhanced magnetic resonance imaging using a convolutional neural network: Feasibility evaluation with multi-sequence data

Dohyun Kim<sup>1</sup>, Yongwon Cho<sup>1</sup>, Yeo Eun Han<sup>2</sup>, Min Ju Kim<sup>2</sup>, Beom Jin Park<sup>2</sup>, Yang Shin Park<sup>3</sup><sup>1</sup>Department of ICT Convergence, Soonchunhyang University, Asan, 31538, Republic of Korea

<sup>1</sup>Department of Computer Science and Engineering, Soonchunhyang University, South Korea, Republic of Korea

<sup>2</sup>Department of Radiology, Korea University Anam Hospital, Korea University College of Medicine, 73, Goryeodae-ro, Seongbuk-gu, Seoul 02841, Republic of Korea

<sup>3</sup>Department of Radiology, Korea University Guro Hospital, Korea University College of Medicine, 148, Gurodong-ro, Guro-gu, Seoul, 08308, Republic of Korea

\*Contact: [dragon1won@sch.ac.kr](mailto:dragon1won@sch.ac.kr)

**Abstract— Background and Objectives:** Analyzing multiple sequences of liver MRI is essential for diagnosing hepatocellular carcinoma (HCC). However, developing computer-aided detection (CAD) for each sequence is time-consuming and labor-intensive due to image segmentation. Therefore, we devised a CAD system specifically for HCC detection on the hepatobiliary phase (HBP) of gadoxetic acid-enhanced MRI, employing a convolutional neural network (CNN). We assessed its viability across various sequences, units, and multi-centers.

**Methods:** We conducted a review of patients who underwent both gadoxetic acid-enhanced MRI and surgery for hepatocellular carcinoma (HCC) at Korea University Anam Hospital (KUAH) and Korea University Guro Hospital (KUGH). Finally, our study included 170 nodules from 155 consecutive patients at KUAH and 28 nodules from 28 randomly selected patients at KUGH. Regions of interest were delineated across the entire volume of HCC on images captured during the hepatobiliary phase (HBP), T1-weighted (T1WI), T2-weighted (T2WI), and portal venous phase (PVP). A computer-aided detection (CAD) system was developed using customized-nnUNet based on the HBP images from KUAH and refined to reduce false positives. Both internal and external validation of the CAD system was conducted using images from HBP, T1WI, T2WI, and PVP acquired at both KUAH and KUGH. **Results:** The figure of merit and recall of the jackknife alternative free-response receiver operating characteristic of the CAD for HBP, T1WI, T2WI, and PVP at false-positive rate 0.5 were (0.87 and 87.0), (0.73 and 73.3), (0.13 and 13.3), and (0.67 and 66.7) in KUAH and (0.86 and 86.0), (0.61 and 53.6), (0.07 and 0.07), and (0.57 and 53.6) in KUGH, respectively.

**Conclusions:** The computer-aided detection (CAD) system for hepatocellular carcinoma (HCC) on gadoxetic acid-enhanced MRI, created through convolutional neural network (CNN) analysis of the hepatobiliary phase (HBP), effectively identified HCCs on both HBP and additional sequences like T1-weighted imaging (T1WI) and portal venous phase (PVP). This suggests that a CAD system developed using a single MRI sequence may be transferable to similar sequences, streamlining the process and saving time and effort in multi-sequence MRI CAD development.

**Keyword:** Abdominal Image Analysis Computer-Aided Diagnosis Deep Learning Hepatocellular Carcinoma Magnetic Resonance Imaging

## 1. Introduction

Hepatocellular carcinoma (HCC) is the most common type of primary liver cancer. HCC is the fifth most common cancer worldwide and the third most common cause of cancer-related deaths, according to the World Health Organization [1]. HCC is most commonly diagnosed based on typical imaging findings from computed tomography (CT) and magnetic resonance imaging (MRI). Liver MRI with gadoxetic acid has advantages in small lesion detection and consists of multi-sequence images including T2-weighted imaging (T2WI), T1-weighted imaging (T1WI), dynamic enhancement study (arterial phase, portal venous phase, transitional phase), hepatobiliary phase (HBP), chemical shift imaging, and diffusion-weighted imaging. The HBP is taken 20 min after intravenous gadoxetic acid administration, and hepatocytes show hyperintensity due to the uptake of gadoxetic acid. HBP significantly improves the sensitivity of HCC detection [2]. Typically, HCCs show hypointensity on HBP and T1WI, hyperintensity on T2WI, enhancements on arterial phase, and wash-out on portal venous phase (PVP) or transitional phase. HBP is the most sensitive sequence for HCC detection [2]. Nevertheless, information from all sequences should be used for the accurate diagnosis of HCC.

After the introduction of machine learning in radiology, computer-aided detection (CAD) for focal liver lesions has been developed. Recently, deep learning algorithms have shown promise in medical images [3], [4], [5], [6]. Deep learning has successfully developed automatic liver lesion segmentation algorithms using CT. Notably, a method using a convolutional neural network (CNN) [7] was superior to others in the 2017 liver tumor segmentation (LiTS) challenge [8] with 130 contrast-enhanced abdominal CT scans. Sun et al. [9] introduced the method of deep CNNs on multi-phase contrast-



enhanced CT images and showed outstanding results compared to previous studies employing monophasic images for detection. CT imaging [10] can help distinguish between healthy tissue, cirrhotic tissue, and HCC converging CNN and semi-automatic algorithms. For MRI, Bousabarah et al. [11] demonstrated the automated detection of HCC on multiphasic contrast-enhanced MRI using deep learning. Another study [12] successfully applied an automated deep learning model to detect HCC on the HBP of gadoxetic acid-enhanced MRI. Deep learning using CNN is a reasonable strategy for developing CAD to detect HCC on gadoxetic acid-enhanced MRI. Some researched multiple segmentation networks using various loss function on ultrasound images [13] and developed fusion model for accurate HCC detection using ultrasound and CT images [14].

For accurate computer-aided detection (CAD) of hepatocellular carcinoma (HCC) on gadoxetic acid-enhanced MRI, the ability to detect HCC across multiple image sequences is crucial. However, training CAD models for each sequence individually demands significant time and effort for image segmentation. If CAD models trained specifically on the hepatobiliary phase (HBP), known for its high sensitivity in HCC detection, could effectively identify HCC on other sequence images, it would streamline CAD development, reducing both time and costs. To date, no study has investigated the feasibility of CAD trained on a single sequence for analyzing multi-sequence liver MRI. In our research, we developed a CNN-based CAD system for HCC, trained exclusively on HBP images from gadoxetic acid-enhanced MRI, and assessed its viability for detecting HCC across HBP, T1-weighted imaging (T1WI), T2-weighted imaging (T2WI), and portal venous phase (PVP) sequences obtained from various units and centers.

## 2. Methods

The institutional review board approved this retrospective cohort study, and the requirement for informed consent was waived (2021AN0221 in Korea University Anam Hospital, 2021GR0311 in Korea University Guro Hospital).

### 2.1. Study participants

This retrospective study reviewed 324 consecutive cases (298 patients) who underwent surgery for suspected HCC between January 2015 and March 2020 at Korea University Anam Hospital (KUAH). Patients who underwent gadoxetic acid-enhanced MRI within two months before surgery and were pathologically diagnosed with primary HCC after surgery were included. Exclusion criteria were post-treatment state of HCC (transarterial chemoembolization, radiofrequency ablation), inadequate image quality for analysis (poor image quality, inadequate MRI protocol), and difficulty in drawing region of interest (ROI) (smaller than 1 cm, no lesion consistent with pathologic reports, infiltrative HCC, overlapped with biliary hamartoma). Finally, 163 cases (155 patients) were included, as presented in Fig. 1 and Table 1.

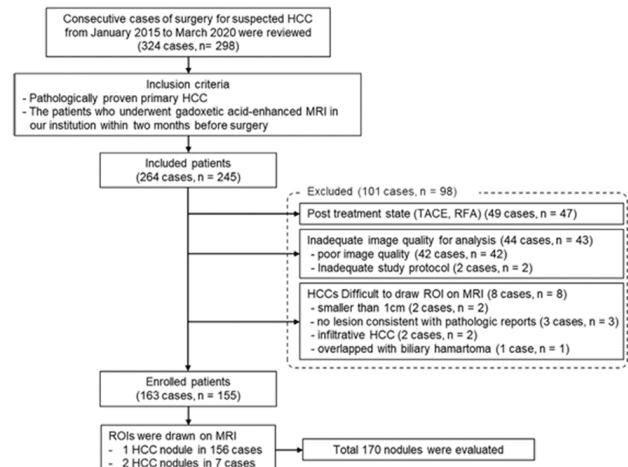
**Table 1.** Performance of computer-aided hepatocellular carcinoma detection on internal and external test datasets in HBP, T1WI, T2WI, and PVP.

	JAFROC FOM	Recall (%) <sup>*</sup>	Rate of FP (%) <sup>†</sup>
Internal, HBP	0.87	87.0 (80.0–87.0)	0.5
Internal, T1WI	0.73	73.3 (70.0–73.3)	0.5
Internal, T2WI	0.13	13.3 (13.0–13.3)	0.5
Internal, PVP	0.67	66.7 (36.0–66.7)	0.5
External, HBP	0.86	86.0 (68.0–86.6)	0.5
External, T1WI	0.61	53.6 (11.0–57.2)	0.5
External, T2WI	0.07	7.0 (6.0–7.0)	0.5
External, PVP	0.57	53.6 (28.6–57.1)	0.5

**Note:** internal datasets were from Korea University Anam Hospital (KUAH) and external datasets were from Korea University Guro Hospital (KUGH). HBP: hepatobiliary phase, T1WI: T1-weighted imaging, T2WI: T2-weighted imaging, PVP: portal venous phase, FOM: figure of merit, FP: false-positive, JAFROC: jackknife alternative free-response receiver operating characteristic; only the HBP datasets from KUAH were used for training and HBP, T1WI, T2WI, and PVP datasets from KUAH and KUGH were used for test.

<sup>\*</sup> Per-HCC-based sensitivities of HCC detection were calculated by dividing the number of detected HCC by the number of patients, for which the threshold of confidence score was set at 0.1.

<sup>†</sup> Rates of false-positive were calculated as the total number of HCCs with false-positives divided by the total number of patients, for which the confidence score threshold was set at 0.1.



**Fig.1** – Flowchart of participant selection. HCC: hepatocellular carcinoma, MRI: magnetic resonance imaging, TACE: Transcatheter arterial chemoembolization, RFA: radiofrequency ablation, ROI: region of interest.

### 2.2. CAD algorithm for HCC detection

Customized nnU-net [15] was used to detect HCC on MRI. Fig. 2 shows this architecture comprising an encoder network with 30 convolutional filters, a pooling layer (max-pooling:  $3 \times 3 \times 3$ ) per layer, and a decoder network with transposed convolutional layers for backpropagation.

First, whole volumes, including HCC, were reduced for input, whereas the last layer of the decoder restored the original volumes. Second, the inference of HCC in the decoder network was cropped for input to the center of the lesion, and our architecture was retrained using these datasets. A prominent feature of the customized architecture is the concatenation of the encoder and decoder networks to avoid missing segmentation information. For training, 120 batches were conducted for an epoch, and random rectified linear unit (ReLU) was used instead of the leaky rectified linear unit activation functions. The loss function sums cross-entropy, dice, and boundary loss. In addition, adaptive layer-instance normalization (AdaLin) [16] was applied to help the attention-guided model correspond to shape transformation.

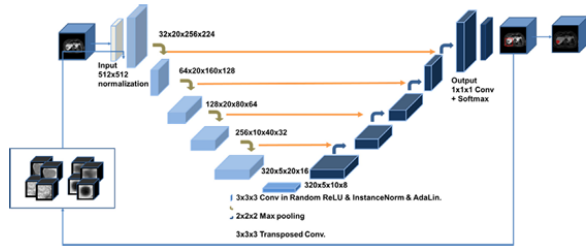


Fig.2 – Architecture of Customized nnU-net.

The initial learning rate and l2 weight decay were  $3 \times 10^{-4}$  and  $3 \times 10^{-5}$  as Adam optimization. If the validation loss did not improve within the previous 30 epochs, the learning rate was decreased by 0.2 times, and training was stopped after approximately 1000 epochs or if the learning rate fell below  $10^{-6}$ . First, all inputs were resized to  $208 \times 208$  pixels (XY spatial size), including the number of slices along the Z direction with intensity normalization by subtracting the mean and dividing by the standard deviation. Second, cropped inputs were resized to  $100 \times 100 \times 100$  volumes. Various augmentations have been used with photographic and geographic methods. The dice similarity coefficient (DSC) was analyzed for segmentation and detection, as shown in Eq. (1). The loss functions, including the dice loss (DLS), boundary loss (BLS) [17], and binary cross-entropy, are defined in Eqs. (2), (3), and (4):  $V_{seg}$  and  $V_{seg}$  are defined as the parameters of the ground truth and CNN inference.

$$DSC(V_{seg}, V_{gt}) = \frac{2|V_{seg} \cap V_{gt}|}{|V_{seg}| + |V_{gt}|} \quad (1)$$

$$DLS = 1 - \frac{2|V_{seg} \cap V_{gt}|}{|V_{seg}| + |V_{gt}|} \quad (2)$$

$$BLS(\partial G, \partial S) = 2 \int_{\Delta S} \|q - Z\partial G(q)\| dq \quad (3)$$

Here,  $\Delta S$  denotes the region between  $\|q - Z\partial G(q)\|$  and the two contours,  $\Omega \rightarrow \mathbb{R}^+$  is a distance map with respect to boundary  $\partial G$ , that is,  $\|q - Z\partial G(q)\|$  evaluates the distance between point  $q \in \Omega$  and the nearest point  $z\partial G(q)$  on contour  $\partial G$ :  $\|q - Z\partial G(q)\|$ .

$$L(y, f) = -y \log f - (1 - y) \log(1 - f) \quad (4)$$

where  $y$  and  $f$  denote the inferred probability and corresponding desired output, respectively.

### 2.3. Evaluation metrics for HCC detection

For the evaluation of CAD on the four sequences of MRI, the recall from the free-response receiver operating characteristic (FROC) curve (python-sci-kit-learn library) which was graphed to show the relationship between per-HCC recall depending on the threshold of the algorithm changes from 0 to 1.0. and the average false-positive (FP) numbers (FP rates). The intersection of union was over 0.5 of the box coordinates of HCC detection between inferences of CAD and gold standards. Furthermore, the ROI-wise classification (per HCC-based analysis) was conducted by activation values using figure of merit (FOM) of the jackknife alternative FROC (JAFROC) (version 4.2.1; <http://www.devchakraborty.com>), using the test data set from KUAH (internal validation) and KUGH (external validation). The FOM is defined as the probability that a true-positive lesion will be rated higher than the highest-rated FP lesion in normal cases.

The performance of the CAD system for HCC detection was evaluated using a FROC curve in Fig. 3 (a) and (b). We calculated the recall of HCC detection on the internal and external datasets.

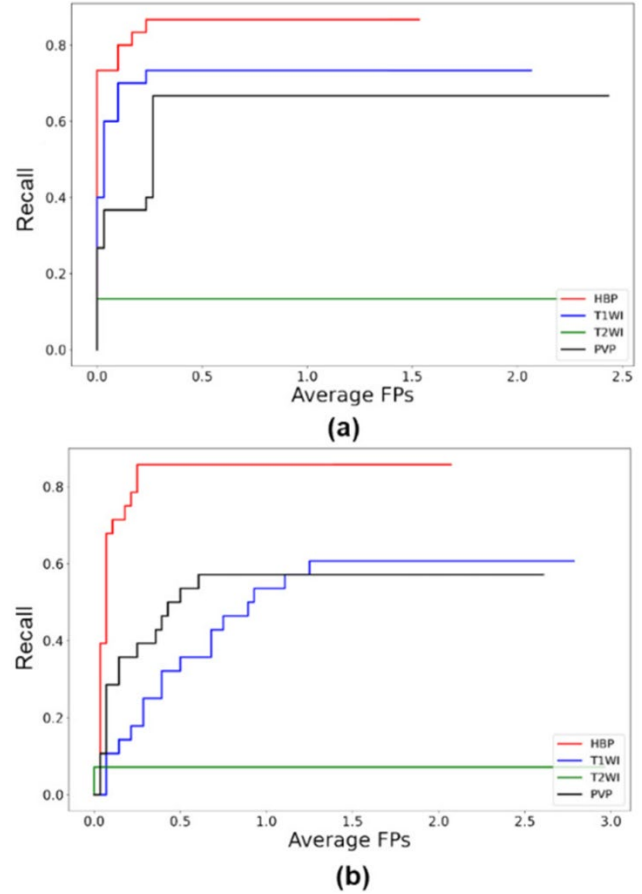


Fig. 3. Free-response receiver operating characteristic curves of computer-aided hepatocellular carcinoma detection developed using customized-nnUNet and HBP images. (a) internal dataset (Korea University Anam Hospital) (b) external dataset (Korea University Guro Hospital); HBP: hepatobiliary phase, T1WI: T1-weighted imaging, T2WI: T2-weighted imaging, PVP: portal venous phase, FP: false-positive.

To evaluate the performance of CAD for HCC detection on multi-sequence MRI in KUAH, the cut-off threshold (0.5) was determined using recall and average FPs in our algorithm. These cut-off thresholds for HCC detection performance were selected empirically as average FPs in the FROC curve of the test set of HBP in Fig. 3(a). At this cut-off threshold (0.5), the recalls of CAD in the internal test set (KUAH) were 87.0%, 73.3%, 13.3%, and 66.7%, in HBP, T1WI, T2WI, and PVP, respectively (Table 3). The recalls of CAD in the external test set (KUGH) were 86.0%, 61.0%, 7.0%, and 57.0%, in HBP, T1WI, T2WI, and PVP, respectively (Table 3). In addition, we evaluated JAFROC FOM. Table 3 shows the FOM of the JAFROC in KUAH and KUGH. The FOMs in the internal test set (KUAH; HBP\*, T1WI, T2WI, and PVP;  $P = .498$ ,  $P = .009$ , and  $P = .301$ ) were 0.88, 0.73, 0.13, and 0.67, respectively. The values in the external test set (KUGH; HBP\*, T1WI, T2WI, and PVP;  $P = .180$ ,  $P = .001$ , and  $P = .137$ ) were 0.86, 0.61, 0.01, and 0.57, respectively.

Table 3. Performance of computer-aided hepatocellular

carcinoma detection on internal and external test datasets in HBP, T1WI, T2WI, and PVP.

Empty Cell <sup>⊙</sup>	JAFROC FOM <sup>⊙</sup>	Recall (%) <sup>⊙</sup>	Rate of FP (%) <sup>⊙</sup>
Internal, HBP <sup>⊙</sup>	0.87 <sup>⊙</sup>	87.0 (80.0–87.0) <sup>⊙</sup>	0.5 <sup>⊙</sup>
Internal, T1WI <sup>⊙</sup>	0.73 <sup>⊙</sup>	73.3 (70.0–73.3) <sup>⊙</sup>	0.5 <sup>⊙</sup>
Internal, T2WI <sup>⊙</sup>	0.13 <sup>⊙</sup>	13.3 (13.0–13.3) <sup>⊙</sup>	0.5 <sup>⊙</sup>
Internal, PVP <sup>⊙</sup>	0.67 <sup>⊙</sup>	66.7 (36.0–66.7) <sup>⊙</sup>	0.5 <sup>⊙</sup>
External, HBP <sup>⊙</sup>	0.86 <sup>⊙</sup>	86.0 (68.0–86.6) <sup>⊙</sup>	0.5 <sup>⊙</sup>
External, T1WI <sup>⊙</sup>	0.61 <sup>⊙</sup>	53.6 (11.0–57.2) <sup>⊙</sup>	0.5 <sup>⊙</sup>
External, T2WI <sup>⊙</sup>	0.07 <sup>⊙</sup>	7.0 (6.0–7.0) <sup>⊙</sup>	0.5 <sup>⊙</sup>
External, PVP <sup>⊙</sup>	0.57 <sup>⊙</sup>	53.6 (28.6–57.1) <sup>⊙</sup>	0.5 <sup>⊙</sup>

**Note:** internal datasets were from Korea University Anam Hospital (KUAH) and external datasets were from Korea University Guro Hospital (KUGH). HBP: hepatobiliary phase, T1WI: T1-weighted imaging, T2WI: T2-weighted imaging, PVP: portal venous phase, FOM: figure of merit, FP: false-positive, JAFROC: jackknife alternative free-response receiver operating characteristic; only the HBP datasets from KUAH were used for training and HBP, T1WI, T2WI, and PVP datasets from KUAH and KUGH were used for test.

\* Per-HCC-based sensitivities of HCC detection were calculated by dividing the number of detected HCC by the number of patients, for which the threshold of confidence score was set at 0.1.

† Rates of false-positive were calculated as the total number of HCCs with false-positives divided by the total number of patients, for which the confidence score threshold was set at 0.1.

### 3. Discussion and conclusion

In this study, we devised a computer-aided detection (CAD) system for hepatocellular carcinoma (HCC) on gadoxetic-enhanced MRI utilizing customized-nnUNet. Our training solely employed the hepatobiliary phase (HBP) images from Korea University Anam Hospital (KUAH), yet the CAD effectively identified HCCs on HBP, portal venous phase (PVP), and T1-weighted imaging (T1WI) at both KUAH and Korea University Guro Hospital (KUGH). To our knowledge, this is the first examination assessing the feasibility of CAD for HCC, originating from a single sequence image (HBP) of liver MRI and tested across multiple centers and sequences.

In conclusion, our deep learning-based CAD system, trained solely on the HBP sequence, proficiently detected HCCs on T1WI and PVP sequences. Our findings indicate that CAD trained on a single sequence MRI could be readily applicable to other sequences, particularly those employing similar imaging parameters. This outcome promises to streamline the training and development process of CAD for MRI, encompassing multiple sequences, thereby reducing time and resources.

### ACKNOWLEDGMENT

This research was supported by a grant of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea [grant number HI22C1302]. This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (RS-2023-00239603, RS-2023-00218176) and the Soonchunhyang University Research Fund.

### REFERENCES

- [1] References
- [2] [1] J.M. Llovet, A. Burroughs, J. Bruix, Hepatocellular carcinoma, *Lancet* 362 (9399)
- [3] (2003) 1907–1917, doi:10.1016/s0140-6736(03)14964-1.
- [4] [2] D.M. Koh, A. Ba-Ssalamah, G. Brancatelli, G. Fananapazir, M.I. Fiel, S. Goshima,
- [5] S.H. Ju, N. Kartalis, M. Kudo, J.M. Lee, et al., Consensus report from the 9(th)
- [6] International Forum for Liver Magnetic Resonance Imaging: applications of gadoxetic acid-enhanced imaging, *Eur. Radiol.* 31 (8) (2021) 5615–5628, doi:10.1007/s00330-020-07637-4.
- [7] [3] G. Litjens, T. Kooi, B.E. Bejnordi, A.A.A. Setio, F. Ciompi, M. Ghafoorian, J. van
- [8] der Laak, B. van Ginneken, C.I. Sánchez, A survey on deep learning in medical
- [9] image analysis, *Med. Image Anal.* 42 (2017) 60–88, doi:10.1016/j.media.2017.07.005.
- [10] [4] F. Li, R. Engelmann, C.E. Metz, K. Doi, H. MacMahon, Lung cancers missed
- [11] on chest radiographs: results obtained with a commercial computer-aided
- [12] detection program, *Radiology* 246 (1) (2008) 273–280, doi:10.1148/radiol.2461061848.
- [13] [5] C.S. White, T. Flukinger, J. Jeudy, J.J. Chen, Use of a computer-aided detection
- [14] system to detect missed lung cancer at chest radiography, *Radiology* 252 (1) (2009) 273–281, doi:10.1148/radiol.2522081319.
- [15] [6] S. Schalekamp, B. van Ginneken, E. Koedam, M.M. Snoeren, A.M. Tiehuis,
- [16] R. Wittenberg, N. Karssemeijer, C.M. Schaefer-Prokop, Computer-aided detection improves detection of pulmonary nodules in chest radiographs beyond
- [17] the support by bone-suppressed images, *Radiology* 272 (1) (2014) 252–261, doi:10.1148/radiol.14131315.
- [18] [7] X. Han, MR-based synthetic CT generation using a deep convolutional neural
- [19] network method, *Med. Phys.* 44 (4) (2017) 1408–1419, doi:10.1002/mp.12155.
- [20] [8] P. Christ, F.G. Ettlinger, J.K. Lipkova, LiTS - Liver Tumor Segmentation Challenge
- [21] (2017) <http://www.litschallenge.com/>. (accessed).
- [22] [9] C. Sun, S. Guo, H. Zhang, J. Li, M. Chen, S. Ma, L. Jin, X. Liu, X. Li, X. Qian,
- [23] Automatic segmentation of liver tumors from multiphase contrast-enhanced
- [24] CT images based on FCNs, *Artif. Intell. Med.* 83 (2017) 58–66, doi:10.1016/j.artmed.2017.03.008.
- [25] [10] A. Nayak, E. Baidya Kayal, M. Arya, J. Culli, S. Krishan, S. Agarwal, A. Mehndiratta, Computer-aided diagnosis of cirrhosis and hepatocellular carcinoma using multi-phase abdomen CT, *Int. J. Comput. Assist. Radiol. Surg.* 14 (8) (2019) 1341–1352, doi:10.1007/s11548-019-01991-5.
- [26] [11] K. Bousabarah, B. Letzen, J. Tefera, L. Savic, I. Schobert, T. Schlachter, L.H. Staib,
- [27] M. Kocher, J. Chapiro, M. Lin, Automated detection and delineation of hepatocellular carcinoma on multiphase contrast-enhanced MRI using deep learning,
- [28] *Abdom. Radiol. (NY)* 46 (1) (2021) 216–225, doi:10.1007/s00261-020-02604-5.
- [29] [12] J. Kim, J.H. Min, S.K. Kim, S.Y. Shin, M.W. Lee, Detection of Hepatocellular Carcinoma in Contrast-Enhanced Magnetic Resonance Imaging Using Deep Learning Classifier: A Multi-Center Retrospective Study, *Sci. Rep.* 10 (1) (2020) 9458, doi:10.1038/s41598-020-65875-4.
- [30] [13] Flaviu Vancea, et al., Hepatocellular Carcinoma Segmentation within Ultrasound Images using Convolutional Neural Networks, *IEEE, ICCP*, 2019, doi:10.

- [39] 1109/ICCP48234.2019.8959687.
- [40] [14] Corina Radu, et al., Integration of Real-Time Image Fusion in the RoboticAssisted Treatment of Hepatocellular Carcinoma, biology MDPI 9 (11) (2020)
- [41] 397, doi:10.3390/biology9110397.
- [42] [15] F. Isensee, J. Petersen, S.A. Kohl, P.F. Jäger, K.H. Maier-Hein, nnU-Net: a selfconfiguring method for deep learning-based biomedical image segmentation, Nature Methods 18 (2021) 203–211, doi:10.1038/s41592-020-01008-z.
- [43] [16] H. Kervadec, J. Bouchtiba, C. Desrosiers, E. Granger, J. Dolz, I. Ben Ayed, Boundary loss for highly unbalanced segmentation, Med. Image Anal. 67 (2021)
- [44] 101851, doi:10.1016/j.media.2020.101851.
- [45] [17] C.P. Hess, D.D. PurcellT.P. Naidich, M. Castillo, S. Cha, J.G. Smirniotopoulos
- [46] (Eds.), Analysis of density, signal intensity, and echogenicity, Imaging of the
- [47] Brain (2013) 45–66.
- [48] [18] J.L. Bloem, M. Reijnierse, T.W.J. Huizinga, A.H.M. van der Helm-van Mil, MR
- [49] signal intensity: staying on the bright side in MR image interpretation, RMD Open 4 (1) (2018) e000728, doi:10.1136/rmdopen-2018-000728.
- [50]

# A Study of a Speech Recognition Model for Patients with Speech Disorders

Dayeong So<sup>1</sup>, Shilong Liu<sup>1</sup>, Sreypov Van<sup>1</sup>, Chomyong Kim<sup>1</sup>,  
Yunyoung Nam<sup>1,2,\*</sup>, Jiyoung Woo<sup>1,3,\*</sup>, and Jihoon Moon<sup>1,3,\*</sup>

<sup>1</sup> Department of ICT Convergence, Soonchunhyang University, Asan 31538, Republic of Korea

<sup>2</sup> Department of Computer Science and Engineering, Soonchunhyang University, Asan 31538, Republic of Korea

<sup>3</sup> Department of AI and Big Data, Soonchunhyang University, Asan 31538, Republic of Korea

\*Contact: {ynam, jywoo, jmoon22}@sch.ac.kr, phone +82 41 530 4956

**Abstract**—In this study, we developed a technology to augment children's speech data using the MaskCycleGAN-VC model. During the data preprocessing phase, we converted adult and child speech data into log Mel spectrograms. During model training, these log Mel spectrograms were used as input. Each generator inferred the speech frequency for silent frames using the Filling in Frame (FIF) technique to produce speech, while discriminators evaluated the authenticity of the generated voices. Efforts were made to improve the quality difference between the original and generated voices by reducing the loss function values between the original and generated data. This approach effectively addresses the scarcity of children's speech data and increases the diversity of training data by transforming existing speech data into different forms. We expect that this method can serve as a preliminary study for the development of automatic speech recognition (ASR) systems for children with developmental speech disorders.

## I. INTRODUCTION

In the context of the Fourth Industrial Revolution, the advent of artificial intelligence (AI) has had a profound impact on our daily lives, fundamentally transforming the way we interact with technology. Despite these advances, a significant digital divide persists, with children with developmental disabilities being particularly affected. This divide not only restricts their access to emerging technologies but also limits their ability to engage fully in increasingly digital educational environments. While automatic speech recognition (ASR) technology has undergone significant advancements, its application has been predominantly geared towards the general population, with the specific needs of children with speech and language disorders being largely overlooked [1].

This oversight is of critical importance, as children with developmental speech impairments frequently encounter difficulties with conventional ASR systems that fail to account for their unique speech patterns, which may include atypical phonetics and inconsistent speech dynamics. Traditional ASR technologies, based on hidden Markov models (HMM) and Gaussian mixture models (GMM), while effective under standard conditions, frequently fail to capture the nuanced variations in speech exhibited by children with developmental challenges [2]. Moreover, the full potential of recent enhancements in computing power and model accuracy has yet

to be realized in addressing the specific needs of children with developmental speech impairments.

The objective of our research is to bridge this gap by employing the MaskCycleGAN-VC model, an innovative approach that adapts the timbre of child speech data sourced from AI Hub to create more inclusive and effective ASR systems [3]. By focusing on the distinct and diverse speech characteristics of children with developmental disabilities, this study aims to refine how ASR technology recognizes and processes their speech. This not only aids in better communication but also enhances their ability to express their thoughts and needs more clearly and accurately.

The contributions of this research are numerous and include three primary advancements:

- This study broadens the reach of AI and ASR technologies to include children with developmental disabilities, thereby facilitating their access to essential digital tools.
- Our research develops ASR systems tailored for greater engagement of children with developmental disabilities in both social and educational environments.
- The project bridges a critical technological divide, showcasing how advanced technologies can empower marginalized communities and foster societal equity.

The remainder of this paper is organized as follows: Session 2 describes the dataset configuration and model setup; Session 3 presents the experiments and their results; and Session 4 provides the conclusions and discusses future work.

## II. MATERIALS AND METHODS

### A. Data Collection and Preprocessing

The research uses a robust dataset of free conversational speech data collected from the AI Hub. This provides a diverse range of audio samples reflecting a variety of linguistic characteristics relevant to different demographic groups, particularly children. This conversion is essential as it represents the original form of the raw audio data in a format suitable for further digital processing. A mathematical algorithm is used to decompose each audio waveform into its component frequencies.



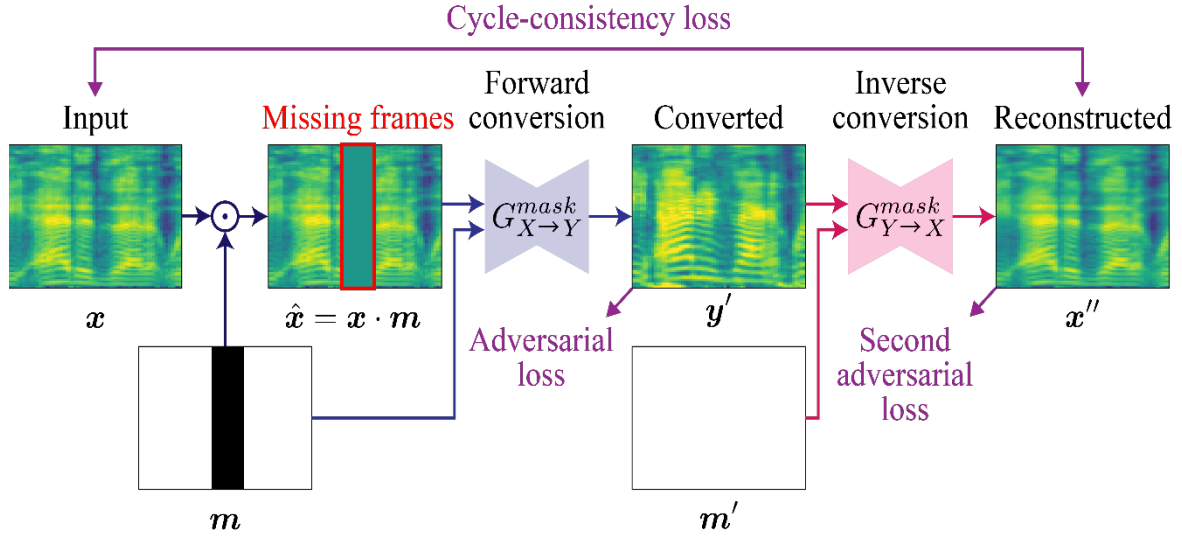


Fig. 1 Architecture of the MaskCycleGAN-VC Model

The Fast Fourier Transform (FFT) is a fundamental signal processing technique that transforms time-domain data into frequency-domain data, revealing the spectral components of the audio signal. To analyze audio signals in shorter segments and capture the time-varying characteristics of speech, the Short-Time Fourier Transform (STFT) is used. This method makes it easier to understand how the frequencies of the audio signal fluctuate over time, which is critical for detecting nuances in speech that may occur over short intervals.

Following the spectral analysis, Mel spectra are generated to represent the speech signals. These are subjected to exploratory data analysis (EDA) to identify patterns, detect anomalies, and test hypotheses about the underlying structures of the data. This step is critical for tailoring the preprocessing and model training phases to the specific characteristics of the speech data, thereby ensuring optimal performance of the speech recognition model.

### B. Voice Style Transfer Using MaskCycleGAN-VC

The primary goal of this method, as shown in Fig. 1, is to transfer specific speech features such as intonation, speed, and intensity from an adult source speaker to a child target speaker while preserving the natural timbre of the child's voice [4]. The MaskCycleGAN-VC, an advanced variant of the CycleGAN used for nonparallel voice conversion, uses a style transfer mechanism that includes feature mapping and voice conversion. In feature mapping, the model learns the characteristics of the source speaker's speech style using Mel spectrograms. The loss functions employed—adversarial loss, identity mapping loss, and cycle loss—each ensure the quality and consistency of the voice style transfer. The results and evaluation of the project, including both subjective listening tests and objective measures such as Mel Cepstral Distortion (MCD), assess the quality and effectiveness of the voice style transfer.

### C. Nonparallel Data Augmentation with FIF Technique

The objective of this study is to improve the robustness and versatility of the MaskCycleGAN-VC Model by enabling it to perform data augmentation using the Filling in Frames (FIF) technique. This approach is advantageous for maintaining the continuity and natural flow of speech in augmented data. The

FIF technique uses advanced algorithms to infer the characteristics of the missing audio based on the characteristics of adjacent frames. The effectiveness of the FIF technique is evaluated by contrasting the synthesized speech with the original recordings, facilitating fine-tuning of the model for improved accuracy and naturalness.

### D. Discriminator Training and Loss Optimization

The task of the discriminator is to distinguish between real and synthetic audio samples. In the generative adversarial network (GAN) framework of MaskCycleGAN-VC, the discriminator plays a central role in this task. The training of the discriminator involves adversarial principles, where it competes with the generator to accurately identify authentic and synthetic audio. This competition improves both the accuracy of the discriminator and the output quality of the generator. Feedback from the discriminator is used to adjust the parameters of the generator. Loss functions for the discriminator include adversarial loss, which measures its ability to distinguish between real and false samples; identity loss, which preserves the voice characteristics of the target speaker; and cycle consistency loss, which ensures that the original audio can be reconstructed from the converted audio. Gradient descent is used to optimize these loss functions and regularization. Techniques such as dropout and L2 regularization prevent overfitting. The learning rate of the discriminator is controlled by adaptive methods such as Adam or RMSprop. Regular validation checks using a held-out data set assess the performance of the discriminator in real scenarios, monitoring metrics such as accuracy, precision and recall.

## III. RESULTS AND DISCUSSION

Despite the critical demand for tailored speech recognition technologies for children with developmental disabilities, challenges persisted due to limited active data collection efforts. To address these gaps, this study successfully collected and analyzed voice data from children aged 3 to 10 years nationwide. The study employed the MaskCycle GAN for voice conversion, utilizing adult voice data from free conversation samples provided by AI Hub. The comprehensive dataset included



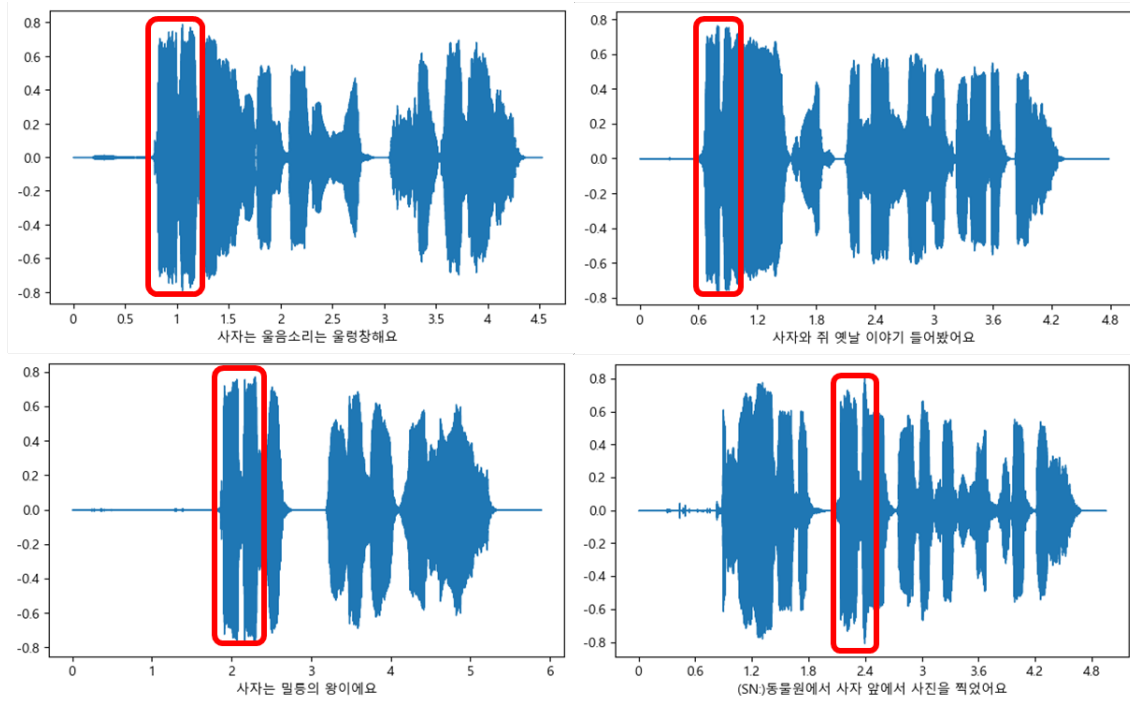


Fig. 2 Visualization of Wave Form results

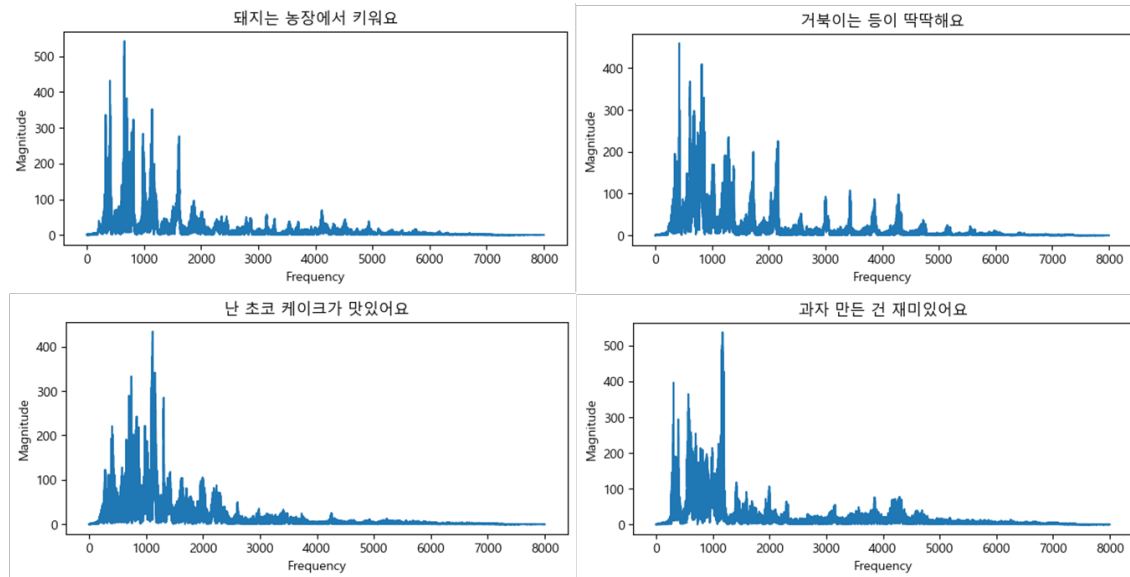


Fig. 3 Visualization of FFT results

voices of individuals ranging in age from teenagers to those in their fifties, with a total duration of over 4,000 hours from more than 2,000 speakers. Prior to the generation of synthetic child speech, an exhaustive analysis of the existing child voice signals was conducted.

The results of this analysis were meticulously depicted in Fig. 2–5, which illustrate the distinctive waveform characteristics of child speech. Fig. 1 demonstrated a consistent waveform pattern when children pronounced specific words, such as “lion.” Fig. 2 indicated that children's voice frequencies were higher than the typical adult male frequency range of 85 Hz to 180 Hz, which highlights considerations for effective voice conversion regarding sound quality, intonation, and tone. Fig. 3 revealed

distinctive frequency signatures when children articulated words like “lion” and “-yo.” Fig. 4 demonstrated the efficacy of this analytical approach, as it revealed distinct patterns in the Mel spectrogram when these specific words were pronounced.

This structured presentation ensures clarity in conveying the results, highlighting the rigorous methodologies employed and the significant insights derived from the visual data analysis. The findings clearly substantiate the initial hypotheses and the effectiveness of the techniques used for speech characteristic analysis and voice conversion within this demographic. A total of 35 generated speech samples were evaluated, and the average MCD was found to be 519.8648, while the average Kernel Density Spectral Distance (KDSD) was 1.0289. While the

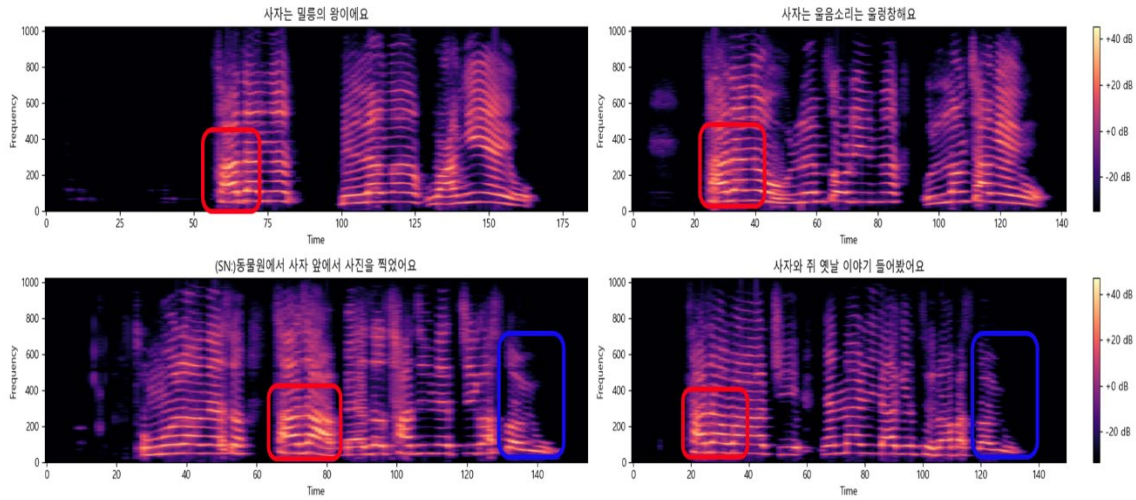


Fig. 4 Visualization of STFT results

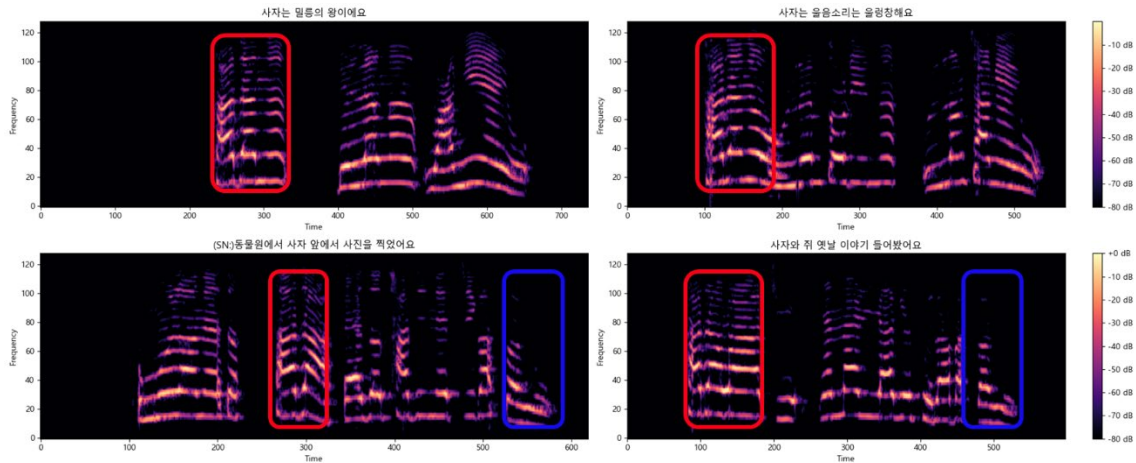


Fig. 5 Visualization of Mel Spectrogram results

KDSD values were relatively low, the MCD values were abnormally high. It is anticipated that this limitation will be gradually ameliorated by incrementally increasing the mask size from its current value of 50 to a higher value during training.

#### IV. CONCLUSIONS

In this study, a technology for augmenting speech data targeting children was developed using the MaskCycleGAN-VC model. During the data preprocessing phase, voice data from both adults and children were transformed into log Mel-spectrograms. In the model training phase, the extracted log Mel-spectrograms were utilized as inputs. The generators were tasked with inferring the speech frequencies that would fill silent frames using the FIF method to generate speech, while the discriminator was responsible for determining the authenticity of the generated voices. Efforts were made to reduce the loss function values between the original and generated data, with the objective of improving the quality difference between the original and generated speech. This approach successfully addressed the issue of insufficient child voice data and enhanced the diversity of training data by transforming existing voice data into various forms. It is anticipated that this methodology could serve as a precursor

study for the development of ASR systems for children with developmental disabilities.

#### ACKNOWLEDGMENT

This research was supported by the MSIT (Ministry of Science and ICT), Korea, under the ICAN (ICT Challenge and Advanced Network of HRD) support program (IITP-2024-2020-0-01832), supervised by the IITP (Institute for Information & Communications Technology Planning & Evaluation).

#### REFERENCES

- [1] K. Singh, et al., "Data Augmentation Using CycleGAN for End-to-End Children ASR," in *Proc. 29th European Signal Processing Conference (EUSIPCO)*, 2021.
- [2] N. Hailu, I. Siegert, and A. Nürnberger, "Improving Automatic Speech Recognition Utilizing Audio-codecs for Data Augmentation," in *Proc. IEEE 22nd International Workshop on Multimedia Signal Processing (MMSP)*, 2020.
- [3] J. Mun, et al., "Deep learning-based speech recognition for Korean elderly speech data including dementia patients," *The Korean Journal of Applied Statistics*, vol. 36, no. 1, pp. 33–48, 2021.
- [4] T. Kaneko, et al., "Maskcyclegan-VC: Learning Non-Parallel Voice Conversion with Filling in Frames," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021.

# A Fusion of Residual Blocks and Stack Auto Encoder Features for Stomach Cancer Classification

Abdul Haseeb<sup>1</sup>, Majed Alhaisoni<sup>3</sup>, Areej<sup>4</sup>, Mazah<sup>4</sup>, Usman Tariq<sup>5</sup>, Muhammad H Ali

<sup>1</sup>Department of CS, HITEC University, 47080, Taxila, Pakistan

<sup>2</sup> Department of Computer Science and Mathematics, Lebanese American University, Beirut, Lebanon

<sup>3</sup>College of Computer Science and Engineering, University of Ha'il, Ha'il 81451, Saudi Arabia

<sup>4</sup>Department of Information Systems, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, PO Box 84428, Riyadh 11671

<sup>5</sup>Prince Sattam Bin Abdulaziz University, Al-Kharaj 11942, Saudi Arabia

**Abstract—** Background- Several computerized solutions have been introduced for stomach disease detection and classification in recent years. The existing techniques faced several challenges, such as irrelevant feature extraction, high similarity among different disease symptoms, and the least important features from a single source. An enormous number of people are affected by gastrointestinal cancer. Diagnosing gastrointestinal cancer by classical means is a hazardous procedure. Deep learning has shown tremendous performance in the recent years for the classification tasks. Method- In this paper, a new deep learning based architecture is designed that is based on the fusion of two models- Residual blocks and Auto Encoder. The hyper-Kvasir dataset, which includes more than 20 classes, has been employed to evaluate the proposed work. A pre-trained CNN model has been selected and then improved by the addition of several residual blocks. This process aims to improve the learning capability of deep models and lessen the number of parameters. In addition, an Auto-Encoder-based network that consists of five convolutional layers in the encoder stage and five in the decoder phase is designed. The global average pooling and convolutional layers were selected for the feature extraction that was optimized using a hybrid Marine Predator optimization and Slime Mould optimization algorithm. The selected features of both models are fused using a novel fusion technique that is later classified using the Artificial Neural Networks classifiers. The experimental process is conducted on the HyperKvasir dataset, which consists of 23 stomach-infected classes. The proposed method obtained an improved accuracy of 93.90% on this dataset. Comparison is also

conducted with some recent techniques and shows that the proposed method's accuracy is improved.

**Keywords—** Stomach cancer; contrast enhancement; deep learning; residual blocks; Information fusion; feature selection; machine learning

## 1 Introduction

Gastrointestinal cancer, also known as digestive system cancer, refers to a group of cancers that occur in the digestive system or gastrointestinal tract, which includes the esophagus, stomach, small intestine, colon, rectum, liver, gallbladder, and pancreas [1, 2]. These cancers develop when cells in the digestive system grow abnormally and uncontrollably, forming a tissue mass known as a tumor [3]. Depending on the type and stage of the disease, the symptoms of gastrointestinal cancer might include stomach discomfort, nausea, vomiting, changes in bowel habits, weight loss, and exhaustion [4]. Gastrointestinal Tract cancer may be treated by surgery, chemotherapy, radiation therapy, or a combination. Detection and treatment at an early stage can enhance survival chances and minimize the risk of complications [5]. Despite a gradual decrease in gastric cancer incidence and mortality rates over the past 50 years, it remains the second most frequent cause of cancer-related deaths globally. However, from 2018 to 2020, both colorectal and stomach cancer have shown an upward trend in their rates [6]. Global Cancer Statistics shows that 26.3 percent of total cancer cases are from Gastrointestinal cancer, whereas the mortality rate is 35.4 percent among all cancers [7].

Identifying and categorizing gastrointestinal disorders subjectively is time-consuming and difficult, requiring much clinical knowledge and skill [8]. Yet, the development of effective computer-aided diagnosis (CAD) technologies that can identify and categorize numerous gastrointestinal disorders in a fully automated manner might reduce these diagnostic obstacles to a great extent [9]. Computer-aided diagnosis technologies can be of great value by aiding medical personnel in making accurate diagnoses and identifying appropriate therapies for serious medical diseases in their early stages [10, 11]. Over the past few years, the performance of diagnostic-based artificial intelligence (AI) computer-aided diagnosis tools in various medical fields has been significantly improved with the use of deep learning algorithms, particularly artificial neural networks (ANNs) [12]. Generally, these ANNs are trained using optimization algorithms such as stochastic gradient descent [13] to achieve the best accurate representation of the training dataset.

DL, which refers to deep learning, is a statistical approach that enables computers to automatically detect features from raw input, such as structured information, images, text, and audio [14, 15]. Many areas of clinical practice have been profoundly influenced by the significant advances made in AI based on DL [16, 17]. Computer-aided diagnosis systems in gastroenterology increasingly use artificial intelligence (AI) to improve the identification and characterization of abnormalities during endoscopy [18]. The CNN, a neural network influenced by the visual cortex of life forms, uses convolutional layers with common two-dimensional weight sets. This enables the algorithm to recognize spatial data and employ layer clustering to filter out less significant information, eventually conveying the most pertinent and focused elements [19]. However, these classifiers face a challenge in interpretability because they are often seen as "black boxes" that deliver accurate outcomes without explaining them [20]. Despite technological developments, image classification for lesions of the gastrointestinal system remains difficult due to a lack of databases containing sufficient images to build the models. In addition, the quality of accessible images has impeded the application of

CNN models [21].

### *1.1 Major Challenges*

In this work, Artificial Neural Networks (ANN) and Deep Neural Networks (DNN) extract the features of images from the Hyper-Kvasir dataset. The dataset contains twenty-three gastrointestinal tract classes with images in each class. However, some classes have only a few images, creating a data misbalancing problem. Data augmentation techniques are used for classes with fewer images to address this issue. Furthermore, feature selection techniques are implied to obtain the best features among feature sets.

### *1.2 Major Contributions*

The major contributions of the proposed method are described as follows:

- Proposed a fusion based contrast enhancement technique based on the mathematical formulation of local and global information enhanced filters, called Duo-contrast.
- A new CNN architecture is designed based on the concept of pre-trained NasnetMobile. Several residual blocks have been added to increase the learning capability and reduction of parameters.
- A stack Auto Encoder-Decoder network is designed that consists of five convolutional layers in the encoder phase and five in the decoder phase.
- The extracted features have been optimized using improved Marine Predator optimization and Slime Mould optimization algorithm.
- A new parallel fusion technique is proposed to combine the important information of both deep learning models.
- A detailed experimental process in terms of accuracy, confusion matrix, and t-test-based analysis has been conducted to show the significance of the proposed framework.

The rest of the manuscript is structured as follows: Section 2 describes the significant related work relevant to the study. Section 3 outlines the methodology utilized in the research, including

the tools, methods, and resources employed. Section 4 comprises a discussion of the findings acquired from the study. Section 5 provides the conclusions of the research.

## 2 Related Work

Gastrointestinal tract classification is a hot topic in research. In recent years, researchers have achieved important milestones in this work domain [22]. In their article, Borgli et al. introduced the Hyper-Kvasir dataset, which contains millions of images of gastrointestinal endoscopy examinations from Baerum Hospital located in Norway. The labeled images in this dataset can be used to train neural networks for discrimination purposes. The authors conducted experiments to train and evaluate classification models using two commonly used families of neural networks, ResNet and DenseNet, for the image classification problem. The labeled data in the Hyper-Kvasir dataset consists of twenty-three classes of gastrointestinal disorders. While the authors achieved the best results by combining ResNet-152 and DenseNet-161, the overall performance was still unsatisfactory due to imbalanced development sets [23]. In their proposal, Igarashi et al. employed AlexNet architecture to classify more than 85000 input images from Hirosaki University Hospital.

Moreover, the input images were categorized into 14 groups based on pattern classification of significant anatomical organs, with manual classification. To train the model, the researchers used 49,174 images from patients with gastric cancer who had undergone upper gastrointestinal tract endoscopies. In comparison, the remaining 36000 images were employed to test the model's performance. The outcome indicated an impressive overall accuracy of 96.5%, suggesting its potential usefulness in routine endoscopy image classification [24]. Gómez-Zuleta developed a deep learning (DL) methodology to detect polyps in colonoscopy procedures automatically. For this task, three models were used, namely Inception-v3, ResNet-50, and VGG-16. Knowledge transfer through transfer learning was adopted for classification, and the resultant weights were used to commence a fresh training process utilizing the fine-tuning technique with colonoscopy images. The training data consisted

of a combined dataset of five databases comprising more than 23000 images with polyps and more than 47000 images without polyps for validation, respectively. The data was split into a 70 by 30 ratio for training and testing purposes. Different metrics such as accuracy, F1-score, and receiver operating characteristic curve, commonly known as ROC, were employed to evaluate the performance. Pretrained model such that Inceptionv3, VGG16, and Resnet50 models achieved accuracy rates of 81%, 73%, and 77%, respectively. The authors described that pretrained network models demonstrated an effective generalization ability towards the high irregularity of endoscopy videos, and their methodology may potentially serve as a valuable tool in the future [25]. The authors employed three networks to classify medical images from the Kvasir database. They began using a preprocessing step to eliminate noise and improve image quality. Then, they utilized data augmentation methods to progress the network's training and a dropout method to prevent overfitting. Yet, the researchers acknowledged that this technique resulted in a doubling of the training time. The researchers also implemented Adam to optimize the loss to minimize error. Also, transfer learning and fine-tuning techniques are implied. The resulting models were then used to categorize 5,000 images into five distinct categories, with eighty percent of the database allocated for training and twenty percent for validation. The accuracy rates achieved by the models were 96.7% for GoogLeNet, 95% for ResNet-50, and 97% for AlexNet [26].

The Kvasir-Capsule dataset, presented in [27], includes 117 videos captured using video-capsule endoscopy (VCE). The dataset comprises fourteen different categories of images and a total of more than 47000 identified categorized images. VCE technology involves a small capsule with a camera, batteries, and other components. To validate the labeled dataset, two convolutional neural networks (CNNs), namely DenseNet\_161 and ResNet\_152, were used for training. The study utilized a cross-validation technique with definite cross-entropy-based loss to validate the models. They implemented this technique without class and with class weight and also used weight-based sampling to balance the dataset by

removing or adding the images for every class. After evaluating the models, the best results were obtained by averaging the outcomes of both CNNs. The resulting accuracy rates were 73.66% for the micro average and 29.94% for the macro average.

Overall, the researchers improved their categorization of the Hyper Kvasir data set. Yet, a significant gap in the subject matter must be filled. So, it must utilize a wonderful hybrid strategy incorporating deep learning and machine learning methodologies to get exceptional outcomes. Using machine learning approaches to discover key characteristics and automated deep feature extraction to uncover them may help increase classification accuracy.

### 3 Proposed Methodology

The dataset used in this manuscript is highly imbalanced as some classes have few images. To resolve this problem, data augmentation techniques are adopted. Nasnetmobile and Stacked Autoencoders are used as feature extractors. Furthermore, extracted feature vectors  $eV_1$  from Nasnetmobile and  $eV_2$  from Stacked Auto-encoder are reduced by applying feature optimization techniques.  $eV_1$  is fed to the Marine Predator Algorithm (MPA) [28] while  $eV_2$  is given as input to the Slime Mould Algorithm (SMA) [29] to extract selected features vectors  $S(eV_1)$  and  $S(eV_2)$ , respectively. Selected feature vectors  $S(eV_1)$  and  $S(eV_2)$  are fused. Moreover, artificial neural networks are used as classifiers to achieve results. Fig. 1 shows the proposed methodology used in this paper.

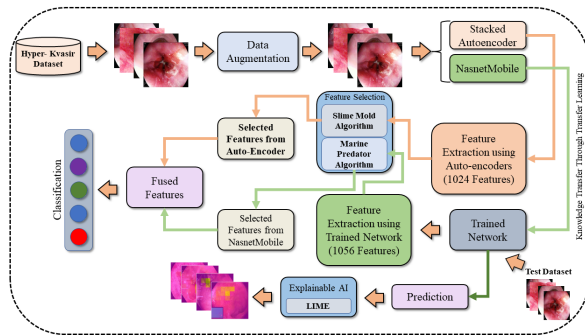


Figure 1: Proposed methodology of stomach cancer classification and polyp detection

#### 3.1 Dataset Description

The Hyper Kvasir dataset used in this study is a public dataset collected from Baerum Hospital

in Norway [23]. The dataset contains 10662 gastrointestinal endoscopy images categorized into 23 classes. Among twenty-three classes, sixteen belong to the lower gastrointestinal area, while seven are related to the upper gastrointestinal segment. Table 1 describes the data misbalancing problem, as some of the classes have very few numbers of images. To nullify the issue, data augmentation techniques are applied. Fig. 2 shows the sample images for each class.

Table 1: Classes of Hyper Kvasir dataset and number of images in each class

Class	Number of Images
barretts	41
barrettes-short-segment	53
bbps-0-1	646
bbps-2-3	1148
cecum	1009
dyed-lifted-polyps	1002
dyed-resection-margins	989
esophagitis-a	403
esophagitis-b-d	260
hemorrhoids	6
ileum	9
impacted-stool	131
polyps	1028
pylorus	999
retroflex-rectum	391
retroflex-stomach	764
ulcerative-colitis-grade-0-1	35
ulcerative-colitis-grade-1	201
ulcerative-colitis-grade-1-2	11
ulcerative-colitis-grade-2	443



ulcerative-colitis-grade-2-3	28
ulcerative-colitis-grade-3	133
z-line	932

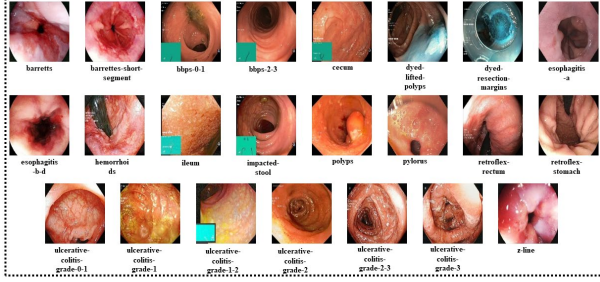


Figure 2: Sample images of each class of the Hyper Kvasir dataset

### 3.2 Proposed Contrast Enhancement

Data is augmented by applying three different image enhancement techniques, as these techniques change spatial properties but do not affect the image orientation. Brightness Preserving Histogram Equalization (BPHE) [30] and Dualistic Histogram Equalization (DHE) [31] are used in preprocessing. Moreover, the haze removal technique is also used in image augmentation procedures.

BPHE is a method employed in image processing to enhance an image's visual quality by improving its contrast. This approach involves adjusting the distribution of intensity levels to generate a more uniform histogram. Unlike conventional histogram equalization techniques, brightness-preserving histogram equalization considers both bright and dark regions in an image. It independently adjusts the histograms of each region to retain the details in both bright and dark areas while enhancing overall contrast. This technique is particularly useful in applications such as medical imaging, where preserving the details in both bright and dark regions is crucial. The input image is divided into two sub-parts; the first consists of pixels with low contrast values, while the second consists of pixels with high contrast values. Mathematically it is denoted as:

$$M_{Inp} = (M_{lower}) \cup (M_{higher})$$

$$(1)$$

Here,

$$(M_{lower}) = \{M(j, k) | M(j, k) \leq M_{mean}, \forall M(j, k) \in M\} \quad (2)$$

and

$$(M_{higher}) = \{M(j, k) | M(j, k) \leq M_{mean}, \forall M(j, k) \in M\} \quad (3)$$

Also, a function of probabilistic density for both sub-parts is derived as:

$$(Dens_{prob})_l(M_t) = \frac{O_l^t}{O_l}, \text{ Where } t = 0, 1, 2, \dots, N \quad (4)$$

and

$$(Dens_{prob})_h(M_t) = \frac{O_h^t}{O_h}, \text{ Where } t = N + 1, N + 2, \dots, N - K \quad (5)$$

Where,  $O_l^t$  and  $O_h^t$  are the number of  $M_p$  in  $(M_{lower})$  and  $(M_{higher})$ , respectively. Also, cumulative density functions for sub-parts are derived as:

$$Func_{dl}(M_t) = \sum_{k=0}^t (Dens_{prob})_l(M_k) \quad (6)$$

and

$$Func_{dh}(M_t) = \sum_{k=N+1}^p (Dens_{prob})_h(M_k) \quad (7)$$

The transform function for sub-parts is as follows:

$$Trf_l(M_t) = M_0(M_N - M_0)Func_{dl}(M_t) \quad (8)$$

and

$$Trf_h(M_t) = M_0(M_N - M_0)Func_{dh}(M_t) \quad (9)$$

The final image having an equalized histogram with preserved brightness can be obtained by combining both equations, that is:

$$Img_{BPHE} = Trf_l(M_t) \cup Trf_h(M_t) \quad (10)$$

In the above equation,  $Img_{BPHE}$  is the Brightness Preserved Histogram Equalized image.

DSIHE is an image enhancement approach that increases an image's contrast by separating it into two sub-images depending on a threshold

value and then applying histogram equalization independently to each sub-image. The significance of DSIHE resides in its capacity to improve the contrast of images containing dark and light areas. Contrast enhancement is done worldwide using classic histogram equalization, which can result in over-enhancing bright parts and under-enhancement of dark regions in an image. DSIHE tackles this issue by separating the picture into two sub-images based on a threshold value that distinguishes between the light and dark regions. Afterward, histogram equalization is applied separately to each sub-image, which helps to achieve an equilibrium across the two regions' contrast enhancement. It has been demonstrated that the DSIHE technique enhances the aesthetic quality of medical images. It is an easy, computationally efficient, and straightforward strategy to implement in image processing systems.

Let  $M_{Inp}$  is an input image that is given to apply DSIHE, and the grey level of that image is  $M_{Inp} = M_{grey}$ . Sub-images are denoted by  $M_{S1}$  and  $M_{S2}$ . The center pixel index is denoted by  $C_{px}$ .

$$M_{Inp} = M_{S1} \cup M_{S2} \quad (11)$$

$$M_{S1} = \{M_{Inp}(i, j) | M_{Inp}(i, j) < M_{grey}, \forall M_{Inp}(i, j) \in M_{Inp}\} \quad (12)$$

$$M_{S2} = \{M_{Inp}(i, j) | M_{Inp}(i, j) \geq M_{grey}, \forall M_{Inp}(i, j) \in M_{Inp}\} \quad (13)$$

Upper transformation is used for less bright images.

$$M_{S1} = \{m_0, m_1, m_2 \dots \dots m_{M_{grey}-1}\} \quad (14)$$

$$M_{S2} = \{M_{grey}, M_{grey+1}, \dots \dots M_{S1-1}\} \quad (15)$$

Aggregation of the grey-level original image is as follows:

$$\{A_{g0}, A_{g1}, \dots \dots A_{M_{grey}-1}\} \quad (16)$$

$$\{A_{M_{grey}}, A_{M_{grey}+1}, \dots \dots A_{m_{M_{grey}-1}}\} \quad (17)$$

The aggregated PDF for grey levels of the original image will be:

$$\{P_{d0}, P_{d1}, P_{d2} \dots \dots P_{d_{grey-1}}\} \quad (18)$$

$$\{P_{d_{grey}}, P_{d_{grey}+1}, \dots \dots P_{d_{grey-1}}\} \quad (19)$$

Suppose

$$p_d = \sum_{i=0}^{grey-1} A_{gi} \quad (20)$$

$$p_d = \sum_{i=grey}^{d_{grey}-1} A_{gi} \quad (21)$$

$$A_g(M_{S1}) = \frac{p_i}{A_{S1}}, i = 0, 1, 2, \dots \dots, grey - 1 \quad (22)$$

$$A_g(M_{S2}) = \frac{p_i}{A_{S2}}, i = grey, grey + 1, grey + 2, \dots \dots, m_{M_{grey}-1} \quad (23)$$

To evaluate CDF:

$$CD_{S1}(M_{inp_k}) = \frac{1}{A_g} \sum_{i=0}^{grey-1} A_{gi} \quad (24)$$

$$CD_{S2}(M_{inp_k}) = \frac{1}{A_g} \sum_{i=grey}^{d_{grey}-1} A_{gi} \quad (25)$$

For both sub-images, the transformation function is given by:

$$F_{trans_{S1}}(M_{inp}) = M_{grey_0} + (M_{grey-1} - m_0) \times CD_{S1}(M_{inp_k}) \quad (26)$$

$$F_{trans_{S2}}(M_{inp}) = M_{grey} + (M_{grey-1} - M_{grey}) \times CD_{S2}(M_{inp_k}) \quad (27)$$

The output image is mathematically denoted by:

$$M_{out} = F_{trans_{S1}}(M_{inp}) \cup F_{trans_{S2}}(M_{inp}) \quad (28)$$

### 3.3 Novelty: Designed CNN Model

Feature extraction is extracting a subset of relevant features from raw data useful for solving a particular machine-learning task [32]. In deep learning, feature extraction involves taking a raw input, such as an image or audio signal, and automatically extracting relevant features or

patterns using a series of mathematical transformations. Deep learning relies on feature retrieval to help the network concentrate on the essential data and simplify the input, making it simpler to train and more accurate. In some cases, feature extraction can also help to reduce overfitting and improve generalization performance. In many deep learning applications, the network performs feature extraction automatically, typically using convolutional layers for image processing or recurrent layers for natural language processing. However, in some cases, manual feature extraction may be necessary, particularly when working with smaller datasets or trying to achieve high levels of accuracy on a specific task. In this study, two feature extractors are used to extract features. Stacked Auto-Encoder and Nasnetmobile are two frameworks that are used to extract features.

CNNs have become a popular tool in the field of medical image processing. A neural network can be classified as a CNN if it contains at least one layer that performs convolution operations. During a convolution operation, a filter with multiple parameters of a specific size is applied to an input image using a sliding window approach. The resulting image is then passed on to the next layer for further processing. This operation can be represented mathematically as follows:

$$M_{out}\{Horz_{out} \times Vert_{out}\} = (M_{inp} * Fil_{op}) \quad (29)$$

Above,  $M_{out}$  is the output matrix having  $Horz_{out}$  and  $Vert_{out}$  rows and columns, respectively. Furthermore, the rectified linear unit function is applied to obtain the negative feature's value as zero, which can be represented in the equation below:

$$Act_{ReLU} = Maximum\_of(0, a), a \in M_{out} \quad (30)$$

Furthermore, a pooling operation reduces computational complexity and improves processing time. This operation involves extracting the maximum or average values from a specific region and replacing them with the central input value. A fully connected layer then flattens the features to produce a one-dimensional vector. Mathematically, this can be represented as:

$$(Vc_{flat})_0^{out} = M_{out}\{Horz_{out} \times Vert_{out}\} \quad (31)$$

$$(Vc_{flat})_i^{in} = (Vc_{flat})_{i-1}^{out} * M_i + Vert_i \quad (32)$$

$$(Vc_{flat})_i^{out} = \Delta_i \left( (Vc_{flat})_i^{in} \right) \quad (33)$$

Where,  $(Vc_{flat})_i^{out}$  is flattened layer vector,  $\Delta$  represents the activation function, and  $i$  is the layer on which the operation is performed. SoftMax is implemented to achieve probability for the feature to obtain the classification results that are shown as:

$$SOFTMAX \left( (Vc_{flat})_i^{out} \right) = \frac{\exp((Vc_{flat})_i^{out})}{\sum_k (Vc_{flat})_k^{out}} \quad (34)$$

### 3.3.1 Stacked Auto-Encoder

A type of neural network known as a stacked autoencoder utilizes unsupervised learning to develop a condensed representation of input data. The architecture consists of multiple layers, each learning a compressed representation called a "hidden layer" of the input data. The output of one layer is used as input for the subsequent layer, and the final output layer generates the reconstructed data. To create a deeper architecture capable of learning more complex and abstract representations, hidden layers are added to the network. During training, the difference between the input and the reconstructed output data, known as the reconstruction error, is minimized using backpropagation to adjust the neural network's weights[33]. Stacked autoencoders are used in various applications, including speech and image recognition, anomaly detection, and data compression.

Let  $X_{inp}$  be the input data and  $Y_{out}$  be the reconstructed data. Let the stacked autoencoder have  $L_{last}$  layers, with the hidden layers denoted as  $h\_1_{layer}, h\_2_{layer}, \dots, h\_L_{last} - 1$ . The output layer is denoted as  $h\_L_{last}$ . A transformation function can represent each layer of the stacked autoencoder  $f_{trans}$  that maps the input to the output. The transformation function for the  $n$ -th layer is denoted as  $f_{transl}$ . The input data is fed into the first layer, which learns a compressed input representation. The output of the first layer is then

passed as input to the next layer, which learns a compressed representation of the output from the first layer. This process continues until the final layer produces the reconstructed data  $Y_{out}$ . The compressed representation learned by each hidden layer can be represented as follows:

$$h_k = f_{trans}(X_{inp}W_k + b_k) \quad (35)$$

where  $h_k$  is the output of the  $k$ th hidden layer,  $W_k$  is the weight matrix connecting the input to the  $k$ th hidden layer, and  $b_k$  is the bias vector for the  $k$ -th hidden layer. The reconstructed output  $Y_{out}$  can be calculated by passing the compressed representation of the input through the decoder network, which is essentially the reverse of the encoder network:

$$Y_{out} = f_{trans}(h_{L_{last}}W_{last} + b_{last}) \quad (36)$$

where  $W_{last}$  is the weight matrix connecting the last hidden layer to the output layer, and  $b_{last}$  is the bias vector for the output layer. Minimizing reconstruction error between input and output trains the stacked autoencoder. Features vector named as  $Feat\_AE_{vec}$  is extracted through the Stacked Auto-Encoder that consists of 1024 features.

### 3.3.2 Feature Extraction using Proposed CNN

Nasnetmobile is a pre-trained neural network model [34] that has been trained using transfer learning. Transfer learning is the method that involves the knowledge transfer learned from a pre-trained model on a new task. In the case of Nasnetmobile, it has been trained on the ImageNet dataset, divided into 70% training and 30% testing images. To adapt the pre-trained model for a new task, transfer learning principles shown in Fig. 3 are used to refine the model. However, since the pre-trained model has been trained on a subset of classes, it is not directly applicable to a medical image classification task. Therefore, the network needs to be trained on a new Hyper-Kvasir dataset. To train the network on the Hyper-Kvasir dataset, the classification layer, soft-max layer, and last fully connected layer of the Nasnetmobile model are replaced with new layers called "new\_classification," "new\_softmax," and "new\_Prediction," respectively. This allows the model to learn to classify medical images using

the features extracted from the original pre-trained model. Furthermore, features are extracted through a trained network and obtained deep feature vectors.  $Feat\_NNMobile_{vec}$  containing 1056 features.

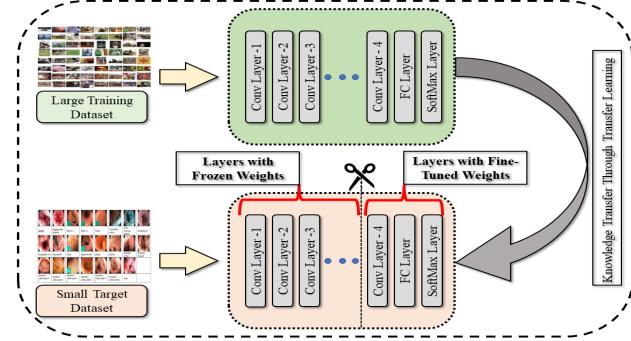


Figure 3: Generalization through transfer learning technique

### 3.4 Novelty: Proposed Features Selection

Feature selection is the operation of identifying a subset of appropriate features from a dataset's larger set of features [35]. Feature selection improves model performance and data interpretation and reduces computational resources. Two feature selection algorithms are used to tackle the curse of dimensionality. Slime Mould Algorithm (SMA) is used to select important features in vector.  $S(Feat\_AE_{vec})$  from  $Feat\_AE_{vec}$  extracted through the Stacked Auto-Encoder while the Marine Predator Algorithm (MPA) is used to extract selected features vector  $S(Feat\_NNMobile_{vec})$  from  $Feat\_NNMobile_{vec}$  that is obtained through Nasnetmobile.  $S(Feat\_AE_{vec})$  consists of 535 features where as  $S(Feat\_NNMobile_{vec})$  has 366 features.

The Slime Mold Algorithm is a feature selection technique influenced by nature and centered around slime mold behavior. The method employs a system of artificial particles that interact with one another to identify the ideal solution. SMA approaches the food according to the strength of the odor the food source spreads. The following equations describe the behavior of the method for slime mold:

$$F_s((i+1)) = \begin{cases} \overrightarrow{F_a(i)} + \overrightarrow{pb} \cdot (\overrightarrow{X} \cdot \overrightarrow{F_a(i)} - (\overrightarrow{F_b(i)})), & s < q \\ \overrightarrow{pc} \cdot \overrightarrow{F(i)}, & s \geq q \end{cases} \quad (37)$$

Here,  $\overrightarrow{pb}$  is the parameter ranging from  $-G_p$  to  $G_p$ . Also,  $\overrightarrow{pc}$  is the parameter that goes from zero to one in descending order.  $i$  represents the iteration number. Moreover,  $\overrightarrow{F_a}$  shows the location of the source that has the highest odour.  $\vec{F}$  is the location where slime mould is located.  $\overrightarrow{F_a}$  and  $\overrightarrow{F_b}$  is randomly selected food sources at the initial time. Furthermore,  $\vec{X}$  is the weight of mould, and  $q$  is derived as:

$$q = \tan|T(j) - EH| \quad (38)$$

Where,  $j \in 1, 2, 3, \dots, n$  and  $T(j)$  represents the fitness for  $\vec{F}$ .  $G_p$  is represented as:

$$G_p = \tanh^{-1} \left( -\left( \frac{i}{\max of i} \right) + 1 \right) \quad (39)$$

The weight of the mould is calculated mathematically as:

$$\vec{X} = \begin{cases} 1 + q \cdot \log \left( \frac{bd - T(j)}{bd - \omega d} + 1 \right), & \text{For some condition} \\ 1 - q \cdot \log \left( \frac{bd - T(j)}{bd - \omega d} + 1 \right), & \text{Other than Condition} \end{cases} \quad (40)$$

$q$  is the random number from the range zero to one.  $bd$  is the best fit for the current iteration, as  $\omega d$  is the worst fit in the current iteration. The position updating is derived as:

$$F_s^* = \begin{cases} rand \cdot (bound_{upper} - bound_{lower} + bound_{lower}), & rand < y \\ \overrightarrow{F_a(i)} + \overrightarrow{pb} \cdot (\vec{X} \cdot \overrightarrow{F_a(i)} - (\overrightarrow{F_b(i)})), \\ \overrightarrow{pc} \cdot \overrightarrow{F(i)}, \end{cases} \quad (41)$$

The Marine Predator Optimization Algorithm (MPO) is a metaheuristic optimization algorithm based on the foraging strategies of aquatic predators. MPO is an algorithm replicating the searching and preying behavior of deep-sea predatory animals such as sharks, orcas, and other ocean animals. Like most metaheuristic algorithms, MPA is a population-based approach in which the baseline answer is dispersed equally over the search area, as in the first experiment. Mathematically it is dented by:

$$A_0 = A_0 + rand (A_{max} - A_{min}) \quad (42)$$

Here,  $A_{min}$  is the lower bound, whereas  $A_{max}$  is the upper bound for variables.  $Rand$  stands for a randomly chosen vector ranging from zero to one. Based on the notion of survival of the fittest, it is considered that the most efficient hunters in nature are the strongest predators. As a result, the top predator is regarded as the most efficient means of generating an elite matrix. These elite matrices are meant to detect and track prey by leveraging their location data. Each element in the elite matrix denotes predators in a position to search for food. The second matrix is called the prey matrix, where each element represents the prey also looking for food. Both matrices have  $r \times c$  dimensions where  $r$  shows the number of searching agents, whereas  $c$  represents the number of dimensions. At each iteration, the fittest predator substitutes the previous fittest predator.

There are three phases that MPA contains. Phase one is considered when a predator is moving faster than prey, and velocity is ( $V \geq 10$ ). In this scenario, the best possible solution could be to stop the updating positions of predators. Mathematically it can be represented as:

$$\begin{aligned} & \text{if } iteration < \frac{1}{3} \text{ of } Max_{Iteration} \\ & \overrightarrow{Stp_{itr}} = \overrightarrow{Rand_{N-dist}} \otimes (\overrightarrow{elite_{itr}} - \overrightarrow{Rand_{N-dist}} \oplus \overrightarrow{Pr_{itr}}), itr = 1, \dots, N_{total} \\ & \overrightarrow{Pr_{itr}} = \overrightarrow{Pr_{itr}} + Const. \overrightarrow{Rand_{N-dist}} \otimes \overrightarrow{Stp_{itr}} \end{aligned} \quad (43)$$

In this scenario,  $\overrightarrow{Rand_{N-dist}}$  represents a normal distribution-generated vector of random integers  $s \leq q$  simulating Brownian motion. The symbol " $\otimes$ " represents entry-wise multiplication. Prey is multiplied by the vector.  $\overrightarrow{Rand_{N-dist}}$  to imitate its movement.  $\overrightarrow{Rand_{N-dist}}$  is a vector of uniform random integers ranging from 0 to 1, whereas  $Const$  is a constant with a value of 0.5. This situation happens during the first one-third of iterations when the size of each step is large due to a greater capacity for exploration.  $iteration$  denotes the current iteration, and  $Max_{Iteration}$  is the total number of possible iterations.

Phase two is considered as unit velocity ratio when both prey and the predators have the same velocity, is ( $V \approx 10$ ). In this phase, the prey is in exploitation mode and levy motion while the

predator is in exploration mode with Brownian motion. For half of the population, this could be denoted by:

$$\begin{aligned} & \text{if } iteration < \frac{2}{3} \text{ of } Max_{Iteration} \\ & \overrightarrow{Stp_{itr}} = \overrightarrow{Rand_{L-dist}} \otimes (\overrightarrow{elite_{itr}} - \\ & \overrightarrow{Rand_{L-dist}} \oplus \overrightarrow{Pr_{itr}}), itr = 1, \dots, N_{total}/2 \\ & \overrightarrow{Pr_{itr}} = \overrightarrow{Pr_{itr}} + Const. \overrightarrow{Rand_{N-dist}} \otimes \overrightarrow{Stp_{itr}} \end{aligned} \quad (44)$$

Above,  $\overrightarrow{Rand_{L-dist}}$  denotes the random number based on Levy distribution. For another half of the population, it is represented as:

$$\begin{aligned} & \overrightarrow{Stp_{itr}} = \overrightarrow{Rand_{N-dist}} \otimes (\overrightarrow{Rand_{N-dist}} \otimes \\ & \overrightarrow{elite_{itr}} - \overrightarrow{Rand_{N-dist}} \oplus \overrightarrow{Pr_{itr}}), itr = \\ & N_{total}/2, \dots, N_{total} \\ & \overrightarrow{Pr_{itr}} = \overrightarrow{elite_{itr}} + Const. \overrightarrow{Adp_{prm}} \otimes \overrightarrow{Stp_{itr}} \end{aligned} \quad (45)$$

Whereas  $Adp_{prm} = \left(1 - \frac{iteration}{Max_{iteration}}\right)^{\left(2 \left(\frac{iteration}{Max_{iteration}}\right)\right)}$  is adaptive control parameter to control the step size.

In phase three prey has low velocity as compared to predator's velocity. In low ratio velocity, the value will be ( $V = 0.1$ ). In this scenario, the best motion for the predator will be the Levy motion, as shown in the equation (46).

$$\begin{aligned} & \text{if } iteration < \frac{2}{3} \text{ of } Max_{Iteration} \\ & \overrightarrow{Stp_{itr}} = \overrightarrow{Rand_{L-dist}} \otimes (\overrightarrow{Rand_{L-dist}} \otimes \\ & \overrightarrow{elite_{itr}} - \overrightarrow{Rand_{L-dist}} \oplus \overrightarrow{Pr_{itr}}), itr = \\ & 1, \dots, N_{total}/2 \\ & \overrightarrow{Pr_{itr}} = \overrightarrow{elite_{itr}} + Const. \overrightarrow{Rand_{N-dist}} \otimes \overrightarrow{Stp_{itr}} \end{aligned} \quad (46)$$

The reason the change in marine predators' behavior is environmental changes inserted in the algorithm as eddy formation and Fish Aggregating Device (FAD) manipulation. These two effects are denoted by:

$$\begin{aligned} & \overrightarrow{Pr_{itr}} = \\ & \left( \overrightarrow{Pr_{itr}} + \overrightarrow{Adp_{prm}} [\overrightarrow{bd_{up}} + \overrightarrow{Rand} \otimes (\overrightarrow{bd_{up}} - \overrightarrow{bd_{lr}})] \otimes \overrightarrow{Vec_{bin}}, \text{ if } u_{rand} \leq FADs \right. \\ & \left. \overrightarrow{Pr_{itr}} + [FADs(1 - u_{rand}) + 1] (\overrightarrow{Pr_{u_{rand1}}} - \overrightarrow{Pr_{u_{rand2}}}), \text{ if } u_{rand} > FADs \right. \end{aligned} \quad (47)$$

Here,  $FADs = 0.20$  represents the likelihood of FADs' influence in the optimization procedure. A binary vector U is created by randomly creating

a vector in the interval  $[0,1]$  and replacing its elements with zero if they are less than 0.2 and with one if they are more than 0.2. The subscript r denotes a uniformly random number in the interval  $[0,1]$ . The vectors  $\overrightarrow{bd_{up}}$  and  $\overrightarrow{bd_{lr}}$  contain the minimum and maximum dimensions.  $u_{rand1}$  and  $u_{rand2}$  denote the random indices of the prey matrix.

### 3.5 Novelty: Proposed Feature Fusion

The significance of feature fusion resides in its capacity to extract more meaningful information from numerous sources, which can lead to improved accuracy in classification, identification, and prediction [36]. Feature fusion can increase the resilience and reliability of machine learning systems, especially in cases when data is few or noisy, by merging complementary information from many sources. As stated before, two feature vectors,  $S(Feat\_AE_{vec})$  and  $S(Feat\_NNMobile_{vec})$ , are retrieved from both networks utilized in this process; hence, it is important to merge both vectors to create a larger, more informative feature vector. A correlation extended serial technique is utilized to combine both vectors, which can be mathematically represented as follows:

$$Co_{rel} = \frac{\sum (Rw_i - \overline{Rw})(Xt_j - \overline{Xt})}{\sqrt{\sum (Rw_i - \overline{Rw})^2 \sum Xt_j - \overline{Xt}}} \quad (48)$$

With this procedure, the features with a positive correlation (+1) are chosen into a new vector labeled.  $Vec_3$  and the features with a correlation value of 0 or -1 are added to  $Vec_4$ . Then, the mean value of  $Vec_4$  is calculated as follows:

$$Co_{rel}T = \begin{cases} Vec_{upd}, & Vec_4 \geq 0 \\ Ignore\_feat, & Vec_4 < 0 \end{cases} \quad (49)$$

Both vectors  $Vec_{upd}$  and  $Vec_4$  are fused using the following formulation.

$$\overrightarrow{Vec_{Fused}} = \begin{pmatrix} \overrightarrow{Vec_{upd}} \\ \overrightarrow{Vec_3} \end{pmatrix} \quad (50)$$

The final fused vector  $Vec_{Fused}$  has 901 features.



#### 4 Results and Discussion

The hyper-Kvasir dataset is used for results and analysis purposes. The dataset contains 10662 images categorized into twenty-three classes. Data is highly imbalanced, so to cater to this issue, data is augmented. The augmented dataset contains 24000 training images, while 520 are obtained for testing. The implementation uses a system with a core i7 Quad-core processor with 16 GB of RAM. Also, the system contains a graphics card with 4GB of VRAM. MATLAB R2021a is used to achieve results.

##### 4.1 Numerical Results

Results are shown in tabular and graphical form. Table 2 represents results for extracted features through Nasnetmobile that are given as input to classifiers. The analysis shows that Wide Neural Network (WNN) has given the best overall accuracy of 93.90. percent, while Narrow Neural Networks, Bilayered Neural Networks, and Trilayered Neural Networks have the lowest accuracy of 93.10 percent. Time taken by WNN is also the highest among all other classifiers, while the lowest time cost is for Narrow Neural Networks. The confusion matrix for WNN is shown in Fig. 4.

Table 2: Performance for ANN classifiers using NasNet Mobile features (1056 Features)

Classifier	Accuracy	Precision	Recall	F1 Score	Time
Narrow Neural Network	93.10	93.09	92.78	92.93	402.6
Medium Neural Network	93.40	93.60	93.47	93.53	457.25
<b>Wide Neural Network</b>	<b>93.90</b>	<b>93.88</b>	<b>93.79</b>	<b>93.84</b>	<b>866.42</b>
Bilayered Neural Network	93.10	92.97	92.89	92.93	418.90

k					
Trilayered Neural Network	93.10	93.02	92.88	92.95	411.31

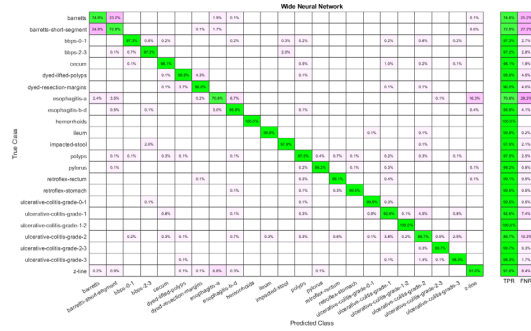


Figure 4: Confusion matrix for WNN using NasNet Mobile features

Similarly, Table 3 shows results obtained by feeding the features extracted by implementing Stacked Auto-Encoders to classifiers. Analysis shows that WNN has the best performance with 80.50 percent accuracy, yet time cost is also highest in the case of WNN and lowest for Narrow Neural Networks. Moreover, the lowest accuracy is achieved by implementing a Narrow Neural Network. The confusion matrix for WNN is shown in Fig. 5.

Table 3: Performance for ANN classifiers using autoencoder features (1024 Features)

Classifier	Accuracy	Precision	Recall	F1 Score	Time
Narrow Neural Network	71.69	68.50	69.03	68.76	382.67
Medium Neural Network	76.70	75.23	75.44	75.34	411.22
<b>Wide Neural Network</b>	<b>80.50</b>	<b>79.74</b>	<b>79.92</b>	<b>79.83</b>	<b>796.32</b>
Bilayered Neural Network	72.00	68.67	69.4	69.0	402.

ed Neural Network			6	7	54
Trilayered Neural Network	71.70	68.43	69.10	68.77	392.06

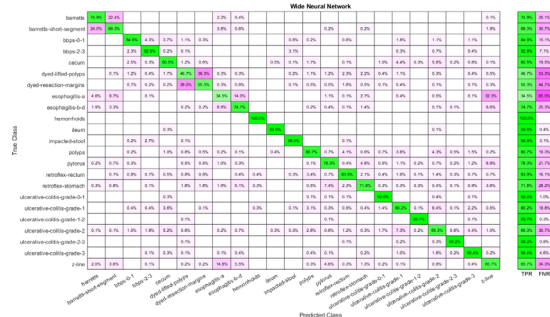


Figure 5: Confusion matrix for WNN using auto-encoder features

Feature selection has given reduced features from the feature vector extracted through Nasnetmobile. Table 4 shows the results for selected features using the Marine Predator Algorithm (MPA). Selected features are given to the classifiers to obtain results. Analysis shows that WNN has the highest accuracy, 93.40, and the highest time cost. Furthermore, the lowest accuracy is obtained through a Trilayered Neural Network. A Narrow Neural Network has given the best time cost among all classifiers. The confusion matrix for WNN is shown in Fig. 6.

Table 4: Performance for ANN classifiers using selected NasNet Mobile features (366 Features)

Classifier	Accuracy	Precision	Recall	F1 Score	Time
Narrow Neural Network	92.40	92.22	92.06	92.14	260.82
Medium Neural Network	93.00	92.99	92.92	92.95	295.29

Wide Neural Network	93.40	93.54	93.45	93.49	511.87
Bilayered Neural Network	92.40	92.14	92.07	92.12	279.42
Trilayered Neural Network	92.30	92.10	91.95	92.02	285.61

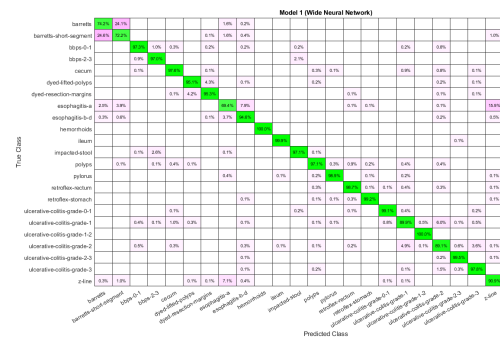


Figure 6: Confusion matrix for WNN using NasNet Mobile selected features.

Table 5 shows the results achieved using selected features from Stacked Auto-Encoder. The features are selected using the Slime Mold Algorithm. WNN has the best performance as the accuracy achieved is 78.40 percent. Also, the time cost is highest for WNN and lowest for Narrow Neural Networks. In addition, Trilayered Neural Network has given the lowest accuracy. The confusion matrix for WNN is described in Fig.7.

Table 5: Performance for ANN classifiers autoencoder selected features (535 Features)

Classifier	Accuracy	Precision	Recall	F1 Score	Time
Narrow Neural Network	69.90	65.91	66.88	66.39	348.91
Medium Neural Network	73.60	71.31	71.67	71.49	402.3

Network					
<b>Wide Neural Network</b>	<b>78.40</b>	<b>76.75</b>	<b>76.97</b>	<b>76.86</b>	<b>651.80</b>
Bilayered Neural Network	69.70	65.38	66.51	65.94	359.24
Trilayered Neural Network	67.90	63.47	64.61	64.04	366.21

m Neural Network			4	1	
<b>Wide Neural Network</b>	<b>93.80</b>	<b>93.81</b>	<b>93.73</b>	<b>93.77</b>	<b>772.93</b>
Bilayered Neural Network	92.30	92.27	92.22	92.24	398.12
Trilayered Neural Network	92.40	92.26	92.20	92.23	372.53

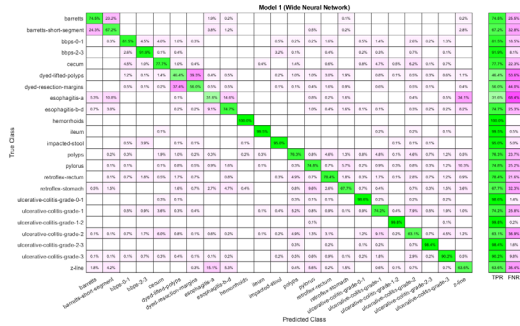


Figure 7: Confusion matrix for WNN using auto-encoder selected features

The best performance obtained using fused features is shown in Table 6. Features are given to ANNs, and analyzed the results. Analysis shows that the highest accuracy of 93.60 is achieved through WNN. Again, the time cost is high in the case of WNN, yet it is best for Narrow Neural Networks. Also, the lowest accuracy is obtained through a Narrow Neural Network. The confusion matrix for WNN is depicted in Fig.8.

Table 6: Performance for ANN classifiers using fused features (901 features)

Classifier	Accuracy	Precision	Recall	F1 Score	Time
Narrow Neural Network	92.20	92.19	92.12	92.15	376.15
Medium Neural Network	93.10	93.28	93.1	93.2	386

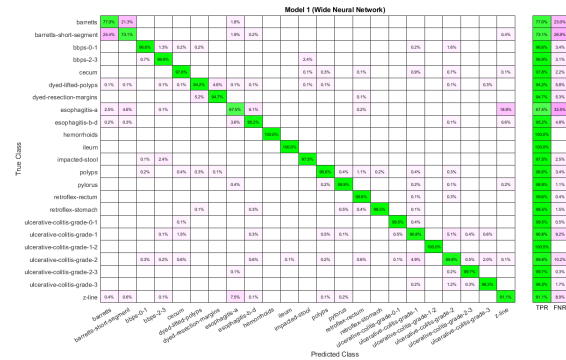


Figure 8: Confusion matrix for WNN using fused features.

## 4.2 Graphical Results

This section shows the graphical representation of the results. Fig. 10 shows the bar chart for all classifiers using the proposed fusion approach. In this figure, each classifier's accuracy is plotted with different colors, and Wide Neural Network shows the best accuracy of 93.8%, which is improved than the other classifiers. Fig. 11 shows the bar chart for the time cost for all classifiers after employing the final step of the proposed approach. Wide Neural Network (WNN) consumed the highest time of 772.93 (sec), whereas the trilayered neural network spent a minimum time of 372.53 (sec). Based on Figures 10 and 11, it is clearly observed that the wide neural network gives better accuracy but

consumes more time due to additional hidden layers. Fig. 12 shows the time-based comparison of the proposed method. This figure shows that the time is significantly reduced after employing the feature selection step; however, a little increase occurs when the fusion step is performed. Overall, it is observed that the reduction of features impacts the computational time, which is a strength of this work.

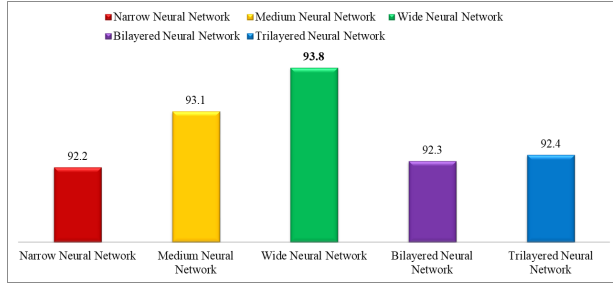


Figure 9: Accuracy bar for all selected classifier using proposed method.

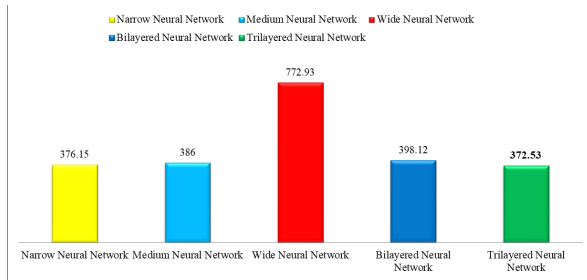


Figure 10: Time bar for classifiers used in the proposed methodology.

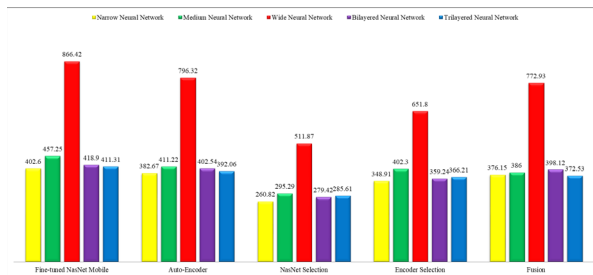


Figure 11: Overall time based comparison among all classifiers using proposed method.

A detailed comparison is also conducted among all classifiers of the middle steps employed in the proposed method. Figure 13 shows the insight view of this comparison. This figure shows that the original accuracy of the fine-tuned model NasNet Mobile is better, and the maximum is 93.9%; however, this experiment consumes more

time, as plotted in Figure 12. After the selection process, the accuracy is slightly reduced, but the time is significantly dropped. After the fusion process, it is clearly noted that the difference in the classification accuracy of the wide neural network is just 0.1% which is almost the same. Still, time is significantly reduced, which is a strength of this work.

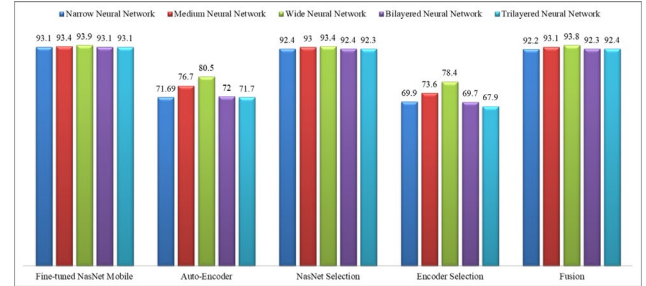


Figure 13: Accuracy comparison of all classifiers using all middle steps of the proposed method.

**LIME based Visualization:** Local Interpretable Model-Agnostic Explanations (LIME) [37] is a well-known technique for explainable artificial intelligence (XAI). It is a model-independent technique that may be used to explain the predictions of any machine learning algorithm, including sophisticated models like deep neural networks. LIME aims to produce locally interpretable models that approach the predictions of the original machine learning model in a limited part of the input space. Local models are simpler and easier to comprehend than the original model and can be used to explain specific predictions. The LIME approach generates a large number of perturbed versions of the input data and trains a local model on each disturbed version. Local models are trained to predict the output of the original model for each perturbed version and are then weighted according to their performance and resemblance to the original input. The final explanation offered by LIME is a mix of the weights of the local models and the most significant characteristics of each local model. An explanation can be offered to the user in the form of a heatmap or other visualization, as shown in Figure 14, indicating which input data characteristics were most influential in forming the prediction.

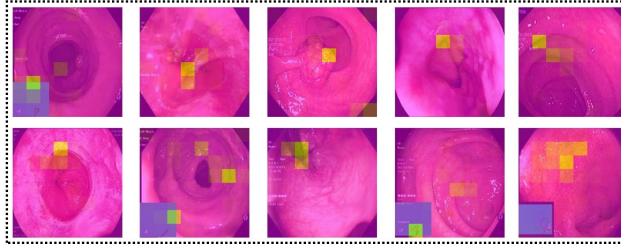


Figure 14: Explanation of network's predictions using LIME

Figure 15 shows the results of the fine-tuned Nasnetmobile deep model employed for infected region segmentation. The segmentation process employs the polyp images with corresponding ground truth images. This fine-tuned model is trained with static hyperparameters by employing original and ground truth images. After that, testing is performed to visualize a few images in binary form, as presented in Figure 15. For the segmentation process, the weights of the second convolutional layers have been plotted and then converted into binary form.

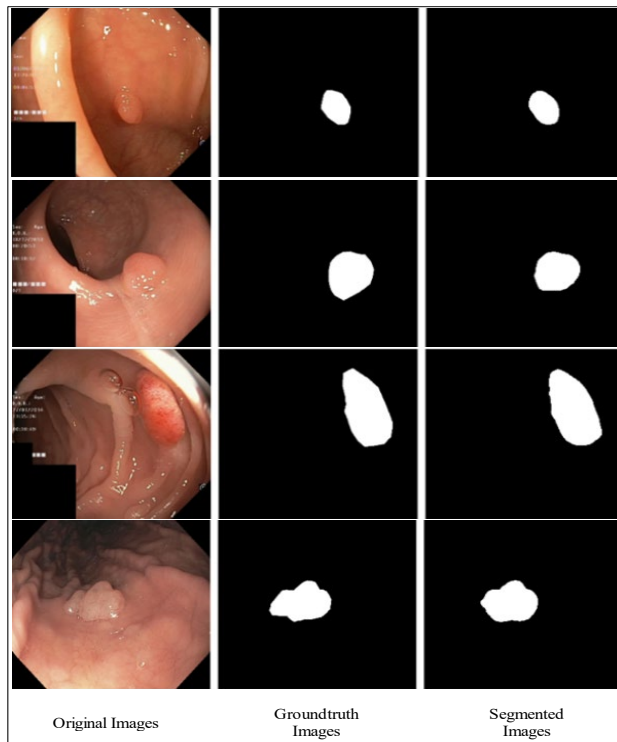


Figure 18: Proposed infection segmentation using fine-tuned NasnetMobile deep model

Table 8 compares the results achieved in this article with recent state-of-the-art works. [38] used self-supervised learning to classify the hyperKvasir dataset. The authors used six classes

and achieved the highest accuracy of 87.45. Moreover, [27] used the hyper Kvasir dataset to classify the gastrointestinal tract and obtained 73.66 accuracy. In the study, the authors only used fourteen classes. In addition, [23] achieved 63 percent accuracy for macro and used all 23 classes. It is clear that the proposed method has outperformed the state-of-the-art methodologies in recent years and achieved the best accuracy of 93.80 percent.

Table 7: Comparison of the proposed framework accuracy with state-of-the-art (SOTA) techniques

Reference	Dataset	Number of Classes	Year	Accuracy (%)
[38]	Hyper - Kvasir	6	2023	87.45
[26]	Kvasir	5	2021	97.00
[27]	Hyper - Kvasir	14	2020	73.66
[23]	Hyper - Kvasir	23	2020	63.00
[39]	Hyper - Kvasir	23	2023	87.1
<b>Proposed</b>	<b>Hyper - Kvasir</b>	<b>23</b>	<b>-</b>	<b>93.80</b>

## 5 Conclusion

Gastrointestinal tract cancer is one of the most severe cancers in the world. Deep learning models are used to diagnose gastrointestinal cancer. The proposed model uses Nasnetmobile and Auto-Encoder to extract deep features and is used as input for Artificial Neural Network classifiers. Moreover, feature selection techniques such as the Marine Predator Algorithm and Slime Mould Algorithm are implemented hybrid to cater



to the curse of dimensionality problems. In addition, selected features are fused and fed for classification. The results analysis shows that classification through features extracted from Nasnetmobile gives the best overall validation accuracy of 93.90. Overall, we conclude the following:

- Data augmentation using contrast enhancement techniques can better impact the learning of deep learning models instead of using flip and rotation-based approaches.
- Extracting encoders and deep learning features give better information on selected disease classes.
- The selection of features using a hybrid fashion impacts the classification accuracy and reduces the time.
- The fusion process improved the classification accuracy.

The drawbacks of this work are: i) segmentation of infected regions is a challenging task due to change of lesion shape and boundary location; ii) manual assignment of hyperparameters of deep learning models is not a good way, and it always affects the learning process of a network. The proposed framework will be shifted to infected region segmentation using deep learning and saliency-based techniques. Also, will opt for a Bayesian Optimization technique for hyperparameter selection. Although the proposed methodology has achieved the best outcomes yet, better accuracy may be achieved through different approaches in the future.

**Acknowledgement:** This work is supported by Princess Nourah bint Abdulrahman University Researchers Supporting Project, Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia.

**Funding Statement:** This work was supported by “Human Resources Program in Energy Technology” of the Korea Institute of Energy Technology Evaluation and Planning (KETEP), granted financial resources from the Ministry of Trade, Industry & Energy, Republic of Korea. (No. 20204010600090). Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2023R387), Princess Nourah bint Abdulrahman University, Riyadh,

Saudi Arabia.

**Author Contributions:** All authors in this work contributed equally. All authors read it and agree for the submission.

**Availability of Data and Materials:** The Kvasir dataset used in this work is publically available. thanks

**Conflicts of Interest:** The authors declare that they have no conflicts of interest to report regarding the present study.

## References

- [1] M. S. Ayyaz, M. I. U. Lali, M. Hussain, H. T. Rauf and B. Alouffi, "Hybrid deep learning model for endoscopic lesion detection and classification using endoscopy videos," *Diagnostics*, vol. 12, no. 3, pp. 43, 2021.
- [2] J. Sharmila Joseph and A. Vidyarthi, "Multiclass gastrointestinal diseases classification based on hybrid features and duo feature selection," *Journal of Biomedical Nanotechnology*, vol. 19, no. 6, pp. 288-298, 2023.
- [3] S. Kashyap, S. Pal, G. Chandan, V. Saini and S. Chakrabarti, "Understanding the cross-talk between human microbiota and gastrointestinal cancer for developing potential diagnostic and prognostic biomarkers," *Seminars in Cancer Biology*, vol. 5, no. 6, pp. 1-11, 2021.
- [4] S. Mohapatra, G. K. Pati, M. Mishra and T. Swarnkar, "Gastrointestinal abnormality detection and classification using empirical wavelet transform and deep convolutional neural network from endoscopic images," *Ain Shams Engineering Journal*, vol. 14, no. 4, pp. 101942, 2023.
- [5] N. Sharma, A. Sharma and S. Gupta, "A comprehensive review for classification and segmentation of gastro intestine tract," in *2022 6th International Conference on Electronics, Communication and Aerospace Technology*, Chennai, IND, pp. 1493-1499, 2022.
- [6] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel and A. Jemal, "GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *Ca Cancer J Clin*, vol. 68, no. 11, pp. 394-424, 2018.
- [7] I. D. Apostolopoulos and T. A. Mpesiana, "Covid-19: automatic detection from x-ray images utilizing transfer learning with convolutional neural networks," *Physical and Engineering Sciences in Medicine*, vol. 5, no. 15, pp. 1, 2020.
- [8] I. Polaka, M. P. Bhandari, L. Mezmale, L. Anarkulova and V. Veliks, "Modular point-of-care breath analyzer and shape taxonomy-based machine learning for gastric cancer detection," *Diagnostics*, vol. 12, pp. 491, 2022.
- [9] X. Pang, Z. Zhao and Y. Weng, "The role and impact of deep learning methods in computer-aided diagnosis using gastrointestinal endoscopy," *Diagnostics*, vol. 11, pp. 694, 2021.
- [10] M. Owais, M. Arsalan, J. Choi, T. Mahmood and K. R. Park, "Artificial intelligence-based classification of multiple gastrointestinal diseases using endoscopy videos for clinical diagnosis," *Journal of clinical medicine*, vol. 8, pp. 986, 2019.
- [11] V. Raut, R. Gunjan, V. V. Shete and U. D. Eknath, "Gastrointestinal tract disease segmentation and classification in wireless capsule endoscopy using intelligent deep learning model," *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, vol. 11, pp. 606-622, 2023.



- [12] H. Ko, H. Chung, W. S. Kang, K. W. Kim and Y. Shin, "Covid-19 pneumonia diagnosis using a simple 2d deep learning framework with a single chest ct image: Model development and validation," *Journal of Medical Internet Research*, vol. 22, pp. e19569, 2020.
- [13] S. Ruder, "An overview of gradient descent optimization algorithms," *Applied Sciences*, vol. 5, no. 2, pp. 1-11, 2016.
- [14] M. Farhad, M. M. Masud, A. Beg, A. Ahmad and L. Ahmed, "A Review of Medical Diagnostic Video Analysis Using Deep Learning Techniques," *Applied Sciences*, vol. 13, pp. 6582, 2023.
- [15] E. Sivari, E. Bostanci, M. S. Guzel, K. Acici and T. Ercelebi Ayyildiz, "A New Approach for Gastrointestinal Tract Findings Detection and Classification: Deep Learning-Based Hybrid Stacking Ensemble Models," *Diagnostics*, vol. 13, pp. 720, 2023.
- [16] K. Sumiyama, T. Futakuchi, S. Kamba and N. Tamai, "Artificial intelligence in endoscopy: Present and future perspectives," *Digestive Endoscopy*, vol. 33, pp. 218-230, 2021.
- [17] R. Zemouri, N. Zerhouni and D. Racocanu, "Deep learning in the biomedical applications: Recent and future status," *Applied Sciences*, vol. 9, pp. 1526, 2019.
- [18] P. Visaggi, N. de Bortoli, B. Barberio, V. Savarino and R. Oleas, "Artificial intelligence in the diagnosis of upper gastrointestinal diseases," *Journal of Clinical Gastroenterology*, vol. 56, pp. 23-35, 2022.
- [19] Y. Song and W. Cai, "Visual feature representation in microscopy image classification," in *Computer Vision for Microscopy Image Analysis*, ed: Elsevier, 2021, pp. 73-100.
- [20] V. Maeda-Gutiérrez, C. E. Galvan-Tejada, L. A. Zanella-Calzada and J. M. Celaya-Padilla, "Comparison of convolutional neural network architectures for classification of tomato plant diseases," *Applied Sciences*, vol. 10, pp. 1245, 2020.
- [21] H. Yu, R. Singh, S. H. Shin and K. Y. Ho, "Artificial intelligence in upper GI endoscopy - current status, challenges and future promise," *Journal of Gastroenterology and Hepatology*, vol. 36, pp. 20-24, 2021.
- [22] M. N. Noor, M. Nazir, I. Ashraf, N. A. Almujaally and S. Fizzah Jilani, "GastroNet: A robust attention - based deep learning and cosine similarity feature selection framework for gastrointestinal disease classification from endoscopic images," *CAAI Transactions on Intelligence Technology*, vol. 1, no. 1, pp. 1-16, 2023.
- [23] H. Borgli, V. Thambawita, P. H. Smedsrud and S. Hicks, "HyperKvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy," *Scientific data*, vol. 7, p. 283, 2020.
- [24] S. Igarashi, Y. Sasaki, T. Mikami, H. Sakuraba and S. Fukuda, "Anatomical classification of upper gastrointestinal organs under various image capture conditions using AlexNet," *Computers in Biology and Medicine*, vol. 124, pp. 103950, 2020.
- [25] M. A. Gómez Zuleta, D. F. Cano Rosales, D. F. Bravo Higuera and J. A. Ruano Balseca, "Detección automática de pólipos colorrectales con técnicas de inteligencia artificial," *Rev Colomb Gastroenterol*, vol. 36, no. 5, pp. 7-17, 2021.
- [26] M. Hmoud Al-Adhaileh, E. Mohammed Senan, W. Alsaade, T. H. H. Aldhyani and N. Alsharif, "Deep learning algorithms for detection and classification of gastrointestinal diseases," *Complexity*, vol. 2021, no. 26, pp. 1-12, 2021.
- [27] P. H. Smedsrud, V. Thambawita, S. A. Hicks, H. Gjestang and O. O. Nedrejord, "Kvasir-Capsule, a video capsule endoscopy dataset," *Scientific Data*, vol. 8, pp. 142, 2021.
- [28] A. Faramarzi, M. Heidarinejad, S. Mirjalili and A. H. Gandomi, "Marine Predators Algorithm: A nature-inspired metaheuristic," *Expert Systems with Applications*, vol. 152, no. 21, pp. 113377, 2020.
- [29] S. B. Stoecklin, P. Rolland, Y. Silue, A. Mailles and C. Campese, "First cases of coronavirus disease 2019 (COVID-19) in France: surveillance, investigations and control measures, January 2020," *Eurosurveillance*, vol. 25, pp. 2000094, 2020.
- [30] Y.T. Kim, "Contrast enhancement using brightness preserving bi-histogram equalization," *IEEE transactions on Consumer Electronics*, vol. 43, pp. 1-8, 1997.
- [31] K. R. Mohan and G. Thirugnanam, "A dualistic sub-image histogram equalization based enhancement and segmentation techniques for medical images," in *2013 IEEE Second International Conference on Image Information Processing (ICIIP-2013)*, 2013, pp. 566-569.
- [32] G. Kumar and P. K. Bhatia, "A detailed review of feature extraction in image processing systems," in *2014 Fourth International Conference on Advanced Computing & Communication Technologies*, Mumbai, India, pp. 5-12, 2014.
- [33] P. Zhou, J. Han, G. Cheng and B. Zhang, "Learning compact and discriminative stacked autoencoder for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, pp. 4823-4833, 2019.
- [34] X. Qin and Z. Wang, "Nasnet: A neuron attention stage-by-stage net for single image deraining," *Applied Sciences*, vol. 22, no. 4, pp. 1-18, 2019.
- [35] J. Li, K. Cheng, S. Wang, F. Morstatter and R. P. Trevino, "Feature selection: a data perspective," *ACM Computing Surveys (CSUR)*, vol. 50, no. 2, pp. 1-45, 2017.
- [36] A. Majid, M. Yasmin, A. Rehman, A. Yousafzai and U. Tariq, "Classification of stomach infections: A paradigm of convolutional neural network along with classical features fusion and selection," *Microscopy Research and Technique*, vol. 83, no. 6, pp. 562-576, 2020.
- [37] S. Khedkar, V. Subramanian, G. Shinde and P. Gandhi, "Explainable AI in healthcare," in *Healthcare (April 8, 2019). 2nd International Conference on Advances in Science & Technology (ICAST)*, NY, USA, pp. 1-6, 2019.
- [38] T. Nguyen-DP, M. Luong, M. Kaaniche and A. Beghdadi, "Self-supervised learning for gastrointestinal pathologies endoscopy image classification with triplet loss," *Sensors*, vol. 4, no. 1, pp. 1-21, 2023.
- [39] X. Wu, C. Chen, M. Zhong and J. Wang, "HAL: Hybrid active learning for efficient labeling in medical domain," *Neurocomputing*, vol. 456, no. 21, pp. 563-572, 2021.

# Development of HCI for Spatial Computing

Sejin Gown<sup>1</sup>, Wangyun Lee<sup>1</sup>, Soki Mai<sup>1</sup>, Seong-A Lee<sup>2</sup>, Yunyoung Nam<sup>1</sup>

<sup>1</sup>Department of ICT Convergence, Soonchunhyang University, Asan 31538, South Korea

<sup>2</sup>Soonchunhyang University, Asan, South Korea

\*Contact: [lovecein4858@naver.com](mailto:lovecein4858@naver.com)

**Abstract** – Vestibular dysfunction and dizziness caused by vestibular abnormalities greatly increase the risk of falls in the elderly. Traditional rotary chair tests and video nystagmus tests to diagnose these problems are expensive and complex procedures. Therefore, it can only be used in locations equipped with the equipment. This puts older people at risk of not receiving treatment in a timely manner. Therefore, a simpler method that can be used at home is needed. In this paper, we present a stable data collection method for an XR-based vestibular dysfunction diagnosis system using Meta Quest Pro and a swivel chair. MetaQuest Pro is an XR device that can be purchased at home that can obtain eye tracking and facial expression data. However, during the system implementation, a phenomenon in which the eye data value bounced was discovered, which was resolved by calculating the collision point with the virtual wall using the eye rotation value. Afterwards, we collected data using existing and new methods to verify the phenomenon of data values bouncing. The collection method involves recording data by rotating the chair at an average of 0.18 Hz while wearing the Meta Quest Pro. As a result of checking the data, it was confirmed that the value splashing phenomenon disappeared. However, the currently collected data is difficult to utilize because it has not been pre-processed. In addition, in this paper, only an explanation of a stable data acquisition method exists, and a vestibular abnormality diagnosis system could not be described for implementation. As a future task, we will proceed with pre-processing so that the data can be utilized. Afterwards, we plan to build an XR-based vestibular system abnormality diagnosis system by calculating gain, symmetry, and phase values using this data.

## I. INTRODUCTION

Vestibular dysfunction can cause an eye movement pattern called nystagmus, which can lead to blurred vision.[1] And 50% of vision can be reduced when caused by shaking vision due to nystagmus [2], which can cause dizziness and significantly increase the risk of falls. In addition, the incidence of vestibular dysfunction is about three times higher in the elderly than in the general adult [3]. The main cause of death in the U.S. over the age of 65 is falls [4], and one in two falls has vestibular dysfunction.[5] It can be seen that the risk of falls in the elderly due to vestibular dysfunction is very high. However, traditional diagnostic methods such as swivel chair testing and video testing are expensive and complex, which can only be accessed by specialized facilities. These restrictions mean that many older adults are not diagnosed in a timely manner, which can exacerbate the risk of falls and related injuries.

To address these challenges, this paper proposes an accessible data collection method for rotating chair inspection and imaging inspection for use at home. Meta Quest Pro is a commercialized XR device that can capture eye tracking and facial expression data. And it can be used without a separate external computing device, and it has a built-in battery, so it can be used without a

wire connection. It also uses a rotating chair in the home. Meta Quest Pro and rotating chairs are easily available at home. These two can be used to collect the data needed to diagnose vestibular dysfunction. MetaQuest only gives data on eye rotation values. However, the rotation value of the head is also included, so the rotation value of the eye alone cannot accurately determine the position of the eye. I made a wall that is fixed in front of my eyes even if my head moves. And the position of the eye was calculated by obtaining the position of the eye's gaze and the collision point of the wall. However, this approach did not immediately update the position of the fixed wall when the head moved. So, the data value was bouncing. We directly calculated the position of the virtual wall fixed to the camera. And it was solved by calculating the collision point of the eye using the rotation value of the eye.

To validate the stability of the solution, we collected data using both conventional and new methods. The data collection method involves rotating a user wearing a Meta Quest Pro to a frequency of 0.18 Hz on average with a chair. This is to reproduce the problem of periodically splashing values when the head rotates at a constant speed. As a result of the test, it was confirmed that the problem of splashing values was solved. As a result, it is confirmed that stable data collection is possible.

However, the currently collected data has no problem with splashing values, but it is difficult to utilize because it has not yet been preprocessed. This paper focuses on explaining methods for stable data acquisition. Implementation of vestibular anomaly diagnosis systems remains a future goal. The next step is to preprocess the collected data to increase utilization. And we aim to build a full XR-based vestibular anomaly diagnosis system by calculating gain, symmetry, and phase values from preprocessed data.

## II. SYSTEM IMPLEMENTATION

### A. First data collection method

In order to measure the Vestibular dysfunction, gain, Symmetry, Phase, Slow Phase must be obtained.[6] Gain is the maximum amplitude of the speed divided by the maximum amplitude of the head angular velocity by extracting the eye slow-phase. Symmetry is the difference between the maximum speed of the value extracted from the slow-phase from the left eye and the right eye. Phase is the value that the extracted slow-phase velocity maximum value is expressed as an angle by multiplying the difference in the head angular velocity maximum value by the frequency. These eye movements in the direction opposite to the direction of rotation of the head are called the slow phases and are followed by rapid phases, re-centering the eye position.[7] To obtain these values, we first need to be able to obtain stable eye position data and head angular velocity data.

MetaQuest Pro returns only the head and eyeball rotation values and locations of the eyeball movement data. And the rotation value of the eye is affected by the rotation value of the head. Unity was used as a tool for data collection programs. First, Unity made a fixed wall in front of the eyes as shown in Fig. 1. The method of obtaining the two-dimensional coordinates penetrating the screen using the rotation value of the eye was calculated using the function provided by Unity. When using this method, there is a problem that the value suddenly bounces when the head moves as shown in Fig. 2. Fig. 2 shows the x-value coordinates of the eye data. This is a problem caused by a mismatch between the movement of the head and the movement of the wall fixed to the gaze.

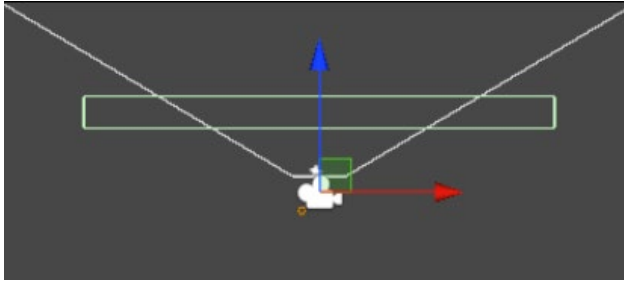


Fig.1 Fixed wall

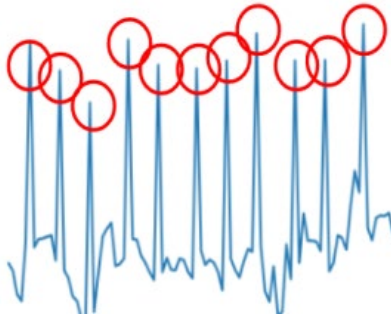


Fig.2 Fixed wall

### B.Second data collection method

Using the rotation values of the eyes, we calculated the collision point of the virtual wall as depicted in Image 3. 500 is the arbitrary value that represents the distance between the p-eye and the p-wall. Here,  $\vec{n}$  represents the normal vector from the eye position to the wall,  $\vec{v}_{\text{sight}}$  denotes the vector in the direction of the gaze,  $\vec{p}_{\text{eye}}$  and  $\vec{p}_{\text{wall}}$  are the respective position values,  $d$  stands for the distance from the eye to the plane, and  $t$  is a parameter used to find the intersection point. Then, use the obtained  $t$ -value to obtain the virtual wall collision point with the eye position and the forward-facing vector. This method does not have a problem of bouncing the value even if the head moves

$$\begin{aligned}\vec{p}_{\text{wall}} &= \vec{p}_{\text{eye}} + \vec{v}_{\text{sight}} \times 500 \\ d &= -(\vec{n} \cdot \vec{p}_{\text{wall}}) \\ t &= -\frac{(\vec{n} \cdot \vec{p}_{\text{eye}}) + d}{\vec{n} \cdot \vec{v}_{\text{sight}}} \\ \vec{p}_{\text{collision}} &= \vec{p}_{\text{eye}} + t \times \vec{v}_{\text{sight}}\end{aligned}$$

Fig.3 Fixed wall

## III. EXPERIMENTAL RESULTS

### A. Check the data splatt

To confirm the stability of the solution, we collected the data using both the first and second methods. We collected the data by rotating the chair at a speed with an average frequency of 0.18 Hz with the user wearing the Meta Quest Pro. By doing this, we reproduced the problem of periodic value bouncing when the head rotates at a constant speed. Through this problem reproduction, we wanted to see how the second method affects problem solving. As a result of the test, it was confirmed that the problem of bouncing values was solved. Accordingly, it was confirmed that stable data collection is possible. These results suggest that the second scheme provides better stability than the first. Fig. 4 is the data of the x-coordinate, and compared with Fig. 5, it can be seen that the problem of the bouncing value has been solved.

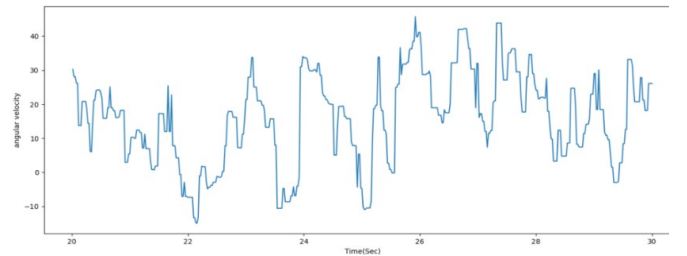


Fig.4 Normal data

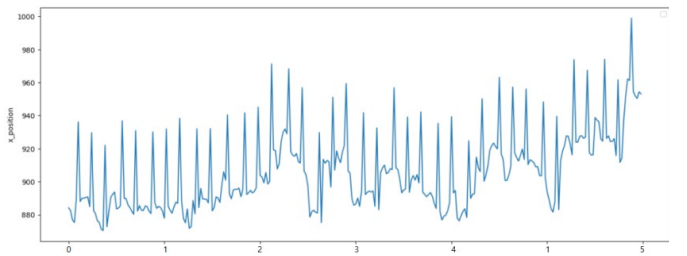


Fig.5 Bouncing data

#### IV. CONCLUSIONS

This paper presents an accessible data collection method for rotating chair inspection and video inspection to be used at home. We present a convenient yet effective way to perform data collection and analysis by using meta-quest pro and rotating chairs. However, the currently collected data has no problem with splashing values, but it is difficult to effectively utilize it because it has not yet been preprocessed. This paper focuses on explaining methods for stable data acquisition. Implementation of vestibular anomaly diagnosis systems remains a future goal. The next step is to preprocess the collected data to increase utilization. We then aim to build a full XR-based vestibular anomaly diagnosis system by calculating gain, symmetry, and phase values from preprocessed data.

#### Acknowledgmen

This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the ICAN(ICT Challenge and Advanced Network of HRD) support program(IITP-2024-2020-0-01832) supervised by the IITP(Institute for Information & Communications Technology Planning & Evaluation)

#### REFERENCES

- [1] JM Epley, Positional vertigo related to semicircular canalithiasis. *OtolaryngolHead Neck Surg.* 112(1), 154–161 (1995)
- [2] DS Zee, *The Neurology of Eye Movements* [Electronic Resource] (Oxford University Press, Oxford, 1999)
- [3] AGRAWAL, Yuri; WARD, Bryan K.; MINOR, Lloyd B. Vestibular dysfunction: prevalence, impact and need for targeted treatment. *Journal of vestibular research: equilibrium & orientation*, 2013, 23.3: 113.
- [4] FULLER, George F. Falls in the elderly. *American family physician*, 2000, 61.7: 2159-2168.
- [5] DONOVAN, Jacquelin, et al. Vestibular dysfunction in people who fall: A systematic review and meta-analysis of prevalence and associated factors. *Clinical rehabilitation*, 2023, 37.9: 1229-1247.
- [6] KONG, Youngsun, et al. A head-mounted goggle-type video-oculography system for vestibular function testing. *EURASIP Journal on Image and Video Processing*, 2018, 2018: 1-10.
- [7] KULDAVLETOVA, Olga, et al. Videoconference-based adapted physical exercise training is a good and safe option for seniors. *International journal of environmental research and public health*, 2021, 18.18: 9439.

# A Novel Deep Learning-Based Model for Classification of Wheat Gene Expression

*Amr Ismail<sup>1</sup>, Walid Hamdy<sup>1,2</sup>, Aya M. Al-Zoghby<sup>3</sup>, Wael A. Awad<sup>3</sup>, Ahmed Ismail Ebada<sup>3</sup>, Yunyoung Nam<sup>4</sup>, Mohamed Abouhawwash<sup>5,6</sup> and Byeong-Gwon Kang<sup>4,\*</sup>*

<sup>1</sup>Faculty of Science, Port Said University, Port Said, 42511, Egypt

<sup>2</sup>Modern Academy for Computer Science and Management Technology, Cairo 11742, Egypt

<sup>3</sup> Faculty of Computers and Artificial intelligence, Damietta University, New Damietta, 34511, Egypt

<sup>4</sup>Department of ICT Convergence, Soonchunhyang University, 31538, South Korea

<sup>5</sup>Department of Mathematics, Faculty of Science, Mansoura University, Mansoura, 35516, Egypt.

<sup>6</sup>Department of Computational Mathematics, Science, and Engineering (CMSE), Michigan State University, East Lansing, MI, 48824 USA.

\*Corresponding Author: Byeong-Gwon Kang. Email: [bgkang@sch.ac.kr](mailto:bgkang@sch.ac.kr)

**Abstract**— Deep learning (DL) plays a critical role in processing and converting data into knowledge and decisions. DL technologies have been applied in a variety of applications, including image, video, and genome sequence analysis. In deep learning the most widely utilized architecture is Convolutional Neural Networks (CNN) are taught discriminatory traits in a supervised environment. In comparison to other classic neural networks, CNN makes use of a limited number of artificial neurons, therefore it is ideal for the recognition and processing of wheat gene sequences. Wheat is an essential crop of cereals for people around the world. Wheat Genotypes identification has an impact on the possible development of many countries in the agricultural sector. In quantitative genetics prediction of genetic values is a central issue. Wheat is an allohexaploid (AABBDD) with three distinct genomes. The sizes of the wheat genome are quite large compared to many other kinds and the availability of a diversity of genetic knowledge and normal structure at breeding lines of wheat. Therefore, genome sequence approaches based on techniques of Artificial Intelligence (AI) are necessary. This paper focuses on using the Wheat genome sequence will assist wheat producers in making better use of their genetic resources and managing genetic variation in their breeding program, as well as propose a novel model based on deep learning for offering a fundamental overview of genomic prediction theory and current constraints. In this paper, the hyperparameters of the network are optimized in the CNN to decrease the requirement for manual search and enhance network performance using a new proposed model built on an improved algorithm and Convolutional Neural Networks (CNN).

**Keywords:** Gene expression; convolutional neural network; optimization algorithm; genomic prediction; wheat

## 1 INTRODUCTION

Cultivated crops must be increased to meet the world's population's food, feed, and fuel demand projected at more than 9 billion by 2050 [1]. One in nine people currently finds themselves living under food insecurity [2]. With limited

opportunities to expand farming on existing land, increasing yields could dramatically reduce the number of people at risk of starvation [3]. Given the need to increase crop production by 50 percent by 2050 [4], our current yield levels are inadequate to achieve this target [5]. Therefore, it is necessary and urgent to find ways to boost crop productivity, such as by genetically modifying cultivars and improving agricultural practices [6, 7]. The plant sector is the center of many countries' production. Growing plant typically has special features, such as habits, morphology, and economic value. According to statistics, several plants are registered and named worldwide [8]. We apply genomic prediction techniques in the plant recognition and identification study to make this industry successful. New approaches and techniques in the detection of plant diseases are being employed in the Genomic processing industry. Therefore, in recent years, researchers have become involved in the detection of plant diseases by using genomic processing technology for their importance and effect on farming's future. However, the prediction of the wheat gene is a new problem in machine learning. Through this method, the goal is to achieve a perfect model for wheat gene expression.

## 2 RELATED WORK

Deep learning is developing into a strong form of machine learning, which benefits both the outstanding computational resources and the very large available datasets [9]. The need to specifically define which features to use or use for data analysis is bypassed by deep learning. Deep learning then optimizes a robust end-to-end cycle by mapping data samples to outputs compatible with the large identified network training data sets. The CNNs practice this end-to-end mapping for image processing activities, by optimizing several layers of filters. The first filters are interpreted simply as low-level image features (e.g., borders, bright spots, color variations), and the subsequent layer combinations are more and more complex. CNN greatly outperforms all current alternative methods for image analysis



where adequate training is given. Results improved from 84.6 percent in 2012 [10] to 96.4 percent in 2015 [11] with benchmark-classification tasks attempting to determine which one thousand different objects are pictures.

Machine Learning technology have be used in a lot of applications in recent years, including image processing. CNN as indicated in [12] is the most common architecture and is primarily used in deep analysis. The CNN is equipped with discriminatory learning features by supervised means. In contrast to other conventional neural networks, CNN utilizes a few artificial neurons that make it suitable for image detection and processing. On the other hand, for training phases, CNN needs a broad sample number. CNN also has hyperparameters and a wide range of special architectures that are considered expensive and difficult to identify manually such as optimum hyperparameters [13]. We are responsive to the planning, which dramatically impacts CNN efficiency, of certain hyperparameters. Moreover, the hyperparameters for each dataset have to be modified because the over-parameters are different from one dataset to another. The correct values for hyperparameters for a certain dataset are calculated by trial and error since a math format is not given to manually change the hyperparameters. Selecting hyperparameter values requires detailed data that forces non-experts to use a random search or a grid seeking to find the better hyperparameters, which achieve the best performance of CNN. In [14] They used six deep neural networks and machine learning techniques to investigate and exploit the methylation patterns of the Chinese spring bread wheat cultivar in order to identify differentially expressed genes (DEGs) between leaves and roots. Genes with increased terms at leaves were mostly engaged in pigment and photosynthesis production activities, as expected, whereas genes with no difference in expression amidst leaves and roots were mostly implicated in protein processing and diaphragm structures. In [15] They used this study to see how well the DL model worked in the spring wheat breeding programme at Washington State University. They compared and evaluated the execution of two DL techniques, the convolutional neural network (CNN) and the multilayer perceptron (MLP), ridge retraction better linear equitable predictor (rrBLUP), which is a popular GS model. They used the nested association mapping (NAM) for the Spring wheat many seeded from the 2014–2016 growth seasons yielded 650 recombinant inbred lines (RILs). They used cross-validations (CVs), alternative sets, and independent validations of single nucleotide polymorphisms (SNP) markers, they made predictions for five various quantitative variables using various genetic architectures. Hyperparameters for models of DL were adjusted by decreasing the root average square in the training dataset and employing dropout and regularization to avoid model overfitting.

In [16] they used R-CNN Faster to verify the spike number by using the dataset for high-density wheat 660K SNP array. they achieved an accuracy of 86.7%. They approve that the R-CNN Faster model is faster and has a high accuracy that may be applied to genetic investigations of SN in wheat.

### 3 DEEP LEARNING PRINCIPLES

A standardized DL architecture consists of a mixture of multiple "neurons" layers. In the 50s, with a prominent

"perceptron" of Rosenblatt, the idea of a nerve network was proposed, inspired by the activity of the brain [17]. In the past decade, the DL resurgence was focused on the development of powerful algorithms which can be used in complex network parameters containing multiple layers of neurons (e.g. backpropagation) [18] and on the fact that they surpass current algorithms in various automated recognizing functions like picture checking [19]. Deep learning is an area of many specific jargon terms, which means that some of the most crucial terms are defined in Fig. 1 to make understanding easier for an inexperienced user.

Fig. 1 Multistage perceptron (MLP) graph displaying the feedback of the simple "Neuron" with  $n$  inputs and four hidden layers of single nucleotide polymorphisms (SNPs). The linear combinations' nonlinear transformations ( $x_i$ ,  $w_i$ , and biases  $b$ ) all culminate in a single neuron. where  $x_i$  represents the neuron's  $i$  input,  $w_i$  represents a weight connected by input  $i$ , and  $b$  represents a time-invariant alignment level.

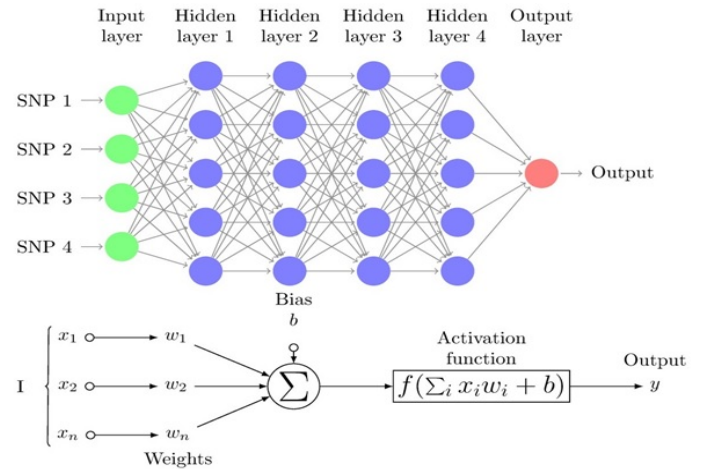


Figure 1: Multistage perceptron (MLP).

The linear combinations' nonlinear transformations ( $x_i$ ,  $w_i$ , and biases  $b$ ) all culminate in a single neuron. where  $x_i$  represents the neuron's  $i$  input,  $w_i$  represents a weight connected by input  $i$ , and  $b$  represents a time-invariant alignment level.

#### 3.1 Deep Learning Architectures

Although all DL techniques generally use stacked neuron layers, they do also include a large architecture. The most prevalent ones are convolutional neuro-networks (CNN), multilayer perceptron (MLP), generative adversarial networks (GANs), and recurrent neural networks (RNNs). These are listed in effect, although the reader should be aware of various additional options [20].

The multi-layer perceptron network (MLP) consists of a set of completely connected layers named hidden and input layers (see Fig. 2) and is one of the most common DL architectures. The first layer receives SNP genotypes ( $x$ ) feedback in the sense of genomic prediction [21], while the initial layer's output is a weighted, non-linear function of all feedback plus a "bias". Then the first output layer is shown in Eq. (1):



$$z(1) = b_0 + W(0) f(0)(x) \quad (1)$$

When  $x$  includes each individual's genotypes,  $b$  is considered a "bias" and is measured along with the remaining weights  $W_0$  and  $f$  is a nonlinear function (activation function available on Keras). The same term is used in successive layers so that the neuron's inputs of a certain layer are the outputs of the preceding layer  $z(k-1)$ :

$$z(k) = b_k + W(k-1) f((k-1)(z(k-1))) \quad (2)$$

The final layer generates a number matrix, whether the goal is a true phenotype, or if the goal is a class (ie a problem with classifying) an array of probabilities for each point. Although MLPs constitute a powerful strategy for managing classification or regression issues, they are not the perfect way to handle space or time sets [22]. In latest years, other methods of DL have been suggested in order to deal with these challenges, such as recurrent neural networks, deep generative networks, or convolutional neural networks.

Input variables have been spread in accordance with space models with one dimension (for example, SNPs or text) and two or three dimensions (for example, images), to conform to the circumstances of the implementation of neural networks. CNN's have been introduced. CNN is a particular type of neural network that uses convolution in hidden layers rather than of full matrix reproduction [23]. A CNN consists of thick layers and "convolutional layers" that are fully connected (Fig. 2). An overall operation as well as the input of predetermined width and steps are done in every convolutionary layer. A 'kernel' or 'buffer' is a collection of convolutional processes that functions similarly to a 'neuron' in an MLP [24].

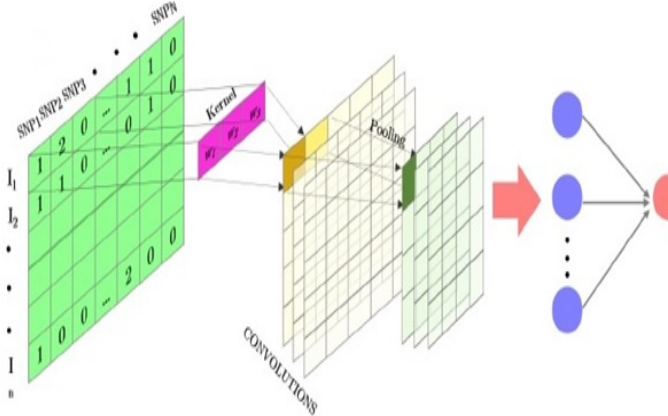


Figure 2: Total view of 1D fully convolutional SNP-matrix neural network.

After each convolution, the output is generated using an activation function. Finally, the results are frequently evened out through a "pooling" method. The kernel outputs of the various positioning positions are combined by using all values of those positions on average, maximum, or minimum. Its capability to which the amount of parameters to be determined is one of the main advantages of convolution networks. These

networks have already restricted connections and are translations similar. Fig. 3 provides an example of a one-dimensional (1D) kernel convolution with a scale of 3K [25].

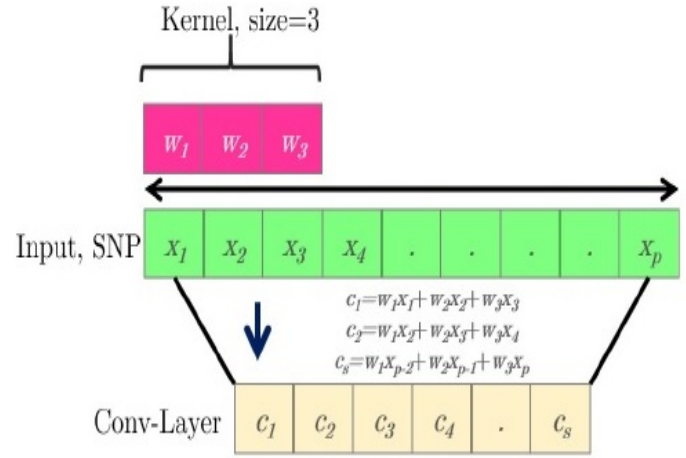


Figure 3: Simple one-dimensional (1D) operation scheme.

### 3.2 Convolutional Neural Network

The CNN is so good at categorizing simple patterns in data, it might be utilized to build additional complicated patterns during higher layers. CNNs are a specific type of multilayer neural network. It is trained using the backpropagation algorithm, which is used by practically all other neural networks. CNN's architecture sets it apart from the competition. In a CNN design, there are input layers, numerous hidden levels, and output layers. The hidden layer is made up of pooling layers, Convolutional layers, and fully connected layers [26].

The input data is received by the convolution layer, which applies a filter to it, essentially, the input data is multiplied by the kernel to generate the adjusted output data. A convolution layer subsampling method is the Pooling layer. The goal is to reduce the number of dimensions. An input layer serves as the first layer in the proposed CNN algorithm used in this study. The second layer makes up a one-dimensional convolution layer with three kernel sizes, a 30 filter, and Rectified Linear Units (RELU) activation. The third layer is the max pooling layer, with two pool sizes. The next layer is a completely connected layer with the ability to activate RELU.

Finally, the output layer is made up of a single sigmoid activation in a neuron. The ADAM optimizer is applied for learning, as a cost function with binary cross-entropy.

### 3.3 Recurrent Neural Network

RNN is the only algorithm with internal memory. Therefore, it is a very powerful and reliable algorithm, the RNN is incredibly powerful since it is still the only algorithm with internal memory. The internal memory of the RNN allows the algorithm to recall and research critical information about the input it receives; this ability allows the program to predict what will happen next with great accuracy [27]. The information in an RNN loops back on itself. As demonstrated in Fig. 4, it considers the current input as well as what it has learned from

previous inputs before making a decision.

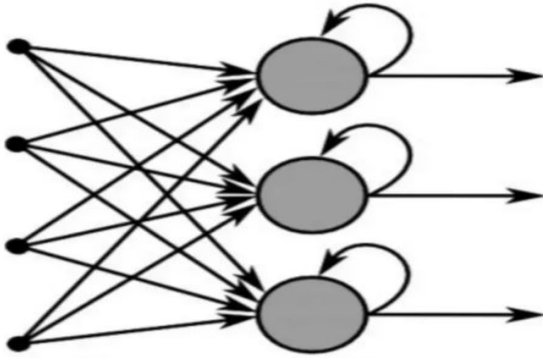


Figure 4: Recurrent Neural Network.

This study employs a simple RNN layer, with the output being fed back into the input. The simple RNN layer is used to apply the RELU activation function. A sigmoid activation algorithm was also employed for the output layer. For learning, the ADAM optimizer is employed, and as a cost function, binary cross-entropy is used.

#### 4 The Proposed Approach

In this section, the dataset used to implement the proposed approach is first described, then the details of the approach proposed are explained.

##### 4.1 Dataset Description

The data for this study is from the Global Wheat data set, which is open to the public [28]. The original genotypic data consisted of 73,345 polymorphic markers anchored to the Chinese Spring Ref Seqv1 map. Before filtering the genotypic data, RILs with lacking phenotypic data in a single setting were deleted. SNP markers having a missing data rate of higher than 20%, minor allele frequencies of less than 0.10, and RILs with more than 10% genotypic data were also eliminated, leaving 40,000 SNP and 635 RILs markers for analysis. Using 635 RILs and 40,000 SNP markers, The demographic structure of the 26 NAM families was investigated using principal component analysis (PCA).

##### 4.2 The Proposed Approach for Classification of Wheat Gene Expression

The solution suggested is an incredibly effective way of optimizing the efficiency of the CNN network by the incorporation of the terminals of two pre-trained CNN networks. In fact, the model's hyperparameters are designed such that each model performs better. As seen in Fig. 5, the proposed algorithm steps will be summed up in four main stages, i.e. (a) stage planning details, (b) stage optimization hyperparameters, (c) learning stage, (d) evaluation stage. The following parts explain additional descriptions for every point

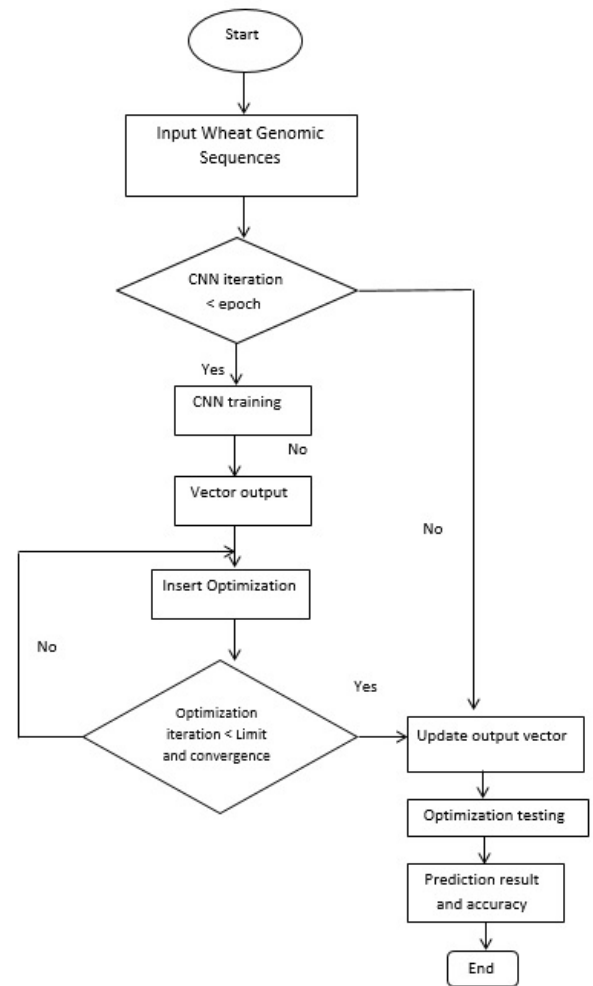


Figure 5: The proposed Approach block diagram.

The Multilayer Perceptron (MLP) with numerous hidden layers is an excellent example of a model with a deep architecture. On huge data, the most recent deep learning algorithm has overcome generalization, training stability, and scale are all issues that need to be addressed. Deep learning algorithms are typically the algorithm of choice for reliable forecast accuracy, and they perform well in a wide range of problems. There are several theoretical frameworks for deep learning, and we adopt the feedforward architecture in this study [29].

MLPs are feedforward neural networks with an architecture consisting of three layers: input, hidden, and output, as shown in Fig. 1. Neurons are little particles that make up each layer. The neurons in the input layer receive the input data  $X$  and forward it to the next layer of the network. The following layer, the hidden layer, receives input from each neuron; these data are a weighted total of the neuron's outputs from the preceding layer. Each neuron uses an activation function to govern the input. A nonlinear mapping of an output vector to an input vector is created by this network, with weights (the vector of weights) as the parameters ( $W$ ). The initial phase is to select the parameters of weight for the model and determine the MLP's right structure, which is dictated by the number of neurons and hidden layers, as well as the number of output and input variables and the kind of activation function. Second, using the training data derives the weight parameters. The training selects

the proper weight vector  $W$  to ensure that the output is as close to the aim as possible [30].

Our suggested MLP approach includes input layers, four hidden layer, and one neuron output layer. Except for the output layers, all levels utilize the rectified linear unit (RELU) activation function, which utilizes the non-linearity sigmoid activation function. The MLP model is trained using the backpropagation algorithm. ADAM: For the learning algorithm, a stochastic gradient descent optimizer was employed, with binary cross entropy as a cost function [31-51].

#### 4.3 Improving MLP by Applying Dropout

Dropout is a model improvement strategy that prevents the model from overfitting. Dropout refers to the process of removing nodes from a neural network. Excluding a node means tentatively removing it from the network, together with all of its incoming and outgoing connections. The nodes are dropped out at random. As can be seen in Fig. 6. The dropout strategy was used in our proposed MLP model with a 20% dropout rate for all hidden layers, and the results were significantly better for our dataset.

expression levels from our dataset. Each time, 95 samples were used as testing data and 285 samples were used as training data to train the convolutional neural network architecture. There were 60 epochs in total. Then, by 95 samples ( $4 \times 95 = 380$ ) and  $k=4$ , we rank-fold cross-validation (CV). The selecting test data were then randomized to a 10-time process of randomization, after which the average value for the following machine learning metrics—accuracy, specificity, recall (sensitivity), and precision—was calculated. These matrices show the relationship between correctly and incorrectly predicted outcomes. TN (True Negative), TP (True Positive), FN (False Negative), and FP (False Positive) are the four categories in the confusion matrix. which were defined as follows:

$$\text{Accuracy} = (\text{TP} + \text{TN}) / (\text{TP} + \text{TN} + \text{FP} + \text{FN})$$

(3)

$$\text{recall} = \text{TP} / (\text{TP} + \text{FN})$$

(4)

$$\text{precision} = \text{TP} / (\text{TP} + \text{FP})$$

(5)

$$\text{specificity} = \text{TN} / (\text{TN} + \text{FP})$$

(6)

The size and number of convolutional filters, as well as the number and size of convolutional layers and hidden layers, were all examined in various combinations. With the architecture, the best outcomes were obtained. Table 1 shows that with our sample, with a mean accuracy of 99.4%, the improved DNN was the most accurate, followed by DNN with 98.2% and 97.5% for CNN and RNN, respectively.

*Table 1: COMPARISON OF CLASSIFICATION ACCURACY RESULTS WITH THE IMPROVED DNN, DNN, RNN, AND CNN.*

Overall, the improved DNN algorithm can be observed that attained maximum accuracy in this study's dataset. There were 100 epochs in total. Fig. 7 shows the curves of the high-accuracy model discovered by Improved DNN on the convex dataset for 100 epochs when compared to DNN, RNN, and CNN models. We can see that accuracy of our models have improvement when compared to other models, implying that the Improved DNN is actually capable of identifying perfect models for a given dataset.

Epoch number	Improved DNN	DNN	RNN	CNN
Epoch 1	99.1	98.1	97.5	97.4
Epoch 2	99.3	98.3	97.4	97.3
Epoch 3	99.4	98.0	97.2	97.5
Epoch 4	99.0	98.2	97.3	97.4
Epoch 5	98.9	98.3	97.6	97.1
Epoch 6	99.0	97.98	97.4	97.5
Epoch 7	98.8	97.99	97.3	97.4
Epoch 8	99.4	98.0	97.5	97.3
Epoch 9	99.3	98.1	97.1	97.2
Epoch 10	99.2	98.2	97.4	97.4

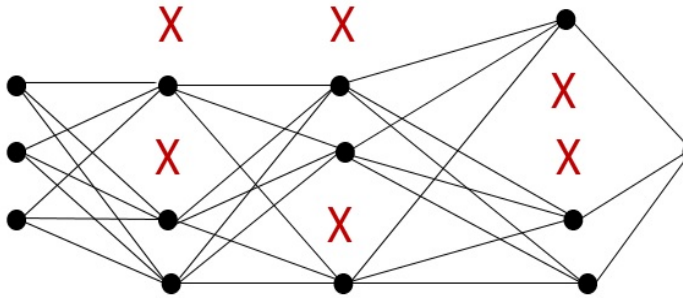


Figure 6: MLP with Dropout.

#### 5 Experiments and Result

The protein interaction network was mapped to the gene

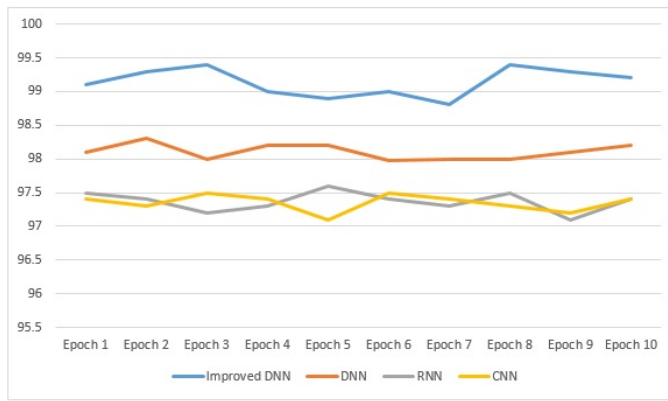


Figure 7: The comparison results.

The performance results based on the three matrices (Accuracy, specificity, and precision) are shown in Table 2. Accuracy, specificity, and precision have respective means of 99.4 %, 98.64 %, and 91.85 %.

Table 2: Results respective Accuracy, specificity, and precision.

Models	Accuracy	specificity	precision
Improved DNN	99.4	98.64	91.85
DNN	98.2	97.8	89.9
RNN	97.5	97.1	89.96
CNN	97.5	97.2	89.5

Fig. 8 shows the curves of the performance results based on the three matrices (Accuracy, specificity, and precision) when compared to DNN, RNN, and CNN models. We can see that accuracy of our models, specificity, and precision have improved when compared to other models.

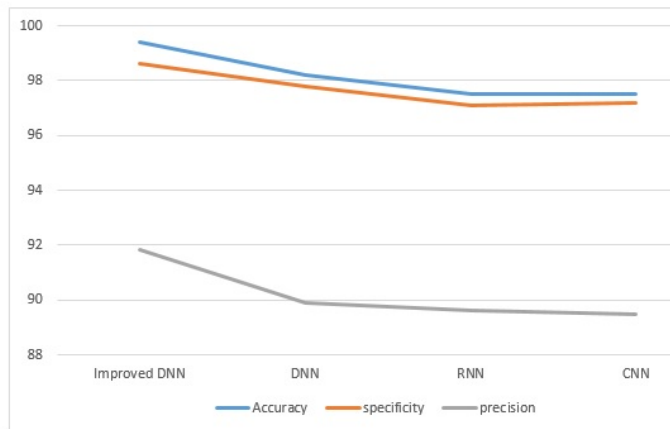


Figure 8: The comparison results (Accuracy, specificity, and precision).

We compared our model to previous work and other models. For training and validation, the majority of the methods used a train/test split. As shown in Table 3, our results outperform other models.

Table 3: Comparison of the suggested model with other proposed models

Year	Existing Work	Accuracy
2021	K. S. Sandhu et al [15]	95
2022	L. Li et al [16]	86.7
2022	L. Mingxuan et al [52]	98.75
Our Proposed model		99.4

## 6 CONCLUSION

In this paper, we have presented a novel deep learning-based model which improves DNN by applying the dropout model to classify Wheat gene expressions. In addition to, the deep learning algorithms CNN, DNN, and RNN, and the proposed model are implemented for the classification of gene expression data. Moreover, the outliers and noisy data are addressed, by using a pre-processing methodology for all features of gene expression, after that we trained all of our models individually using a perfect framework and learning method. Finally, our learned models are applied to testing data to classify it. For all of the datasets studied, the Improving-DNN outperformed other models in accuracy terms from the result illustrated our Improving-DNN has a high accuracy of 99.4%, while DNN has 98.2% accuracy, RNN and CNN have 97.5% accuracy. Therefore, the Improving-DNN model is actually more appropriate for solving the wheat gene expression dataset.

In future work, we will apply our proposed model to another dataset in many fields, especially in agriculture. Furthermore, it would be interesting to study the influence of combining additional deep learning models or using different optimization models.

## Funding Statement

This research was supported by Korea Institute for Advancement of Technology(KIAT) grant funded by the Korea Government(MOTIE) (P0012724, The Competency Development Program for Industry Specialist), the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. RS-2023-00218176), and the Soonchunhyang University Research Fund.

## Availability of Data and Materials

The data presented in this study are available on request from the corresponding author.

## Conflicts of Interest

The authors declare that there is no conflict of interests regarding the publication of the paper.

## REFERENCES

- [1] U.N. Desa, "World population prospects 2019: Highlights," New York (US): United Nations Department for Economic and Social Affairs, vol 11, no. 1, pp. 125. 2019.
- [2] S. McGuire, "FAO, IFAD, and WFP. The state of food insecurity in the world 2015: meeting the 2015 international hunger targets: taking stock of uneven progress. Rome: FAO," *Advances in Nutrition*, vol. 6, no. 5, pp. 623-624, 2015.
- [3] M.W. Rosegrant, S. Tokgoz, P. Bhandary and S. Msangi, "Looking Ahead: Scenarios for the Future of Food. 2012 Global Food Policy Report. IFPRI.

- Washington," International Food Policy Research Institute (IFPRI), vol. 4, no. 3, pp. 1-15, 2013.
- [4] D. Tilman, C. Balzer, J. Hill and B. L. Befort, "Global food demand and the sustainable intensification of agriculture," *Proceedings of the National Academy of Sciences - PNAS*, vol. 108, no. 50, pp. 20260-20264, 2011.
- [5] D. K. Ray, N. D. Mueller, P. C. West and J. A. Foley, "Yield trends are insufficient to double global crop production by 2050," *Plos One*, vol. 8, no. 6, pp. e66428-e66428, 2013.
- [6] H. Spiertz, "Avenues to meet food security. The role of agronomy on solving complexity in food production and resource use," *European Journal of Agronomy*, vol. 43, no. 5, pp. 1-8, 2012.
- [7] J.L. Araus, R. Park, D. Calderini, D. Miralles, T. Shen et al., "Prospects of doubling global wheat yields," *Food and Energy Security*, vol. 2, no. 1, pp. 34-48, 2013.
- [8] W.Hamdy, A. Ismail, W. A. Awad, A.H. Ibrahim and A. Hassanien, "A Support Vector Machine Model for Rice (*Oryza sativa* L.) Leaf Diseases Based on Particle Swarm Optimization." In *Artificial Intelligence: A Real Opportunity in the Food Industry*, Springer, Cham, pp. 45-54, 2023.
- [9] Y. LeCun, Y. Bengio and G. Hinton, "Deep learning," *Nature (London)*, vol. 521, no. 7553, pp. 436-444, 2015.
- [10] A. Elaraby, W. Hamdy and M. Alruwaili, "Optimization of deep learning model for plant disease detection using particle swarm optimizer," *Computers, Materials & Continua*, vol. 71, no. 2, pp. 4019-4031, 2022.
- [11] J. R. Ubbens and I. Stavness, "Corrigendum: deep plant phenomics: A deep learning platform for complex plant phenotyping tasks," *Frontiers in Plant Science*, vol. 8, no. 12, pp. 2245-2245, 2018.
- [12] A. Elaraby, W. Hamdy and S. Alanazi, "Classification of citrus diseases using optimization deep learning approach," *Computational Intelligence and Neuroscience*, vol. 2022, no. 10, pp. 9153207-9153212, 2022.
- [13] N. Ni and S. Xu, "Model optimization strategies based on deep neural networks Learning and application of pruning optimization algorithms," *Journal of Physics, Conference Series*, vol. 2303, Dali City, China, no. 1, pp. 012033, 2022.
- [14] A. N. Diaye, B. Byrns, A.T. Cory, K.T. Nilsen, S. Walkowiak et al., "Machine learning analyses of methylation profiles uncovers tissue-specific gene expression patterns in wheat," *The Plant Genome*, vol. 13, no. 2, pp. e20027, 2020.
- [15] K. S. Sandhu, D. N. Lozada, Z. Zhang, M. O. Pumphrey and A. H. Carter, "Deep learning for predicting complex traits in spring wheat breeding program," *Frontiers in Plant Science*, vol. 11, no. 4, pp. 613325-613335, 2021.
- [16] L. Li, M.A. Hassan, S. Yang, F. Jing, M. Yang et al., "Development of image-based wheat spike counter through a Faster R-CNN algorithm and application for genetic studies." *The Crop Journal*, vol. 12, no. 3, pp. 1-12, 2022.
- [17] P. Matteo, "Machines that morph logic: neural networks and the distorted automation of intelligence as statistical inference." *Glass Bead*, vol. 1, no. 1, pp. 25-36, 2017.L. B. Balzer and M. L. Petersen, "Invited commentary: machine learning in causal inference-how do I love thee? let me count the ways," *American Journal of Epidemiology*, vol. 190, no. 8, pp. 1483-1487, 2021.
- [18] I. H. Sarker, "Deep learning: a comprehensive overview on techniques, taxonomy, applications and research directions," *SN Computer Science*, vol. 2, no. 6, pp. 420-420, 2021.
- [19] G. Jing, P. Li, Z. Chen and J. Zhang, "A survey on deep learning for multimodal data fusion," *Neural Computation*, vol. 32, no. 5, pp. 829-864, 2020.
- [20] G. R. T. de Lima and G. B. Scofield, "Feasibility study on operational use of neural networks in a flash flood early warning system," *Revista Brasileira de Recursos hídricos*, vol. 26, no. 2, pp. 1-10, 2021.
- [21] J. M. Silva, A. Figueiredo, J. Cunha, J.E. Dias, S. Silva et al., "Using rapid chlorophyll fluorescence transients to classify vitis genotypes," *Plants (Basel)*, vol. 9, no. 2, pp. 174-189, 2020.
- [22] M. Mostavi, Y.-C. Chiu, Y. Huang and Y. Chen, "Convolutional neural network models for cancer type prediction based on gene expression," *BMC Medical Genomics*, vol. 13, no.5, pp. 1-13, 2020.
- [23] S.D. O'Donovan, K. essens, D. Lopatta, F. Wimmenauer, A. Lukas et al., "Use of deep learning methods to translate drug-induced gene expression changes from rat to human primary hepatocytes," *Plos One*, vol. 15, no. 8, pp. e0236392, 2020.
- [24] H. Lahmer, A. E. Oueslati and Z. Lachiri, "Classification of DNA microarrays using deep learning to identify cell cycle regulated genes," 5th International Conference on Advanced Technologies for Signal and Image Processing (ATSIP), Sousse, Tunisia, pp. 1-5, 2020.
- [25] B. He, L. Bergenstrahle, L. Stenbeck, A. Abid, A. Andersson et al., "Integrating spatial gene expression and breast tumour morphology via deep learning," *Nature Biomedical Engineering*, vol. 4, no. 8, pp. 827-834, 2020.
- [26] O. Ahmed and A. Brifcani, "Gene expression classification based on deep learning," 4th Scientific International Conference Najaf (SICN), Najaf, Iraq, pp. 145-149, 2019
- [27] K.W. Jordan, S. Wang, F. He, S. Chao, Y. Lun et al., "The genetic architecture of genome - wide recombination rate variation in allopolyploid wheat revealed by nested association mapping," *The Plant Journal: for Cell and Molecular Biology*, vol. 95, no. 6, pp. 1039-1054, 2018.
- [28] W. Hamdy, A. Darwish and A. E. Hassanien, "Artificial intelligence strategy in the age of covid-19: opportunities and challenges," *Digital Transformation and Emerging Technologies for Fighting COVID-19 Pandemic: Innovative Approaches*, Cham: Springer International Publishing, pp. 81-93, 2021.
- [29] Y.C. Chiu, H. Chen, T. Zhang, S. Zhang, A. Gorthi et al., "Predicting drug response of tumors from integrated genomic profiles by deep neural networks," *BMC Medical Genomics*, vol. 12, no. 1, pp. 18-155, 2019.
- [30] H. Wang, R. Liu, P. Schyman and A. Wallqvist, "Deep neural network models for predicting chemically induced liver toxicity endpoints from transcriptomic responses," *Frontiers in Pharmacology*, vol. 10, no. 4, pp. 42-42, 2019.
- [31] A.H. ElBassiouny, M. Abouhawwash and H.S. Shahren, "New generalized extreme value distribution and its bivariate extension," *International Journal of Computer Applications*, vol. 173, no. 3, pp. 1-10, 2017.
- [32] A.H. ElBassiouny, M. Abouhawwash and H.S. Shahren, "Inverted exponentiated gamma and its bivariate extension," *International Journal of Computer Application*, vol. 3, no. 8, pp. 13-39, 2018.
- [33] A.H. ElBassiouny, H.S. Shahren and M. Abouhawwash, "A new bivariate modified weibull distribution and its extended distribution," *Journal of Statistics Applications & Probability*, vol. 7, no.2, pp. 217-231, 2018.
- [34] M. Abouhawwash and M.A. Jameel, "KKT proximity measure versus augmented achievement scalarization function," *International Journal of Computer Applications*, vol. 182, no. 24, pp. 1-7, 2018.
- [35] H.S. Shahren, A.H. El-Bassiouny and M. Abouhawwash, "Bivariate exponentiated modified weibull distribution," *Journal of Statistics Applications & Probability*, vol. 8, no. 1, pp. 27-39, 2019.
- [36] M. Abouhawwash and M.A. Jameel, "Evolutionary multi-objective optimization using benson's karush-kuhn-tucker proximity measure," *International Conference on Evolutionary Multi-Criterion Optimization*, East Lansing, Michigan, USA, Springer, pp. 27-38, 2019.
- [37] M. Abouhawwash, M.A. Jameel and K. Deb, "A smooth proximity measure for optimality in multi-objective optimization using benson's method," *Computers & Operations Research*, vol. 117, no. 2, pp. 104900, 2020.
- [38] M. Abouhawwash, K. Deb and A. Alessio, "Exploration of multi-objective optimization with genetic algorithms for PET image reconstruction," *Journal of Nuclear Medicine*, vol. 61, no. 1, pp. 572-572, 2020.
- [39] S. Ibrahim, H. Alhumyani, M. Masud, S.S. Alshamrani, O. Cheikhrouhou et al., "Framework for efficient medical image encryption using dynamic S-boxes and chaotic maps," *IEEE Access*, vol. 8, no. 13, pp. 160433-160449, 2020.
- [40] M. Rawashdeh, M. Zamil, S. M. Samarah, M. Obaidat and M. Masud, "IOT-based service migration for connected communities," *Computers & Electrical Engineering*, vol. 96, no. 2, pp. 1-10, 2021.
- [41] A. Roozbahani, H. Ghased and M. H. Shahedany, "Inter-basin water transfer planning with grey COPRAS and fuzzy COPRAS techniques: A case study in Iranian Central Plateau," *Sci. Total Environ.*, vol. 726, pp. 138499, 2020
- [42] S. Ibrahim, H. Alhumyani, M. Masud, S. S. Alshamrani, O. Cheikhrouhou et al., "Framework for efficient medical image encryption using dynamic S-boxes and chaotic maps," *IEEE Access*, vol. 8, no. 13, pp. 160433-160449, 2020.

- [43] M. Kumar, K. Venkatachalam, M. Masud and M. Abouhawwash, "Novel dynamic scaling algorithm for energy efficient cloud computing, " *Intelligent Automation & Soft Computing*, vol.33, no.3, pp. 1547-1559, 2022.
- [44] R.S. Ram, K. Venkatachalam, M. Masud and M. Abouhawwash, "Air pollution prediction using dual graph convolution LSTM technique," *Intelligent Automation & Soft Computing*, vol.33, no.3, pp. 1639-1652, 2022.
- [45] A. Mutlag, M. Ghani and M. Mohammed, "A healthcare resource management optimization framework for ECG biomedical sensors," *Proc. Efficient Data Handling for Massive Internet of Medical Things Springer*, vol. 12, no. 5, pp. 229–244, 2021.
- [46] G. Ravikumar, K. Venkatachalam, M.A. AlZain, M. Masud and M. Abouhawwash, "Neural cryptography with fog computing network for health monitoring using IoMT," *Computer Systems Science and Engineering*, vol. 44, no.1, pp.945-959, 2023.
- [47] R. Rajdevi, K. Venkatachalam, M. Masud, M.A. AlZain and M. Abouhawwash, "Proof of activity protocol for IoMT data security," *Computer Systems Science and Engineering*, vol. 44, no. 1, pp.339-350, 2023.
- [48] Y. Wang, J. Ma, A. Sharma, P. K. Singh, G. Singh et al., "An exhaustive research on the application of intrusion detection technology in computer network security in sensor networks," *Journal of Sensors*, vol. 2021, no. 12, pp. 1–11, 2021.
- [49] N. Mittal, H. Singh, V. Mittal, S. Mahajan, A.K. Pandit et al., "Optimization of cognitive radio system using self-learning salp swarm algorithm," *Computers, Materials & Continua*, vol.70, no.2, pp.3821-3835, 2022.
- [50] J. Saini, M. Dutta and G. A. Marques, "Comprehensive review on indoor air quality monitoring systems forenhanced public health," *Sustainable Environment Research*, vol. 30, no. 6, pp. 2–17, 2020.
- [51] L. Mingxuan, G. Zhou, A. Chen, J. Yi, Chao, M. He et al., "FWDGAN-based data augmentation for tomato leaf disease identification." *Computers and Electronics in Agriculture*, vol. 194, pp. 106779, 2022.



# A Gamified Cognitive Behavioral Therapy to Reduce Symptoms of Depression, Anxiety, and Stress

NOURHAN A.AMERI, SAMAA M. SHOHIEB1, WALEED ELADROSY2, SHILONG LIU3, YUNYOUNG NAM3\*, AND SAMIR ABDELRAZEKI

1 Information Systems Department, Faculty of Computers and Information, Mansoura University, Mansoura 35516, Egypt

2 Computer Science Department, Faculty of Computers and Information, Mansoura University, Mansoura 35516, Egypt

3 Department of ICT Convergence, Soonchunhyang University, Asan 31538, Korea

Corresponding author: Yunyoung Nam (ynam@sch.ac.kr).

“This work was supported in part by the U.S. Department of Commerce under Grant BS123456.”

**Abstract**— Depression, anxiety, and stress are common diseases among a large number of individuals from different societies with different social levels and have a significant impact on their life and production. Recently, it has become a common practice to provide online psychological therapies via mobile phones, including cognitive behavioral therapy (CBT). CBT is a successful therapeutic intervention for many ailments. This study investigates the development and evaluation of Sokoon; a gamified CBT application that aims to increase CBT skills adherence and engagement in individuals who have depression, stress, and anxiety by encouraging the user to learn CBT skills through a series of activities and games. Psychiatrists were consulted to determine the appropriate skill set. The application is set in the form of seven planets to learn seven evidence-based skills, each planet represents a specific skill, which are relaxation, behavioral activation, gratitude, problem-solving, self-love, social skills, and cognitive restructuring. This research focuses on four skills initially for application: relaxation, gratitude, behavioral activation, and cognitive restructuring. Sokoon is the first app that integrates many techniques; namely, gamification and Hexad theory, this was applied using a dynamic difficulty adjustment (DDA) algorithm. These techniques can provide a more enjoyable and engaging experience than traditional CBT techniques intervention, while still providing the same level of effectiveness. To determine the efficiency level of Sokoon regarding depression, anxiety, and stress symptoms, a randomized controlled experiment was done. Results revealed that the Sokoon group had dramatically lessened stress, anxiety, and depressive symptoms.

**INDEX TERMS**— Cognitive Behavioral Therapy (CBT), Gamification, Hexad theory, DDA, Depression, Anxiety, Stress.

## I. INTRODUCTION

Mental disorders are illnesses characterized by changes in thought, emotion, or behavior (or any combination of these) that are connected to suffering and/or poor functioning, giving rise to a wide range of issues for people, such as disability, discomfort, or even death [1]. Depression is a widespread and dangerous medical condition that has an adverse impact on how one feels, thinks, and behaves [2]. Symptoms of depression include sadness and/or a loss of interest in previously appreciated activities. It is estimated that in any given year, depression affects one in 15 adults (6.7%), and 16.6% of people (or one in six) will experience depression at some point

in their lives [2]. Researchers have discovered that individuals with social phobia are more likely to develop chronic depression [3]. Chronic stress increases the likelihood of developing psychiatric disorders such as anxiety and depression [4].

Most mental health organizations support Cognitive Behavior Therapy (CBT) as an evidence-based therapy for treating and managing depression and anxiety [5]. However, access to CBT delivered by specialists is still limited for many patients. This limitation stems not only from a scarcity of specialized practitioners but also from other factors, such as the need for patients to travel and attend during normal working hours, as well as the cost [5]. Therefore, the benefits and practicality of self-help treatments such as computerized CBT (CCBT), which is self-help CBT using a program on a website or a computer without Internet access, have been appealing. It is expected that self-help CBT will be an effective intervention, particularly for mild-to-moderate depression [6, 7]. Computerized cognitive behavior therapy (CCBT) can be used as a stand-alone treatment or as part of a stepped-care treatment plan and is beneficial for people with anxiety and/or depression [8].

Gamification and serious games are traditionally defined as the use of game-playing components, such as points and scoring, for goals other than play, most commonly to promote motivation and enhance abilities [9, 10]. According to some research, gamification has the potential to increase user engagement and adherence to mental health applications, which can enhance the effectiveness of therapeutic-based apps (such as CBT) and minimize depression symptoms. It may stimulate reward-mediated brain pathways, prompting pleasurable feelings that may counter some of the negative feelings associated with depression. In their opinion, gamified mental health apps will be superior to those without gamification in terms of reducing depression symptoms and encouraging adherence [11]. The number and sophistication of "mHealth" interventions have grown along with the prevalence of mobile phones in everyday life, and many health-related smartphone apps now feature gamification [12, 13]. Available research indicates that cognitive-behavioral therapy (CBT) techniques used in smartphone-based therapies can dramatically reduce depressive symptoms [14-17].

Recently, gamification has attracted scholarly attention as a tool for changing behavior in a variety of contexts. Several experts emphasize the importance of adapting material to the needs of different users when designing gamification, such as by employing the gamification user types hexad typology [18]. Based on their capacity to be motivated by intrinsic factors (such as self-realization) or extrinsic factors (such as rewards), Marczewski hypothesized six main user categories: Socializers, Free Spirits, Achievers, Players, Disruptors, and Philanthropists [19, 20].

One problem with many traditional games is the fixed difficulty of each level. The difficulty is typically predetermined by the designers, regardless of the skill level of each player. For some players, this makes the game levels either extremely challenging or absurdly simple. Dynamic Difficulty Adjustment (DDA) is a method for resolving this issue [21]. DDA is one of the fundamental methods used in behavior analysis. It is a technique for instantly altering video game characteristics, scenarios, and behaviors based on a player's performance, keeping them from getting bored (when the game is too easy) or frustrated (when the game is too difficult). DDA aims to provide players a challenging experience while keeping them interested until the end [22].

In this paper, we provide an overview of the steps we took to create a gamified mobile health CBT intervention called Sokoon. Our goal is to reduce symptoms of depression (mild to moderate), anxiety, and stress in adults by leveraging the field of computerized cognitive-behavioral therapy. We aim to provide the best results by incorporating techniques such as gamification, which can attract users and encourage them to seek treatment. To further enhance the effectiveness of Sokoon, we incorporated the hexad theory to allow for customization based on different user types, and a dynamic difficulty adjustment algorithm to improve efficiency and results. By adding more personalization, we believe that Sokoon can be more effective in addressing the unique needs of each individual and helping them manage their symptoms of depression, anxiety, and stress. Our hope is that Sokoon can serve as a useful tool in the larger effort to improve mental health outcomes and increase access to evidence-based treatments.

## II. Literature Review

Gamification has recently attracted more and more attention as a cutting-edge method of treating mental illness. Gamification techniques can be successful in enhancing mental health outcomes, such as decreasing symptoms of anxiety and depression [23]. A study found that gamified apps can reduce anxiety and enhance the impact of mobile interventions for health and well-being [24]. Another study suggests that the design of a gamified app, which incorporated game components like points, awards, and progress tracking, may have enhanced participants' motivation and engagement in meditation practice, contributing to the app's efficacy in reducing symptoms of depression [17]. Several studies have reported positive outcomes associated with gamified interventions for depression, anxiety, and stress,

including increased motivation, engagement, and treatment adherence. It is clear from empirical evidence that gamification has a positive impact on mental health domains and is a revolutionary field to explore [25].

A growing body of research has demonstrated the potential benefits of combining gamification and cognitive-behavioral therapy (CBT) to develop engaging and effective mental health interventions. Several studies have shown that gamified CBT interventions can be successful in reducing symptoms of anxiety [12], depression [14, 15], and stress [16]. However, there is limited research on the use of gamification and the Hexad theory in CBT applications. Only one study [26] has utilized the Hexad theory with gamification components, but for diagnosis rather than treatment. Sokoon is one of the first applications to integrate these elements in a CBT intervention.

Gamification has shown promise as a tool for treating depression, anxiety, and stress through cognitive-behavioral therapy (CBT). However, further research is needed to determine the most effective strategies for tailoring gamification elements to meet individual needs and preferences. To address this gap, this study proposes a gamified CBT approach for treating depression and anxiety that incorporates personalized gamification systems.

This approach aims to improve therapy outcomes by using gamification components such as challenges, badges, levels, and prizes that are tailored to each individual's personality. The approach includes a dynamic difficulty adjustment algorithm that adapts to each user's skill to optimize engagement and effectiveness.

## III. Methods

We have completed the pre-design phase for our application, which involved identifying our target audience as adults, with a particular focus on university students who may be more susceptible to depression and anxiety due to psychosocial factors [27]. During this phase, we consulted with psychiatrists to determine the necessary tasks and skills, and covered topics such as design, acoustics, visuals, and content creation. We also agreed on the name for the app, Sokoon, and developed the main layout, appearance, elements, and delivery method incrementally over several weeks of development.

The first prototype of Sokoon includes four modules: Gratitude, Relaxation, Behavior Activation, and Cognitive Restructuring. To encourage engagement, we have incorporated gamification features, applying the Hexad theory to increase customization for each user's personality type. We have also utilized a dynamic difficulty adjustment algorithm (DDA) to adaptively change the difficulty level of the game. Our approach involved creating mini-games, interactive workouts, and stories as part of an Android mobile application using the gamification features provided by a Unity (2D/3D) gaming engine [28]. Since the Unity engine is cross-platform, porting games to other platforms such as the web, PC, and iOS is easier.

The methods we used to incorporate gamification into CBT procedures for Sokoon are described and discussed in the following sections.

### A. Research Design

This study uses a single-group pre-post design to evaluate the impact of sokoon on the treatment outcomes of adults with DASDs. The Patient Health Questionnaire-9 (PHQ-9) and the Generalised Anxiety Disorder-7 (GAD-7) will be completed by participants at baseline (pre-intervention) and right after the intervention.

#### Research Question 1:

The first research question for this study is: "In what ways does the use of Sokoon impact the treatment outcomes of adults with DASDs?" This question seeks to understand how the Sokoon intervention affects the overall treatment outcomes of individuals with DASDs. The study will measure changes in various outcomes, such as symptom severity, before and after the Sokoon intervention using the PHQ-9 and GAD-7.

#### Research Question 2:

The second research question for this study is: "To what degree can sokoon reduce depression, anxiety, and stress symptoms?" This question specifically focuses on the impact of the sokoon intervention on symptoms of depression, anxiety, and stress. The study will measure changes in these symptoms before and after the intervention using the PHQ-9 and GAD-7 to determine the degree to which sokoon can effectively reduce these symptoms in adults with DASDs.

As described in more detail in the following sections, participants were chosen, instructed to use the program, and then results and feedback were gathered from their providers.

#### 1) SAMPLE SELECTION

A random sample of 30 adults over the age of 18 was selected to participate in the study as shown in Tab. 1. Participants were administered the Patient Health Questionnaire-9 (PHQ-9) [29] to assess their level of depression and the Generalized Anxiety Disorder-7 (GAD-7) [30] to measure their level of anxiety. We excluded 15 participants who reported mild to moderate levels of depression and anxiety and retained the remaining 15 participants who reported low levels of depression and anxiety.

The proposed model was then provided to the participants for a two-week trial period, as the psychiatrist overseeing the study recommended. At the end of the trial, results were collected manually via email. Five participants provided incomplete results, and their data were excluded from the analysis. The results of the remaining 10 participants were analyzed to assess the effectiveness of the proposed model.

TABLE I  
INFORMATION ABOUT PARTICIPANTS.

USERS	AGE	SEX	ANXIETY DEGREE	DEPRESSION DEGREE
USER1	21	F	MODERATE	MILD
USER2	35	M	MODERATE	MODERATE
USER3	21	F	MILD	MODERATE
USER4	21	F	MODERATE	MILD
USER5	25	F	MODERATE	MODERATE
USER6	25	F	MILD	MODERATE
USER7	25	F	MODERATE	MODERATE
USER8	20	F	MODERATE	MODERATE
USER9	21	F	MILD	MILD
USER10	32	F	MODERATE	MODERATE

#### 2) ETHICAL CONSIDERATIONS

Permission for the study was granted by the ethics committee at Mansoura University. Caregivers who agreed to participate in the study were provided with a detailed explanation of the study's purpose, methodology, risks, and benefits, and were asked to provide their informed consent before participating. Participants were informed that their participation was voluntary and that declining to participate would not have any negative consequences.

#### 3) DATA ANALYSIS

Means and standard deviations are examples of descriptive statistics, that were calculated to summarize the clinical characteristics of the sample, as well as the PHQ-9 and GAD-7 scores at each time point (baseline, post-intervention). The amount of the intervention impact will be determined using Cohen's d-effect size estimates. We used Cohen's guidelines for interpreting effect sizes, where an effect size of 0.2 is considered small, 0.5 is medium, and 0.8 or higher is large [31].

### B. THE THERAPEUTIC COMPONENTS

In developing our app, Sokoon, we utilized cognitive behavioral therapy (CBT), which is an evidence-based therapy for the treatment and management of depression and anxiety [5]. CBT can be divided into various components, including different skills and learning objectives [12]. In order to determine which skills to include in our app, we conducted extensive research and consulted with a psychiatrist. This process was challenging, but ultimately helped us to identify the most effective therapeutic components to incorporate into Sokoon.

### C. GAMIFICATION ELEMENTS AND HEXAD THEORY

We used the HEXAD framework to identify six user types, where there are several personalities into which the users are divided, the player, socializer, free sprite, achiever, Philanthropist, and disruptor, which makes the experience more customized to the user. There are several widely acknowledged key gamification components and hexad theory techniques that are covered in depth elsewhere [19,20].

In Tab. 2, we'll go through how we included gamification elements for every user personality based on hexad theory into our app, Sokoon.

TABLE 2  
GAMIFICATION ELEMENTS USED IN SOKOON.

Gamification elements	Used in Sokoon	User type
Points	Points are earned daily in each game and activity. Users can save them and spend them in "the home" game, avatar customization, and open message in message for your page.	Player
Badges	When finishing any planet.	Player
Customization	Users can customize their avatars.	Free sprite
Anonymity	The user has the ability to hide his name.	Disruptor
Anarchy	The user can burn all his progress on all planets at any time and start again.	Disruptor

Levels	In the “positive word” game.	Achiever
Random rewards	In all planets.	For all user type.

#### D. DYNAMIC DIFFICULTY ADJUSTMENT(DDA)

We added Dynamic Difficulty Adjustment (DDA) to a game that was played as a part of the intervention. The game's goal was to enhance cognitive abilities, and DDA was used to adjust the difficulty level to each player's performance and skill level.

Level difficulty can be dynamically changed at runtime using a method known as Reference Player's Difficulty (RPD) [21]. RPD is a particular method that falls within the DDA umbrella. To determine the beginning value of Difficulty with equation.1 in the PRD approach, the user should complete the first level. The difficulty for the user at the following level will be reduced if the  $RDP \geq 0.5$ ; otherwise, the difficulty will be increased.

$$RDP = \frac{(\text{The Best Score for the first level} - \text{User's score for the first level})}{\text{Best Score for the first level}}$$

Example: If the user's score for the first level is 15 and the best score is 30, then the RDB is calculated as follows.

$$RDP = \frac{(30 - 15)}{30} = 0.5$$

The next level's difficulty will then be lowered so that the user can become more involved.

The ideal scenario is to get an ease score of 0.5 or close to it, suggesting that the proposed task is appropriate for the current user.

We applied this on the positive planet. We gave more details on how to apply this in the positive planet in the following section.

#### E) THE PROPOSED MODEL(SOKOON)

Our proposed model aims to assist individuals suffering from depression, anxiety, and stress, with a specific focus on psychopaths, by utilizing cognitive behavioral therapy (CBT) skills such as cognitive restructuring, relaxation, self-love, socialization, behavior activation, gratitude, and problem-solving. We incorporate gamification techniques, the user hexad theory, and the DDA algorithm into our model to enhance its effectiveness.

In our prototype, we will introduce and explain four of the CBT skills that we have selected: cognitive restructuring, relaxation, behavior activation, and gratitude [32]. These skills were chosen based on their proven effectiveness in treating depression, anxiety, and stress, and are expected to have a positive impact on the target population.

To make the CBT skills more engaging and enjoyable, we designed them in the form of planets. Each skill was assigned a planet, which contains a set of activities and games related to that particular skill. Our goal was to create a fun and interactive experience for users and to prevent them from feeling bored or disengaged while using the app.

Let me explain each step of our Arabic language model designed to help individuals in Arabic countries dealing with depression,

anxiety, and stress, with the aid of figures. At every stage of the application, we collected and stored specific data for different purposes, which I will explain.

The first page of our application is the registration page, where we collect essential data such as the user's username, age, gender, and user avatar, as shown in Fig. 1(a). The age must be stated as eighteen years or older, and we save each user's data to aid us in behavior analysis. Once the user has entered their data, the application verifies that the age is over eighteen years old and that all data has been entered. Afterward, the user can proceed to the second page, which is the gamified user type test, as shown in Fig. 1(b). Once the registration phase is complete, the main page is loaded, as shown in Fig. 1(c).

we include a gamified user type test that is used by Andrzej Marczewski [19] to classify users in a gamified system. The test consists of 24 questions, and each question has seven options, each with a score ranging from strongly agree (3) to strongly disagree (-3). The questions are designed to evaluate the user's personality type in gaming, and they cover a range of topics, such as helping others, trying new things, following rules, being part of a community, mastering difficult tasks, and winning prizes. By answering the questions and scoring each option, the user's personality type can be classified into one of several categories, including Philanthropists, Socializers, Free Spirits, Achievers, Players, and Disruptors.

In our model, we incorporate various gamification elements such as badges, customization, levels, points, and prizes, which are listed in Tab. 2.

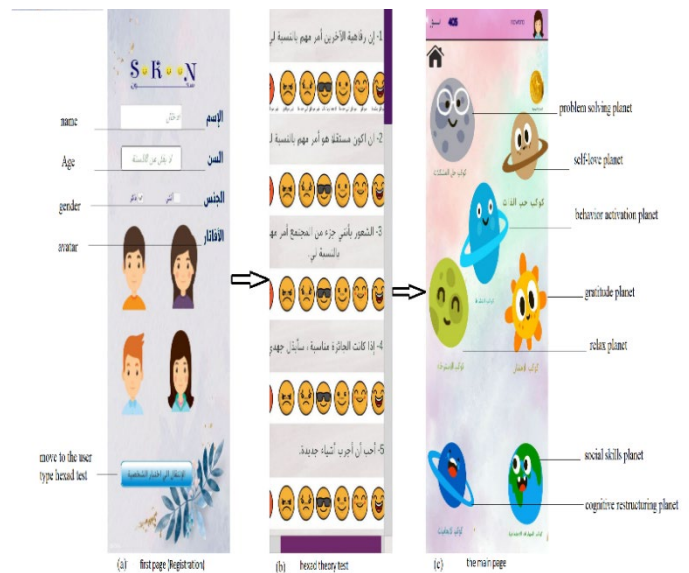


FIGURE 1. Registration pages and the main page of the app. (Design Credits: The author).

#### HOW IS THE GAMIFIED USER TYPE RESULT CALCULATED BASED ON HEXAD THEORY?

The Hexad theory was used to categorize users into six different types based on their motivation for playing games. As part of the registration process, each user was given a Hexad theory test

consisting of 24 questions (as shown in Fig. 1(b)) to determine their gamified user type. Fig. 2 displays the four questions pertaining to each user type.

For each user type (Socializer, Philanthropist, Free sprite, Achiever, Disruptor, Player), identify the four questions associated with that type.

- For each question in the Hexad theory test, assign a numerical value ranging from 1 to 7 to each possible answer (1 = strongly disagree, 7 = strongly agree).
- For each user, record their response to each question in the Hexad theory test.
- For each user type, sum the numerical values of the four questions associated with that type to obtain a total score for that type.
- Calculate the sum of the numerical values for all 24 questions in the Hexad theory test to obtain a total score.
- For each user type, divide the total score for that type by the total score for all 24 questions and multiply the result by 100 to obtain a percentage score.
- Display the percentage score for each user type to the user in the Sokoon application interface Fig. 3.

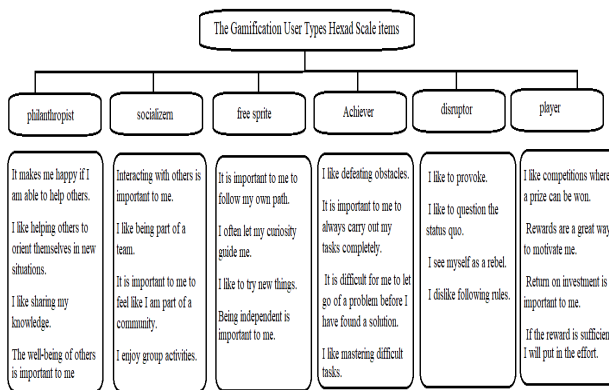


FIGURE 2. The Gamification User Types Hexad Scale items. Adapted from [20].



FIGURE 3. The result of the test. (Design Credits: The researcher).

The main page of the application is divided into three parts, as depicted in Fig. 1(c). At the top of the page, the user's name, avatar photo, and app coin called "nour" are displayed. The main page itself comprises seven planets, each serving a

specific purpose, which are the self-love planet, problem-solving planet, positive planet, relaxation planet, gratitude planet, activities planet, and social planet.

The main icon contains:

- Positive messages: the app features positive messages, and clicking on the message icon displays a new positive message every time.
- Depression test (PHQ-9) (Fig. 4): this test is not a screening tool for depression, but it is used to track the degree of depression and the effectiveness of its treatment [29].
- Anxiety test (GAD-7) (Fig. 5): this test is one of the strategies that can be employed to detect anxiety or evaluate its severity [30].
- Badges page (Fig. 6): this page displays all the badges that the user has earned. Users can earn more badges by playing more games.



FIGURE 4. Depression test [29]. (Design Credits: The researcher).



FIGURE 5. Anxiety test (GAD-7) [30]. (Design Credits: The researcher).





FIGURE 6. Badges page. (Design Credits: The researcher).

As a prototype, we are focusing on only four planets: the Positive Planet, the Relaxation Planet, the Gratitude Planet, and the Activities Planet.

### 1) THE POSITIVE PLANET

This planet focuses on cognitive restructuring skills. Negative emotions, along with their physical and behavioral effects, can be changed using the cognitive restructuring technique. This involves identifying the false negative beliefs that underlie negative emotions and replacing them with more positive coping concepts [33].

Based on this technique, we developed an activity that helps patients replace distorted thoughts with positive ones when they are exposed to negative situations. Figure 7 illustrates the steps of thought analysis used in our app, which are explained in more detail in Appendix A.

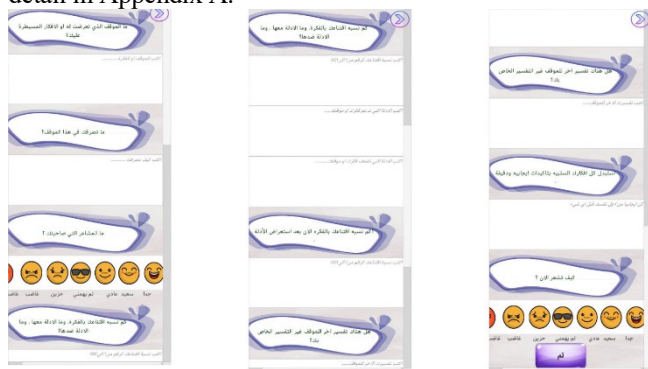


FIGURE 7. Record thoughts on the positive planet. (Design Credits: The author) Appendix A.

This planet also features a game called the Positive Word Game (see Figure 8). The game presents a collection of positive and negative words within a limited time, after which the words disappear and a new collection of words appears. The player's goal is to collect as many positive words as possible before they disappear. To enhance the player's experience, we applied a DDA algorithm in the game.

### HOW IS DDA APPLIED ON THE POSITIVE PLANET?

We implemented a dynamic difficulty adjustment (DDA) algorithm in our gamified CBT application to personalize and

adapt the difficulty level of the game based on the user's performance and preferences. We used Reference Player's Difficulty (RPD) to calculate the initial difficulty level after the user played the first level and divided the resulting range into four ranges based on the user's performance.

The Algorithm steps:

- Calculate the Reference Player's Difficulty (RPD) for the first level based on the user's performance using Eq.1.
- Divide the resulting range (difficulty) into four ranges: 0 to .25, .25 to .5, .5 to .75, and .75 to 1.
- If the RPD falls in the first range (0 to .25), increase the number of words or the time limit for the next level to make it harder.
- If the RPD falls in the second range (.25 to .5), move the user to the next level without any adjustments.
- If the RPD falls in the third range (.5 to .75), ask the user to replay the same level for more practice without moving to the next level.
- If the RPD falls in the fourth range (.75 to 1), ask the user to replay the same level with an increased time limit to make it easier.
- Repeat steps 1-6 for each subsequent level.
- Collect data on the user's performance, satisfaction, and mental health outcomes using surveys.
- Use the collected data to evaluate the effectiveness of the DDA algorithm over a period of two weeks.

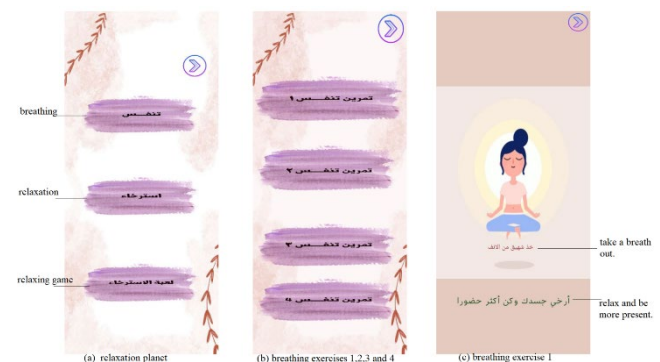


FIGURE 8. The positive word game. (Design Credits: The author)

### 2) RELAXATION PLANET

This planet focuses on relaxation skills and includes breathing exercises Fig. 9(c), relaxation videos, and a relaxation game. The relaxation game involves collecting a series of stars while listening to relaxing music Fig. 10(b).

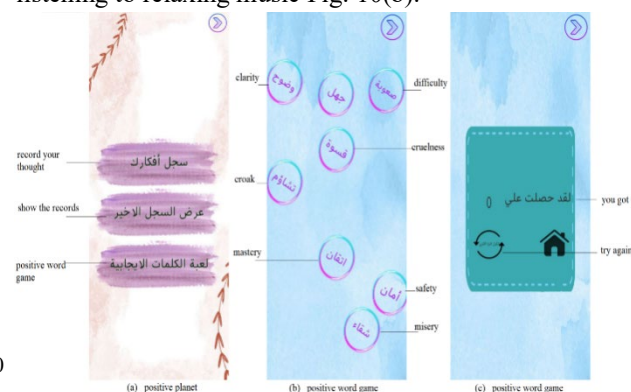




FIGURE 9. Breathing exercises in the relaxing planet. (Design Credits: The authors).



FIGURE 10. relax game in relax planet. (Design Credits: The authors).

### 3) GRATITUDE PLANET

This planet features three games, including the examples shown in Fig. 11 (b and c), which are designed to help users focus on positive aspects of their lives, such as blessings, family, and good memories. Each game involves collecting a series of flowers using a butterfly. When the butterfly lands on a flower, a gratitude sentence appears, and the player earns points. To achieve a high score, the player must collect all of the flowers without missing any.



FIGURE 11. gratitude game. (Design Credits: The authors).

### 4) BEHAVIOUR ACTIVATION PLANET

Individuals with depression often experience reduced interest in routine activities and a decreased capacity for pleasure [34]. To address this issue, we have incorporated a simple feature into our app that suggests different activities for the user to try (Fig. 12). These suggestions can help users discover new hobbies and interests, potentially improving their mood and overall sense of well-being.

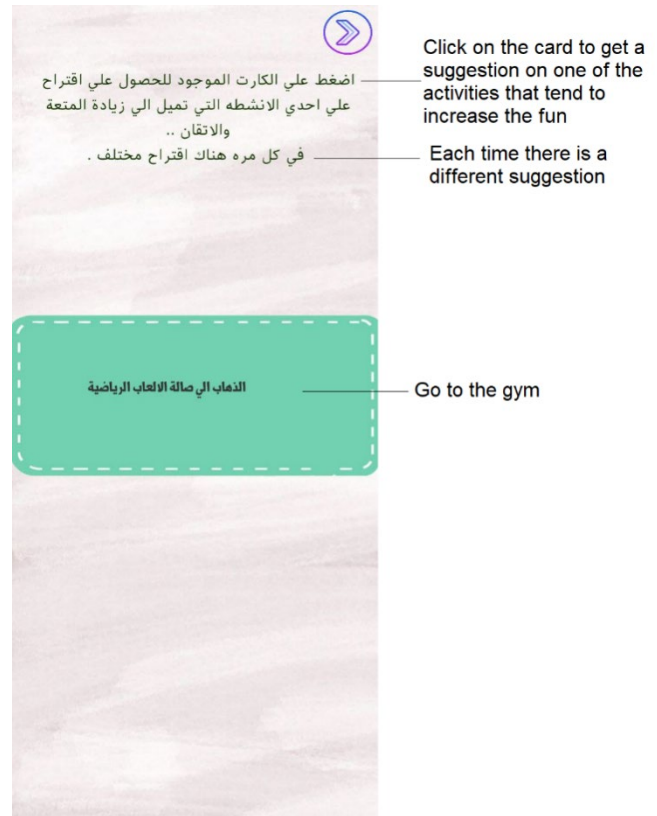


FIGURE 12. behavior activation planet. (Design Credits: The authors)

## IV. Results

### A. The usability metrics results (based on Nielsen Norman Group).

The usability metrics for Task 1 (Registration) and Task 2 were calculated according to the Nielsen Norman Group [35]. For Task 1, the Task Success Rate was 100%, indicating that all users who attempted to register were able to finish the task effectively. The User Satisfaction for Task 1 was 5.3, which is the average satisfaction rating given by users after completing the registration task on a scale of 1 to 7. The Average Task Time for Task 1 was 2.23 minutes, which is the average time taken by users to complete the registration task. The Time-Based Efficiency for Task 1 was 52%, which measures the percentage of time users spent actively completing the task as opposed to waiting for the system to respond or load. The Average Error Occurrence Rate for Task 1 was 0.05, which is the average number of errors encountered per user while attempting to complete the registration task.

For Task 2, the Success Score was 81.5, which is a measure of the overall success rate of the task, taking into account both completed and partially completed attempts. The Average Task Time for Task 2 was 1.85 minutes, which is the average time taken by users to complete the task. The Average Error Occurrence Rate for Task 2 was 0.13, which is the average number of errors encountered per user while attempting to complete the task. The Time-Based Efficiency for Task 2 was 43%, which measures the percentage of time users spent actively completing the task as opposed to waiting for the system to respond or load. The Average Satisfaction (SEQ) for

Task 2 was 5.7, which is the average satisfaction rating given by users after completing the task on a scale of 1 to 7, using the Single Ease Question (SEQ) method. Fig. 13 summarizes the usability metrics applied.

Overall, the findings show that users were able to successfully complete both activities with high success rates and low error occurrence rates. Additionally, the users were able to finish the jobs quickly based on the low average task times. The Time-Based Efficiency for both jobs, however, was comparatively poor, indicating that consumers had to wait a long time for the system to reply or load. Both exercises had above-average User Satisfaction scores, demonstrating that users were generally happy with their experience performing the assignments.

Usability Metrics	The first task (Registration)	The second task (Gratitude Game)
Effectiveness	100%	81.5%
Average error occurrence rate	.05	.13
Average task time	2.23 min	1.85 min
Time based efficiency	52% goals/min	.43% goals/min
Average Satisfaction (SEQ)	5.3	5.7

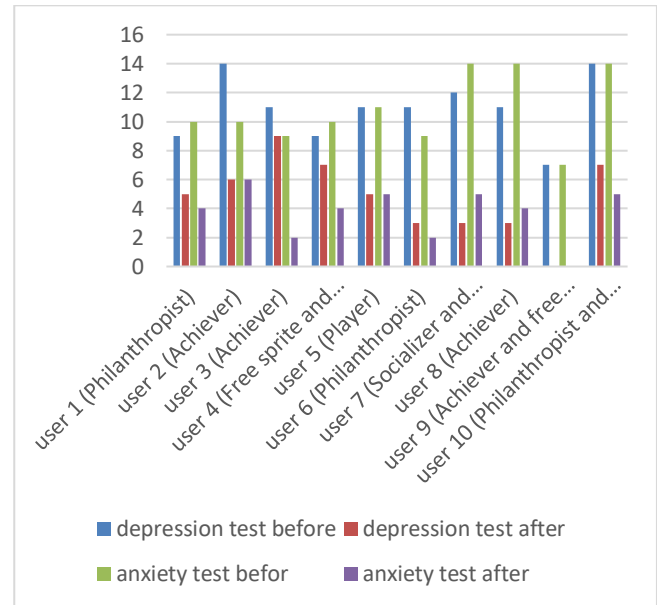
FIGURE 13. Usability metrics of Sokoon.

### B. Participants Results.

The average participant age was 24.6 years ( $SD = 5.12$ ). The results showed that Sokoon's participants had a significant reduction in symptoms of depression and anxiety after using the application as shown in Fig. 14. with a large effect ( $d=2.5$ ,  $d=3.3$ ) for depression and anxiety based on the PHQ-9 test and GAD-7 test results. From the pretest to the posttest, the participants' anxiety symptoms were less severe ( $M = 10.8$ ,  $SD = 2.44$  vs.  $M = 3.7$ ,  $SD = 1.8$ ). Likewise, their depressed symptoms decreased (pretest:  $M = 10.9$ ,  $SD = 2.18$ ), (posttest:  $M = 4.8$ ,  $SD = 2.61$ ).

The Hexad theory test was used to improve the gamification features of a CBT application designed to promote mental health and well-being. The test was administered during the registration process and used to categorize users into six Hexad types based on their motivation for playing games.

The results of the study showed that gamification features tailored to each Hexad type can improve user engagement, motivation, and retention in the CBT exercises. The majority of users belonged to the "Achiever" and "Philanthropist" user types Fig. 14, demonstrating their motivation to utilise the service through collaborative engagement and achievement.



### A. Performance Metrics

Based on how well the players performed, the game's difficulty level was modified using the DDA algorithm. The initial level of difficulty was set based on the performance of a reference player who had achieved the target level of performance. The algorithm then adjusted the difficulty level based on the performance of each participant.

Our results showed that both the DDA algorithm were effective in adapting the difficulty level of the game to each user's needs and preferences. and resulted in higher levels of motivation, engagement, and satisfaction with the game mechanics and content. We noticed that Participants spend more time in positive planet game over other planets. By applying the RPD to our game, we can potentially create a more personalized and engaging experience for our users, which can help to promote learning and skill development. We also found that the users showed improvements in their mental health outcomes over the course of the study.

Overall, these results suggest that the sokoon intervention may be an effective treatment for reducing symptoms of anxiety, stress, and depression but to verify these results and investigate the intervention's long-term consequences, additional study is required.

### B. Comparison with similar apps

In Tab. 3, a comparison was made with other applications that applied gamification in their interventions. Sokoon had a significant effect size compared to them, as it applied a DDA algorithm in its intervention for the first time, highlighting the potential benefits of using such techniques in interventions.

TABLE 3  
COMPARISON OF EFFECT SIZES IN INTERVENTIONS USING GAMIFICATION:  
SOKOON VS. OTHER APPLICATIONS.

The study	N	Cohen's d (Effect size)	
		Depression	Anxiety
Sokoon	10	2.5	3.3
MTPhonix [14]	40	1.02	N/A
Sparx [15]	50	.6	N/A

mindfulness meditation [17]	30	.28	N/A
--------------------------------	----	-----	-----

## V. Discussion

Mental illnesses such as depression, anxiety, and stress have spread widely among adults, and for several reasons, there is no interest in going to a psychiatrist, including that psychological treatment is expensive and requires a lot of time, follow-up, and feelings of shame, which made the problem exacerbate. The mental illness may be in an early stage from mild to moderate, and lack of interest makes it get worse. Sokoon is an app for treating patients with mild to moderate anxiety, depression, and stress based on gamified CBT. This does not mean replacing psychiatrists, but it is a quick and helpful solution. Compared to previous studies that applied cognitive-behavioral therapy and its skills, we added more skills and applied the Hexad theory to make the gamification experience more customized, unlike previous studies that were content with gamification on a regular basis. We also applied dynamic difficulty adjustment Algorithm to one of the games to make the experience far from boring or difficult. To our knowledge, this is the first study of its kind that applies the Hexad theory and combines DDA with CBT to reduce symptoms of stress, anxiety, and depression.

A study was carried out to examine the function of CBT in the management of mental disorders., the most applied skills, and how to apply them to obtain the best results, with some notes obtained by asking the psychiatrist and then these observations were incorporated to be applied in the application. The experiment was conducted on a sample of adults with mild to moderate degrees of depression and anxiety. The results of the study indicate a decrease in the percentage of depression and anxiety. A usability metrics was made on the data that was collected using the application for the volunteer category by collecting it by sending it via e-mail while people used the application, and it was discovered that the design was effective and obtained high user satisfaction. In short, sokoon reduced the symptoms associated with depression and mild to moderate anxiety, and applied gamification and the hexad theory increased interaction and engagement.

In response to the first research query that looked at how Sokoon impacted adults with depression and anxiety, it was noted that the gamification feature led to increased engagement and motivation. The combination of hexad theory with gamification provided a comprehensive framework for delivering treatment according to the individual's specific needs.

To answer the second research question, to what extent does Sokoon reduce symptoms of depression and anxiety? We made the volunteers do depression and anxiety tests (PHQ-9 and GAD-7) before and after using the application, and we collected the results as shown in Figure 17. This shows that Sokoon had a major role in making a difference in the psychological state of the volunteers.

comments from volunteers included that they liked the shorts game, the design, the music we used, and the procedure for obtaining prizes and badges. Some also say that using the application helped them improve their mood in real time and

loved using it.

During the testing procedure, Sokoon's drawbacks were discovered. Our system only supports the Arabic language, as we took care of introducing it to Arab countries, future versions can add more languages to make the app widely used. Because this study uses a single-group pre-post design, there is a lack of a control group. In future research, a randomized controlled trial with a control group could be used to further investigate the effectiveness of this intervention. Not all gamification elements have been applied due to the difficulty of applying them, such as: leaderboards, sharing to social media, teams, and others. As a prototype, we did not activate all the cognitive-behavioral therapy skills that we mentioned and there were a few games, as future versions can activate all the skills, which increases the improvement of cases and gives the experience a lot of pleasure and adds more games that make the experiment more fun. As the application includes some but not all components of evidence-based CBT, it cannot be independently relied upon to effectively recover from depression, anxiety, and depression, in the future, we can further enhance the application by incorporating additional CBT skills and techniques. The sample used was close in age and most of them were female. In the future, it is possible to apply it to a larger sample with more different age groups. A larger sample size would be needed to confirm the findings of this study and to generalize the results to the larger population. While our DDA algorithm was effective in adapting the difficulty level of the game based on the user's performance and preferences, there is still room for improvement. Future research could explore more advanced algorithms, such as deep reinforcement learning, to learn more complex and nuanced patterns in the user's behavior and provide a more personalized and adaptive experience. The psychiatrist can be involved in the application, where patients who want more treatment can communicate and follow up with the psychiatrist, and the psychiatrist can follow the patient's page to see the progress in his psychological condition. Despite these drawbacks, the study will still be helpful to discover more about the possible advantages of Sokoon for treating people's symptoms of stress, anxiety, and depression.

## VI. Conclusions

This paper's objective is to review the supporting data for the efficacy of CBT in treating DASDs and to explore the potential for gamifying CBT to enhance its efficacy. We describe the approach we have taken in designing sokoon, a mobile application that applies CBT skills as a set of planets and uses gamification to increase adult engagement and applies hexad theory to increase customization with dynamic difficulty adjustment to help adults with DASDs. Gamifying CBT appears to increase engagement and motivation, and to reduce the time and cost of treatment. Gamified CBT interventions that target hexad theory may be more effective than traditional CBT interventions. Hexad theory provides a framework for understanding how game elements can be used to motivate people. When combined with CBT, gamification can be used to improve engagement with the treatment and to accelerate the treatment process. This event could lead to greater adherence to CBT, more effective treatment outcomes, and improved quality of life for people suffering from depression, anxiety, and stress.

By harnessing the power of gamification and the hexad theory, CBT can be used more effectively to help people overcome these barriers and live healthier lives.

Evaluation of the use of this application by adults felt it was practical and simple to use, and the application showed effective results in improving the psychological state of adults after using the application, which was determined by the depression and anxiety scale (PHQ-9, GAD-7) before and after using the application. We also used usability metrics to assess the efficiency, effectiveness, and satisfaction of adults who used the app. This shows how easy and effective the app is for users as the results show. Outside of scheduled therapy sessions, sokoon expanded access to evidence-based CBT techniques in a format that was well-liked and utilized by adults. We used several techniques we thought would result in a better outcome such as gamification, hexad theory and DDA. To create a more engaging experience, we used the Hexad theory for personalizing gamified systems to users' personalities. This added a type of customization for each user, as the appropriate gamification elements were selected for each user's personality. The usefulness of sokoon should also be investigated for additional outcomes, such as suicidal ideation, and in understudied populations, such as older persons. Further research is needed to investigate the mechanisms by which gamification enhances CBT for depression, anxiety and stress, and to develop and test more effective gamified CBT interventions. It is anticipated that this technology will advance in the next years, making it easier for people to obtain therapies in the manner that most suits them.

In conclusion, our study has provided a foundation for future research and development in gamified CBT applications using DDA and other advanced algorithms. We hope that our findings will inspire further exploration and innovation in this field to improve the mental health outcomes of individuals suffering from depression, anxiety, and stress.

#### ACKNOWLEDGMENT

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2023-00218176).

#### APPENDIX A

Thought recording steps in the positive planet. [from Psychologist]

1. What situation have you been exposed to or the idea controlling you?
2. What do you do in this situation?
3. What are the feelings that accompanied you?
4. Rate the intensity of your feelings from 1 to 10.
5. How much do you believe in the idea? And what is the evidence behind it? What is the evidence against it?
6. After reviewing the evidence, how much do you believe in the idea?
7. Is there another explanation for the situation other than your own?

8. Replace all your negative thoughts with positive, accurate affirmations.
9. How are you feeling now? Rate your feelings from 1 to 10.

#### REFERENCES

- [1] Larry Gamm, S. Stone, and S. Pittman, "MENTAL HEALTH AND MENTAL DISORDERS—A RURAL CHALLENGE: A LITERATURE REVIEW," *Rural healthy people*, vol. 1, no. 2, pp. 97–114, 2010.
- [2] I. A. Alshawwa, M. Elkahoul, H. Qasim El-Mashharawi, and S. S. Abu-Naser, "An Expert System for Depression Diagnosis," *International Journal of Academic Health and Medical Research (IJAHMR)*, vol. 3, no. 4, pp. 20–27, 2019.
- [3] B. N. Gaynes, K. M. Magruder, B. J. Burns, H. Ryan, Wagner, K. S. H. Yarnall, and W. Eugene. Broadhead, "Does a coexisting anxiety disorder predict persistence of depressive illness in primary care patients with major depression?," *General Hospital Psychiatry*, vol. 21, no. 3, pp. 158–167, May 1999, doi: [https://doi.org/10.1016/s0163-8343\(99\)00005-5](https://doi.org/10.1016/s0163-8343(99)00005-5).
- [4] K.-S. Kim and P.-L. Han, "Optimization of chronic stress paradigms using anxiety- and depression-like behavioral parameters," *Journal of Neuroscience Research*, vol. 83, no. 3, pp. 497–507, Feb. 2006, doi: <https://doi.org/10.1002/jnr.20754>.
- [5] C. Williams and M. Chellingsworth, *CBT : a clinician's guide to using the five areas approach*. London: CRC Press, 2010.
- [6] E. Kaltenthaler *et al.*, "Computerised cognitive behaviour therapy for depression and anxiety update: a systematic review and economic evaluation," *Health Technology Assessment*, vol. 10, no. 33, Sep. 2006, doi: <https://doi.org/10.3310/hta10330>.
- [7] M. So, S. Yamaguchi, S. Hashimoto, M. Sado, T. A. Furukawa, and P. McCrone, "Is computerised CBT really helpful for adult depression?—A meta-analytic re-evaluation of CCBT for adult depression in terms of clinical implementation and methodological validity," *BMC Psychiatry*, vol. 13, no. 1, Apr. 2013, doi: <https://doi.org/10.1186/1471-244x-13-113>.
- [8] K. D. Vallury, M. Jones, and C. Oosterbroek, "Computerized Cognitive Behavior Therapy for Anxiety and Depression in Rural Areas: A Systematic Review," *Journal of Medical Internet Research*, vol. 17, no. 6, p. e139, Jun. 2015, doi: <https://doi.org/10.2196/jmir.4145>.
- [9] A. Miloff, A. Marklund, and P. Carlbring, "The challenger app for social anxiety disorder: New advances in mobile psychological treatment," *Internet Interventions*, vol. 2, no. 4, pp. 382–391, Nov. 2015, doi: <https://doi.org/10.1016/j.invent.2015.08.001>.
- [10] S. Deterding, D. Dixon, R. Khaled, and L. Nacke, "From game design elements to gamefulness," *Proceedings of the 15th International Academic MindTrek Conference on Envisioning Future Media Environments - MindTrek '11*, pp. 9–15, 2011, doi: <https://doi.org/10.1145/2181037.2181040>.
- [11] S. G. Six, K. A. Byrne, T. P. Tibbett, and I. Pericot-Valverde, "Examining the Effectiveness of Gamification in Mental Health Applications for Depression: A Systematic Review and Meta-Analysis (Preprint)," *JMIR Mental Health*, vol. 8, no. 11, Jul. 2021, doi: <https://doi.org/10.2196/32199>.
- [12] G. I. Christie, M. Shepherd, S. N. Merry, S. Hopkins, S. Knightly, and K. Stasiak, "Gamifying CBT to deliver emotional health treatment to young people on smartphones," *Internet Interventions*, vol. 18, p. 100286, Dec. 2019, doi: <https://doi.org/10.1016/j.invent.2019.100286>.
- [13] D. Bakker, N. Kazantzis, D. Rickwood, and N. Rickard, "Mental Health Smartphone Apps: Review and Evidence-Based Recommendations for Future Developments," *JMIR Mental Health*, vol. 3, no. 1, p. e7, Mar. 2016, doi: <https://doi.org/10.2196/mental.4984>.
- [14] C. A. Lukas, B. Eskofier, and M. Berking, "A Gamified Smartphone-based Intervention for Depression: Randomized Controlled Pilot Trial. (Preprint)," *JMIR Mental Health*, Oct. 2019, doi: <https://doi.org/10.2196/16643>.
- [15] K. Yokomitsu *et al.*, "Gamified Mobile Computerized Cognitive Behavioral Therapy for Japanese University Students With Depressive Symptoms: Protocol for a Randomized Controlled Trial," *JMIR Research Protocols*, vol. 9, no. 4, p. e15164, Apr. 2020, doi: <https://doi.org/10.2196/15164>.
- [16] L. P. S. Dias, J. L. V. Barbosa, L. P. Feijó, and H. D. Vianna, "Development and testing of iAware model for ubiquitous care of

- patients with symptoms of stress, anxiety and depression,” *Computer Methods and Programs in Biomedicine*, vol. 187, p. 105113, Apr. 2020, doi: <https://doi.org/10.1016/j.cmpb.2019.105113>.
- [17] M. T. Fish and A. D. Saul, “The Gamification of Meditation: A Randomized-Controlled Study of a Prescribed Mobile Mindfulness Meditation Application in Reducing College Students’ Depression,” *Simulation & Gaming*, vol. 50, no. 4, p. 104687811985182, Jun. 2019, doi: <https://doi.org/10.1177/1046878119851821>.
- [18] J. Krath and H. F. O. von Korflesch, “Player Types and Game Element Preferences: Investigating the Relationship with the Gamification User Types HEXAD Scale,” *Lecture Notes in Computer Science*, pp. 219–238, 2021, doi: [https://doi.org/10.1007/978-3-030-77277-2\\_18](https://doi.org/10.1007/978-3-030-77277-2_18).
- [19] Andrzej Marczewski, *Even Ninja Monkeys Like to Play : gamification, game thinking and motivational design*, Illustrated. CreateSpace Independent Publishing Platform, 2015, p. 220.
- [20] G. F. Tondello, R. R. Wehbe, L. Diamond, M. Busch, A. Marczewski, and L. E. Nacke, “The Gamification User Types Hexad Scale,” *Proceedings of the 2016 Annual Symposium on Computer-Human Interaction in Play*, pp. 229–243, Oct. 2016, doi: <https://doi.org/10.1145/2967934.2968082>.
- [21] G. K. Sepulveda, F. Besoain, and N. A. Barriga, “Exploring Dynamic Difficulty Adjustment in Videogames,” *2019 IEEE CHILEAN Conference on Electrical, Electronics Engineering, Information and Communication Technologies (CHILECON)*, Nov. 2019, doi: <https://doi.org/10.1109/chilecon47746.2019.8988068>.
- [22] M. Zohaib, “Dynamic Difficulty Adjustment (DDA) in Computer Games: A Review,” *Advances in Human-Computer Interaction*, vol. 2018, pp. 1–12, Nov. 2018, doi: <https://doi.org/10.1155/2018/5681652>.
- [23] R. Damaševičius, R. Maskeliūnas, and T. Blažauskas, “Serious Games and Gamification in Healthcare: A Meta-Review,” *Information*, vol. 14, no. 2, p. 105, Feb. 2023, doi: <https://doi.org/10.3390/info14020105>.
- [24] S. Litvin, R. Saunders, M. A. Maier, and S. Lüttke, “Gamification as an approach to improve resilience and reduce attrition in mobile mental health interventions: A randomized controlled trial,” *PLOS ONE*, vol. 15, no. 9, p. e0237220, Sep. 2020, doi: <https://doi.org/10.1371/journal.pone.0237220>.
- [25] N. Sinha, “Introducing Gamification for Advancing Current Mental Healthcare and Treatment Practices,” *IoT in Healthcare and Ambient Assisted Living*, pp. 223–241, 2021, doi: [https://doi.org/10.1007/978-981-15-9897-5\\_11](https://doi.org/10.1007/978-981-15-9897-5_11).
- [26] N. F. Jamaludin, T. S. M. Tengku Wook, S. F. Mat Noor, and F. Qamar, “Gamification Design Elements to Enhance Adolescent Motivation in Diagnosing Depression,” *International Journal of Interactive Mobile Technologies (IJIM)*, vol. 15, no. 10, p. 154, May 2021, doi: <https://doi.org/10.3991/ijim.v15i10.21137>.
- [27] L. M. Farrer, A. Gulliver, K. Bennett, D. B. Fassnacht, and K. M. Griffiths, “Demographic and psychosocial predictors of major depression and generalised anxiety disorder in Australian university students,” *BMC Psychiatry*, vol. 16, no. 1, Jul. 2016, doi: <https://doi.org/10.1186/s12888-016-0961-z>.
- [28] “Download Unity!,” *Unity*, 2019. <https://unity3d.com/get-unity/download>
- [29] K. Kroenke, R. L. Spitzer, and J. B. W. Williams, “The PHQ-9: Validity of a brief depression severity measure,” *Journal of General Internal Medicine*, vol. 16, no. 9, pp. 606–613, Sep. 2001, doi: <https://doi.org/10.1046/j.1525-1497.2001.016009606.x>.
- [30] R. L. Spitzer, K. Kroenke, J. B. W. Williams, and B. Löwe, “A Brief Measure for Assessing Generalized Anxiety Disorder,” *Archives of Internal Medicine*, vol. 166, no. 10, pp. 1092–1097, May 2006, doi: <https://doi.org/10.1001/archinte.166.10.1092>.
- [31] G. M. Sullivan and R. Feinn, “Using effect size—or why the P value is not enough,” *Journal of Graduate Medical Education*, vol. 4, no. 3, pp. 279–282, Sep. 2012, Available: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3444174/>
- [32] N. A. Amer, Samaa Mohammed Shohieb, W. M. Eladrosy, Hazem Mokhtar Elbakry, and Samir M. Abd Elrazek, “Sokoon,” *International Journal of Gaming and Computer-Mediated Simulations (IJGMS)*, vol. 15, no. 1, pp. 1–26, Jun. 2023, doi: <https://doi.org/10.4018/ijgms.324098>.
- [33] J. D. Otis, *Managing chronic pain : a cognitive-behavioral therapy approach : workbook*. Oxford: Oxford University Press, 2007.
- [34] B. Szygula-Jurkiewicz, A. Duszańska, and L. Poloński, “Is depression a problem in patients with chronic heart failure?,” *Polish Archives of Internal Medicine*, vol. 118, no. 1–2, pp. 52–56, Jan. 2008, doi: <https://doi.org/10.20452/pamw.304>.
- [35] “7 Essential Usability Metrics and How to Use Them,” *www.eleken.co*. <https://www.eleken.co/blog-posts/usability-metrics>

# Comparison of Brain Activation According to Sound and Motion Visibility in Videos During Combined Action Observation and Motor Imagery

Si-An Lee<sup>1</sup>, Yunyoung Nam<sup>2</sup>, Seong A Lee<sup>3</sup>, and Jin-Hyuck Park<sup>3\*</sup>

<sup>1</sup>Department of ICT convergence, The Graduate School, Soonchunhyang University, Asan, Republic of Korea

<sup>2</sup>Department of Computer Science and Engineering, Soonchunhyang University, Asan, Republic of Korea

<sup>3</sup>Department of Occupational Therapy, College of Medical Science, Soonchunhyang University, Asan, Republic of Korea

\*Contact: jhpark1217@sch.ac.kr, phone +82-41-530-4773

**Abstract—** This study investigated variation in brain activation during action observation with motor imagery (AOMI) by examining the influence of sound and motion visibility, utilizing functional near-infrared spectroscopy (fNIRS). The final analysis included 28 healthy participants who observed videos on handwashing and mentally simulated the depicted actions. AOMI showed that the greatest number of activated channels occurred under the video condition with sound and unrestricted motion visibility up to the elbow. Furthermore, videos without sound exhibited higher peak values and reached peak activation more rapidly compared to videos with sound. By comparing brain activation under different video conditions, this study could provide insights into optimal video conditions to facilitate brain activation, which could be utilized in future rehabilitation interventions employing AOMI and offering more effective video conditions.

## I. INTRODUCTION

Action observation involves purposefully watching meaningful actions in videos, intending to imitate them [1, 2]. Motor imagery is defined as a dynamic state where individuals mentally engage in a specific action, experiencing it as if they were physically executing it [3]. Similar brain regions, including the bilateral premotor cortex and parietal lobes, are activated during both action observation and motor imagery [4]. While traditionally studied independently or comparatively [5], recent research increasingly underscores the combined effects of action observation and motor imagery [6].

Combined action observation and motor imagery (AOMI) entails the simultaneous observation of actions presented in videos and the mental imagining of the effort and sensations required to perform those actions [7, 8]. This methodology is employed in treating individuals with impaired motor functions, such as those with Parkinson's disease or developmental coordination disorders, and its intervention effects have been validated [9]. AOMI research primarily focuses on the intervention effects in clinical groups or refining the combination of methods [6]. Despite AOMI involving observation through videos, there is limited research on video parameters compared to action observation.

The parameters of videos are essential since they are linked to the most efficient modulation methods for stimulating the motor system [10]. Additionally, in research comparing first-person and third-person perspectives in action observation, it has been noted that the extent of brain activation varies based on video parameters, underscoring the importance of these parameters [11].

In this investigation, our objective was to explore video parameters, specifically focusing on the presence of sound and motion visibility. Despite the prevalent use of videos with sound in the majority of action observation studies [12], there is a scarcity of research on sound as a video parameter. According to a previous study, sound plays a role in enhancing motor learning by establishing a robust auditory-motor coupling [13]. Furthermore, in a study on action observation for gait rehabilitation, significant performance improvements were observed when participants viewed videos with sound compared to those without sound [14]. However, contrasting findings emerged in an electroencephalography (EEG) study, where videos without sound were found to be more effective in activating the mirror neuron system [15]. The activated brain areas may undergo meaningful functional changes due to neuroplasticity during the intervention period [16]. In conclusion, existing research has not produced consistent results.

Moreover, recent studies utilizing transcranial magnetic stimulation (TMS) have revealed an enhancement in cortical-spinal excitability when the observer's gaze is allowed to move freely, as opposed to a fixed gaze condition [17]. Ito et al. [18] validated that restricted motion visibility was more efficacious in increasing cortical-spinal excitability compared to unrestricted motion visibility in their investigation focused on walking activities. However, there is currently no study exploring the disparity in brain activation between restricted and unrestricted motion visibility when observing upper limb activities. Hence, to discern effective video parameters, it is imperative to investigate the variations in brain activation associated with the presence of sound and motion visibility in videos.

On the other hand, the TMS or EEG studies discussed earlier involve simple actions like thumb opposition, diverging from the intricate, daily-life movements emphasized in rehabilitation interventions. The latter frequently adopts task-oriented approaches that concentrate on multi-joint movements [19]. Consequently, there is a demand for research on activities within a task-oriented context. On another note, brain imaging techniques such as EEG, functional magnetic resonance imaging (fMRI), and functional near-infrared spectroscopy (fNIRS) are employed to observe brain activation. Among these, fNIRS stands out as a non-invasive device for monitoring hemodynamic responses, recognized for its portability and exceptional temporal resolution, which makes it widely utilized in measuring brain activation [20].



Hence, in this study, fNIRS was utilized to assess and compare the influence of sound and motion visibility in AOMI on brain activation. The objective was to pinpoint the most effective video parameter for enhancing brain activation. The outcomes have the potential to validate effective video parameters and underscore their significance. Moreover, insights into the optimal application of AOMI can offer guidance for effective interventions in clinical settings. By focusing on the upper limbs, unlike previous research, this study contributes to the field of upper limb motor rehabilitation research.

## II. METHODS

### A. Participants

This study involved the participation of thirty healthy adults (mean age = 21.1 years, 20 females (66.7%)). Inclusion criteria stipulated that participants must have had no impairment in vision or hearing and should be capable of understanding linguistic instructions. Exclusion criteria encompassed individuals diagnosed with psychiatric or neurological disorders and those with an aversion to computer use.

### B. Stimuli

In this study, the video employed for AOMI focused on activities of daily living (ADLs) and specifically featured a recording of the handwashing activity corresponding to basic activities of daily living (BADLs). The handwashing activity video adheres to the guidelines provided by the Korea Disease Control and Prevention Agency for 'Proper Handwashing' [21], with each video having a duration of 30 seconds.

The videos in this study are categorized into four conditions: (1) Videos with sound and restricted motion visibility up to the wrist (Sound & Wrist; SW); (2) Videos with sound and unrestricted motion visibility up to the elbow (Sound & Elbow; SE); (3) Videos without sound and restricted motion visibility up to the wrist (No sound & Wrist; NSW); and (4) Videos without sound and unrestricted motion visibility up to the elbow (No sound & Elbow; NSE) (Fig. 1).

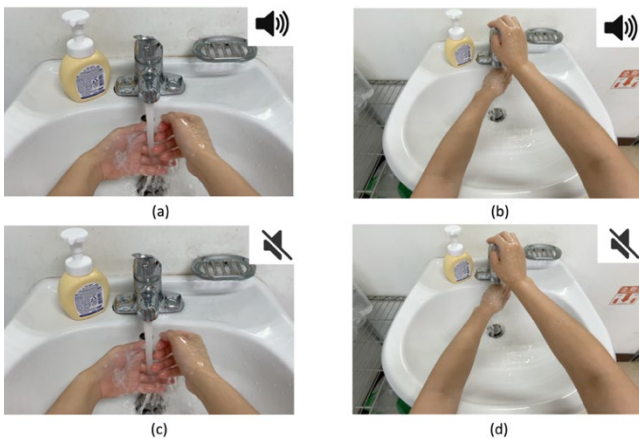


Fig. 1 The Video conditions. (a) SW; (b) SE; (c) NSW; (d) NSE

### C. Procedure

In a relaxed state, participants observed videos under four conditions while wearing fNIRS. Throughout video observation, participants were guided to mentally simulate performing the actions portrayed in the videos. Prior to video observation, a 10-second baseline was recorded. During this baseline measurement, participants were instructed to rest while focusing on a fixation cross displayed on the computer monitor. Each handwashing video had a duration of 30 seconds, followed by a 30-second rest period between videos and a 1-second preparation time before the commencement of each video (Fig. 2). The sequence of the videos was randomly assigned by generating a random number using the Python programming language from a pool of 24 different videos with distinct sequences.

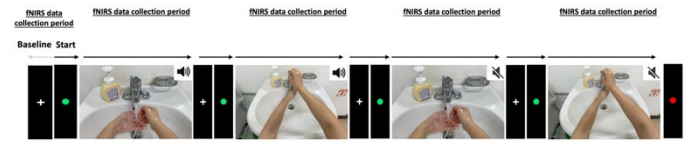


Fig. 2 Stimulus timing

### D. Outcome measurement

The assessment of brain activation involved measuring eight channels strategically positioned in areas associated with action observation and motor imagery, specifically the dorsolateral prefrontal cortex (DLPFC), dorsal premotor cortex (DPMC), and ventrolateral premotor cortex (VLPMC) [4]. The criteria for cortical regions based on the positions of the eight fNIRS channels are detailed in Table 1. The average values of oxygenated hemoglobin (HbO<sub>2</sub>) obtained from the eight fNIRS channels underwent processing using Oxysoft 3.2.51.4 (Artinis Medical Systems, Netherlands), employing a low-pass filter with a cutoff frequency of 0.5 Hz applied to the measured signals [22].

Moreover, considering physiological delays in hemodynamic responses, the initial 5 seconds during the video observation period were excluded. To avoid anticipated responses at the end of the video, the last 5 seconds were also excluded. Additionally, the initial 5 seconds before the video start were excluded from the 10-second baseline period to prevent anticipated responses [23].

TABLE I  
CORTICAL REGION CRITERIA BASED ON CHANNEL LOCATIONS

Region	Channel	Region	Channel
Rt DLPFC	T1	Lt DLPFC	T5
Rt DPMC	T2, T3	Lt DPMC	T6, T8
Rt VLPMC	T4	Lt VLPMC	T7

Rt: right; Lt: left; DLPFC: dorsolateral prefrontal cortex; DPMC: dorsal premotor cortex; VLPMC: ventrolateral premotor cortex

### E. Statistical analysis

This study utilized SPSS version 22.0 (SPSS Inc., USA) to analyze the changes in brain activation across video conditions. Paired sample t-tests were applied to all channels to evaluate the contrast between each of the four conditions and the baseline. Furthermore, the alterations in brain activation during

AOMI were visually examined through graphs generated in Microsoft Excel (Microsoft, Redmond, WA, USA).

TABLE III  
COMPARISON OF BRAIN ACTIVATION CHANNELS BY VIDEO CONDITIONS ( $N = 28$ )

	SW vs. Baseline		SE vs. Baseline		NSW vs. Baseline		NSE vs. Baseline	
	t	p	t	p	t	p	t	p
T1	1.758	0.09	3.24	0.003	2.263*	0.032	2.345*	0.027
T2	-0.029	0.977	0.227	0.822	-0.699	0.491	-0.221	0.827
T3	-1.083	0.288	-0.623	0.539	-0.996	0.328	-0.686	0.498
T4	1.461	0.155	2.548*	0.018	2.012	0.054	1.183	0.247
T5	2.624*	0.014	2.548*	0.017	2.545*	0.017	2.459*	0.021
T6	0.676	0.505	0.085	0.933	0.561	0.579	0.987	0.333
T7	2.549*	0.017	2.578*	0.016	2.489*	0.019	3.045**	0.005
T8	1.619	0.117	1.3	0.205	1.743	0.093	1.848	0.076

\* $p < 0.05$ , \*\* $p < 0.01$ , SW: Sound & Wrist, SE: Sound & Elbow, NSW: No sound & Wrist, NSE: No sound & Elbow

### III. RESULTS

Due to measurement issues with the fNIRS device, data from two participants were excluded, leading to a total analysis involving 28 participants. To exclude the possibility of the video condition order influencing the results, the order was randomly arranged, and no significant differences were observed across conditions.

#### A. Comparison of Brain Activation Channels by Videos Conditions

In comparison to the baseline, the condition that demonstrated the highest activation across channels was the video featuring sound and unrestricted motion visibility up to the elbow (Table 2). Furthermore, activation was noted in the DLPFC and VLPMC regions.

#### B. Time-series Data Across Video Conditions

The data included peak values and the time taken to reach the peak for each video condition. The video condition with the highest peak value was the one without sound and unrestricted motion visibility up to the elbow. Additionally, the video condition that reached the peak value the fastest was the one without sound and restricted motion visibility up to the wrist (Fig. 3). This suggests that videos without sound exhibited higher peak values and reached the peak faster compared to videos with sound.

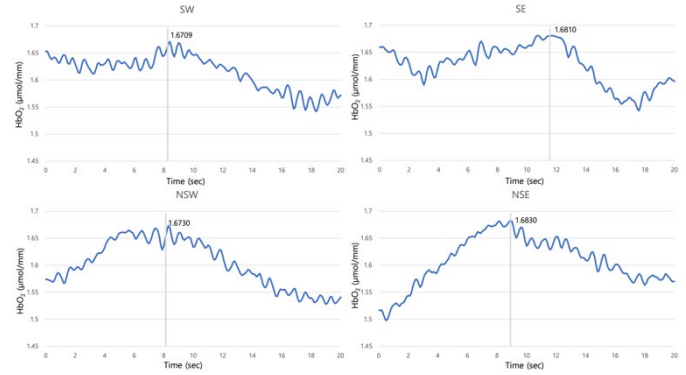


Fig. 3 Time-series data across video conditions

### IV. DISCUSSION

The primary aim of this study was to explore differences in brain activation during AOMI, specifically investigating the impact of sound and motion visibility in videos using fNIRS. The findings indicated that more areas associated with AOMI were activated in videos with sound and unrestricted motion visibility up to the elbow, in comparison to other conditions. As cognitive demands increased, there was a corresponding rise in the activation of brain regions, suggesting greater neural recruitment in the relevant areas [24]. Consequently, this investigation sought to identify conditions that exhibited significantly higher activation in regions related to AOMI compared to the baseline, with a particular focus on the DLPFC, DPMC, and VLPMC.

Concerning the presence or absence of sound, the outcomes of this study are consistent with a prior investigation suggesting that the presence of sound establishes a robust auditory-motor coupling, enhancing motor learning [13]. However, it contradicts previous findings from an EEG study, which proposed that observing videos without sound is more effective in activating the mirror neuron system compared to observing videos with sound [15]. The inconsistency in these results may be attributed to differences in measurement sites. The activation

of the mirror neuron system typically progresses from the occipital cortex to the superior temporal region, inferior parietal lobule, Broca's area, and finally, the primary motor cortex [25]. Notably, previous EEG studies focused on the primary motor cortex and primary somatosensory cortex as measurement locations, diverging from the DLPFC area measured in this study, which corresponds to Broca's area. Additionally, it is crucial to consider the temporal disparity in the measurement periods between the previous EEG study (2 to 4 seconds) and the fNIRS measurement period in this study (30 seconds).

On the other hand, the findings of this study concerning motion visibility contradict a prior discovery that showcased the effectiveness of videos illustrating movements up to the ankle, particularly when comparing brain activation between videos displaying the lower half of a person walking and videos depicting motion up to the ankle [18]. Conversely, according to the attention field theory, neuron firing rates increase more significantly when stimuli are larger compared to when they are smaller. This aligns with the outcomes of our study, where brain activation in the unrestricted motion visibility up to the elbow, representing relatively larger stimuli, was higher. Nonetheless, additional investigation is warranted to substantiate the differences in brain activation between the upper and lower extremities based on various motion visibilities.

In the time series analysis, it was observed that videos without sound exhibited higher peak values and reached the peak faster compared to videos with sound. The elevated HbO<sub>2</sub> levels reaching the peak more rapidly indicate individuals efficiently utilize brain resources when confronted with challenging cognitive tasks [26]. Consequently, this study scrutinized peak values and the time taken to reach the peak for each condition. The results revealed that videos without sound reached the peak faster and had a more prolonged descent phase compared to videos with sound. Thus, when observing short videos of about 10 seconds, it is suggested that watching videos without sound may be more suitable for eliciting brain activation.

Until now, there has been limited research on video conditions during AOMI, underscoring the significance of this study in shedding light on optimal video conditions for AOMI. Moreover, the study is noteworthy for identifying suitable video conditions specifically tailored for observing short-length videos. Executing AOMI under optimal conditions has the potential to amplify activation in relevant brain regions, potentially inducing neuroplasticity [27]. Furthermore, given that fNIRS enables real-time monitoring of cerebral hemodynamic changes [28], it could be utilized to monitor intervention effects during AOMI for motor rehabilitation. With its emphasis on the upper extremities, unlike previous investigations, this study has the potential to enrich the field of upper limb motor rehabilitation research. Additionally, while previous studies have identified effective video conditions through the primary motor cortex, our study is significant in that it identified these conditions through three regions associated with the mirror neuron system (DLPFC, DPMC, VLPMC), revealing significant differences. Moreover, while previous studies often utilized short and simple activities in their videos, making them less applicable for AOMI training in rehabilitation, our study's videos were designed to reflect the typical lengths

observed in clinical settings and consisted of activities of daily living. Therefore, our findings are more suitable for clinical application.

Despite identifying optimal video conditions to enhance brain activation through comparisons across different video conditions, this study comes with several limitations. Firstly, there was an imbalance in the gender distribution of participants, with 10 males and 18 females. Secondly, the study exclusively included healthy young adults as participants, limiting the generalizability of our findings. Future research should consider examining brain activation differences across video conditions in clinical groups such as stroke or Parkinson's patients. Thirdly, the videos used in this study focused solely on the distal side during upper limb movements. Future studies might need to compare brain activation under different conditions by observing not only the proximal part of the upper extremities but also lower limb movements. Additionally, there is a need to investigate brain activation differences between upper limb and lower limb movements. Finally, a limitation is the lack of diversity in video durations. In future studies, it will be essential to diversify the duration of videos and compare the differences in brain activation across various video durations.

#### ACKNOWLEDGMENT

This research was supported by Korea Institute for Advancement of Technology(KIAT) grant funded by the Korea Government(MOTIE) (P0012724, HRD Program for industrial Innovation).

#### REFERENCES

- [1] B. Buchignani et al., "Action observation training for rehabilitation in brain injuries: a systematic review and meta-analysis," *BMC Neurology*, vol. 19, pp. 1-16, 2019.
- [2] B. Neuman and R. Gray, "A direct comparison of the effects of imagery and action observation on hitting performance," *Movement & Sport Sciences-Science & Motricité*, no. 79, pp. 11-21, 2013.
- [3] J. Decety, "The neurophysiological basis of motor imagery," *Behavioural Brain Research*, vol. 77, no. 1-2, pp. 45-52, 1996.
- [4] R. M. Hardwick, S. Caspers, S. B. Eickhoff, and S. P. Swinnen, "Neural correlates of action: Comparing meta-analyses of imagery, observation, and execution," *Neuroscience & Biobehavioral Reviews*, vol. 94, pp. 31-44, 2018.
- [5] J. J. Zhang, K. N. Fong, N. Welage, and K. P. Liu, "The activation of the mirror neuron system during action observation and action execution with mirror visual feedback in stroke: a systematic review," *Neural Plasticity*, 2018.
- [6] S. Romano-Smith, G. Wood, D. J. Wright, and C. J. Wakefield, "Simultaneous and alternate action observation and motor imagery combinations improve aiming performance," *Psychology of Sport and Exercise*, vol. 38, pp. 100-106, 2018.
- [7] M. Scott, S. Taylor, P. Chesterton, S. Vogt, and D. L. Eaves, "Motor imagery during action observation increases eccentric hamstring force: an acute non-physical intervention," *Disability and Rehabilitation*, vol. 40, no. 12, pp. 1443-1451, 2018.
- [8] Y. Sun, W. Wei, Z. Luo, H. Gan, and X. Hu, "Improving motor imagery practice with synchronous action observation in stroke patients," *Topics in Stroke Rehabilitation*, vol. 23, no. 4, pp. 245-253, 2016.
- [9] M. W. Scott, G. Wood, P. S. Holmes, B. Marshall, J. Williams, and D. J. Wright, "Combined action observation and motor imagery improves learning of activities of daily living in children with Developmental Coordination Disorder," *Plos one*, vol. 18, no. 5, e0284086, 2023.
- [10] S. Ge, H. Liu, P. Lin, J. Gao, C. Xiao, and Z. Li, "Neural basis of action observation and understanding from first-and third-person perspectives: an fMRI study," *Frontiers in Behavioral Neuroscience*, vol. 12, p. 283, 2018.
- [11] M. Angelini, M. Fabbri-Destro, N. F. Lopomo, M. Gobbo, G. Rizzolatti, and P. Avanzini, "Perspective-dependent reactivity of sensorimotor mu

- rhythm in alpha and beta ranges during action observation: an EEG study," *Scientific Reports*, vol. 8, no. 1, p. 12429, 2018.
- [12] H. J. Park et al., "Action observation training of community ambulation for improving walking ability of patients with post-stroke hemiparesis: a randomized controlled pilot trial," *Clinical Rehabilitation*, vol. 31, no. 8, pp. 1078-1086, 2017.
- [13] N. Schaffert, T. B. Janzen, K. Mattes, and M. H. Thaut, "A review on the relationship between sound and movement in sports and rehabilitation," *Frontiers in Psychology*, vol. 10, p. 244, 2019.
- [14] N. Schaffert, T. B. Janzen, K. Mattes, and M. H. Thaut, "A review on the relationship between sound and movement in sports and rehabilitation," *Frontiers in Psychology*, vol. 10, p. 244, 2019.
- [15] B. Wang, J. Zhang, C. Wang, and J. Hong, "Effectiveness of action observed for sports function rehabilitation based on mirror neuron," in *2016 13th International Conference on Ubiquitous Robots and Ambient Intelligence (URAI)*, pp. 338-341, IEEE, August 2016.
- [16] V. Demarin and S. Morović, "Neuroplasticity," *Periodicum Biologorum*, vol. 116, no. 2, pp. 209-211, 2014.
- [17] G. D'Innoccenzo, C. C. Gonzalez, A. V. Nowicky, A. M. Williams, and D. T. Bishop, "Motor resonance during action observation is gaze-contingent: A TMS study," *Neuropsychologia*, vol. 103, pp. 77-86, 2017.
- [18] T. Ito et al., "Visual Attention and Motion Visibility Modulate Motor Resonance during Observation of Human Walking in Different Manners," *Brain Sciences*, vol. 11, no. 6, p. 679, 2021.
- [19] S. Héту, M. Gagne, P. L. Jackson, and C. Mercier, "Variability in the effector-specific pattern of motor facilitation during the observation of everyday actions: implications for the clinical use of action observation," *Neuroscience*, vol. 170, no. 2, pp. 589-598, 2010.
- [20] T. Wilcox and M. Biondi, "fNIRS in the developmental sciences," *Wiley Interdisciplinary Reviews: Cognitive Science*, vol. 6, no. 3, pp. 263-283, 2015.
- [21] Korean Disease Control and Prevention Agency, "Press release," August 24, 2016. [Online]. Available: [https://www.kdca.go.kr/gallery.es?mid=a20503020000&bid=0003&act=view&list\\_no=136782](https://www.kdca.go.kr/gallery.es?mid=a20503020000&bid=0003&act=view&list_no=136782).
- [22] R. A. Khan et al., "Cortical tasks-based optimal filter selection: an fNIRS study," *Journal of Healthcare Engineering*, vol. 2020, pp. 1-15, 2020.
- [23] P. Nóbrega-Sousa et al., "Prefrontal cortex activity during walking: effects of aging and associations with gait and executive function," *Neurorehabilitation and Neural Repair*, vol. 34, no. 10, pp. 915-924, 2020.
- [24] F. G. Metzger et al., "Functional brain imaging of walking while talking—an fNIRS study," *Neuroscience*, vol. 343, pp. 85-93, 2017.
- [25] N. Nishitani and R. Hari, "Viewing lip forms: cortical dynamics," *Neuron*, vol. 36, no. 6, pp. 1211-1220, 2002.
- [26] D. Yang, K. S. Hong, S. H. Yoo, and C. S. Kim, "Evaluation of neural degeneration biomarkers in the prefrontal cortex for early identification of patients with mild cognitive impairment: an fNIRS study," *Frontiers in Human Neuroscience*, vol. 13, p. 317, 2019.
- [27] S. C. Cramer et al., "Harnessing neuroplasticity for clinical applications," *Brain*, vol. 134, no. 6, pp. 1591-1609, 2011.
- [28] P. Pinti et al., "Using fiberless, wearable fNIRS to monitor brain activity in real-world cognitive tasks," *JoVE (Journal of Visualized Experiments)*, no. 106, e53336, 2015