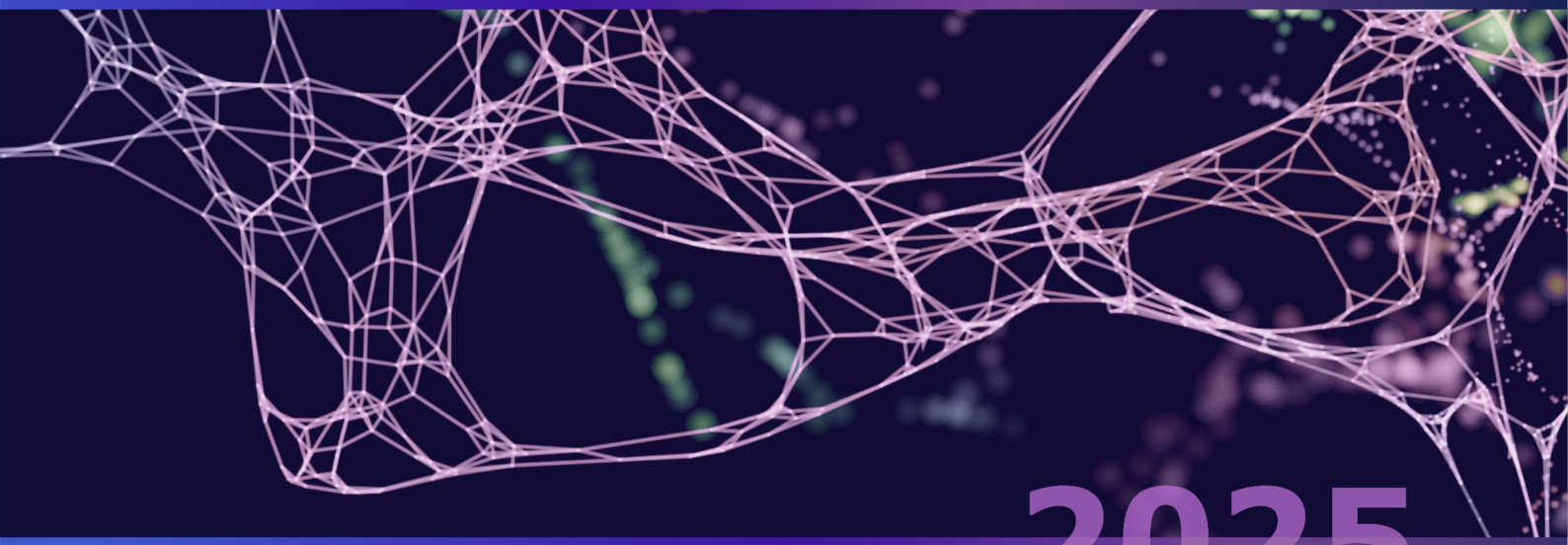
AiMH Innsbruck 2025

Artificial Intelligence in Medicine and Healthcare

Proceedings

1st International Conference on AI in Medicine and Healthcare (AiMH' 2025)



2025

8-10 April 2025, Innsbruck, Austria

IFSA

Artificial Intelligence in Medicine and Healthcare

Proceedings of the 1st International Conference on AI in Medicine and Healthcare (AiMH' 2025)

8-10 April 2025
Innsbruck, Austria

Edited by
Dr., Prof. Sergey Y. Yurish
IFSA,
Barcelona, Spain

IFSA Publishing

Sergey Y. Yurish, *Editor*
AiMH' 2025 Conference Proceedings

E-mail (for orders and customer service enquires): ifsa.books@sensorsportal.com

Visit our Home Page on http://www.sensorsportal.com

ISBN: 978-84-09-71190-1
BN-20250402-XX
BIC: UYQ

Contents

Foreword

It is both an honor and a privilege to present the proceedings of the *1st International Conference on Artificial Intelligence in Medicine and Healthcare (AiMH' 2025)*, held in Innsbruck, Austria, from April 8 to 10, 2025. This conference represents a pivotal convergence of computational innovation and medical science, designed to explore the multifaceted and rapidly evolving interface between artificial intelligence (AI) methodologies and their translational applications in medicine, clinical diagnosis, and healthcare systems.

The contributions compiled herein reflect a broad and rigorous spectrum of inquiry, ranging from deep learning-based diagnostic algorithms, generative models for data augmentation, and multimodal fusion techniques, to interpretable machine learning, federated and privacy-preserving AI frameworks, and AI-augmented clinical workflow systems. Such diversity highlights not only the interdisciplinary nature of this research domain but also its increasing clinical relevance and potential for real-world deployment in diverse healthcare settings.

Particular attention in this year's submissions has been given to the methodological robustness, clinical interpretability, and generalizability of AI models—cornerstones of trustworthy AI in healthcare. Several presented works advance the state of the art in performance benchmarking of AI against human experts in radiology, oncology, pathology, and surgery. Others propose novel algorithmic frameworks for early disease prediction, patient stratification, digital twin generation, and intelligent assistance in cognitive and behavioral therapy.

The significance of these developments cannot be overstated. They exemplify how data-driven insights and algorithmic intelligence are beginning to redefine conventional paradigms in medical science—improving diagnostic accuracy, enabling precision medicine, and enhancing operational efficiency in clinical environments. These proceedings are a testament to the collective effort of an international research community committed to the responsible and impactful integration of AI into healthcare systems. Each paper has been rigorously peer-reviewed and selected for its originality, scientific merit, and potential impact.

As Chairman of AiMH' 2025, I wish to extend my deepest appreciation to the authors for their excellent contributions, the program committee for their meticulous and scholarly peer review, and all session chairs and organizing staff for their tireless dedication to ensuring the success of this conference. I am confident that the research contained in this volume will catalyze further advancements in the field and foster collaborative partnerships that will shape the next generation of intelligent healthcare systems.

Prof. Dr. Sergey Y. Yurish
AiMH' 2025 Chairman

(007)

Assessing Artificial Intelligence and Radiologist Performance in Musculoskeletal Fracture Detection: Multi-Reader, Multi-Case Study

J. Dandár [1], S. Klíčník [2], Z. Straka [1] and D. Kvak [1,4,*]
[1] Carebot s.r.o., Rašínovo Nábřeží 71/10, 128 00 Prague, Czechia
[2] Department of Radiology and Nuclear Medicine, University Hospital Královské Vinohrady, Šrobárova 1150/50, 100 00 Prague, Czechia
[3] Department of Simulation Medicine, Faculty of Medicine, Masaryk University, Kamenice 5, 625 00 Brno, Czechia
Tel.: + 420 739 174 316
E-mail: daniel.kvak@carebot.com

Summary: Fracture detection on musculoskeletal (MSK) radiographs is critical for both emergency and routine care, yet diagnostic errors remain common due to high workloads and limited radiological expertise. This study evaluates the diagnostic performance of an artificial intelligence (AI) system in detecting fractures on MSK X-rays, comparing its performance to six radiologists of varying experience levels in a blinded multi-reader, multi-case (MRMC) study. A total of 489 radiographs were retrospectively analyzed from routine clinical practice, with ground truth established for 448 images through consensus among three experienced radiologists. Diagnostic performance was assessed using sensitivity (Se), specificity (Sp), positive (PLR), and negative likelihood ratio (NLR), with statistical analysis including McNemar's test and Holm's method for multiple comparisons. The AI system achieved a Se of 0.921 (95% CI: 0.846–0.961) and Sp of 0.897 (0.861–0.924). Radiologists' Se ranged from 0.663 to 0.933, and Sp ranged from 0.916 to 0.989. The AI demonstrated consistent high sensitivity across body parts, particularly for elbow and hand/wrist fractures, often exceeding radiologists' performance. Specificity was slightly lower but acceptable, supporting AI's potential as a complementary diagnostic tool. These findings highlight the clinical utility of AI in MSK fracture detection, particularly in settings with limited resources or high diagnostic workloads. Future research should validate these results in larger, multicentric studies to ensure broader generalizability and evaluate AI integration in real-world workflows.

Keywords: Artificial intelligence, Fracture detection, Musculoskeletal radiographs, Diagnostic performance, Multi-reader study.

1. Introduction

Fractures represent a significant burden in emergency medical care, with skeletal injuries being among the most common reasons for consultations. Radiographs remain the first-line imaging modality for the detection of fractures and are one of the most utilized diagnostic tools worldwide [1,2]. Meanwhile, missed fractures account for up to 80% of diagnostic errors in emergency settings [3], posing a significant challenge to healthcare systems. In the United States, extremity fractures are the second most commonly missed diagnosis leading to medical-legal claims, following breast cancer [4]. These errors stem from several factors, including high workload, insufficient radiologic expertise, and the cognitive biases inherent in high-pressure environments. Despite their ubiquity and essential role in trauma assessment, the interpretation of radiographs is a demanding task, requiring specialized radiologic expertise often unavailable in overburdened emergency departments [5]. Consequently, there is an urgent need for innovative solutions to enhance diagnostic accuracy and efficiency in fracture detection.

Over the past two decades, computer-aided detection (CAD) systems have been developed to support radiologists, initially focusing on applications such as breast cancer screening and lung nodule detection. These advancements have reinvigorated interest in AI-based solutions for fracture detection [6], addressing limitations of traditional CAD systems [7]. AI-based fracture detection systems have demonstrated promising results, achieving high sensitivity and specificity in identifying fractures on radiographs [8,9]. Studies have shown that incorporating AI aids can significantly improve the diagnostic performance of both radiologists and emergency physicians, enhancing sensitivity and specificity while reducing false-positive rates [10]. Despite these advancements, significant challenges remain in translating AI solutions into clinical practice. Many prior studies have focused on isolated body parts or used internal test sets, limiting their generalizability.

This study aims to evaluate the diagnostic performance of a deep learning-based AI system for fracture detection in a multi-reader, comparative setting. By assessing its impact on the real-world prevalenced, the research seeks to provide robust evidence for the clinical utility of AI in trauma care.

2. Material and Methods

2.1. Algorithm

The proposed AI system (Carebot AI Bones 1.8.10; Carebot s.r.o.) is an advanced deep learning-based

software developed to assist in the detection of fractures on musculoskeletal radiographs. Leveraging the object detection YOLO architecture [11], the software implements a single-stage regression model optimized for real-time object detection. It is complemented by a body part classifier to ensure analysis tailored to specific anatomical regions. The AI system was trained on a robust dataset comprising 115,213 annotated radiographic images, curated by a diverse group of 21 radiologists with varying levels of experience. The annotations included a wide range of fracture types and anatomical complexities, ensuring the development of a comprehensive detection algorithm.

To address the diversity in musculoskeletal imaging, seven body part-specific models were independently trained and validated, with model selection driven by performance metrics emphasizing both sensitivity (Se) and specificity (Sp). The software was designed for seamless integration into clinical workflows, including compatibility with picture archiving and communication systems (PACS) and hospital information systems (HIS). Upon detecting a potential fracture with a confidence level exceeding the predefined threshold for a specific body part, the software categorizes the image as "Abnormal" or "Normal" and highlights the suspected fracture site with a bounding box (Fig. 1).

Fig. 1. An example of the examined AI software (Carebot AI Bones 1.8.10; Carebot s.r.o.) predictions.

2.2. Data Collection and Demographics

A total of 585 musculoskeletal (MSK) radiographs from both pediatric and adult patients were retrospectively collected, representing a routine clinical practice at Šumperk Hospital. The data were collected from a one-week period, from September 11 to September 17, 2023, reflecting the real-world, local prevalence of MSK conditions and injuries. Radiographs were sourced from diverse clinical settings, including the emergency department and routine outpatient examinations. Exclusion criteria, specifically excluding radiographs of the spine, head, and nasal bones, were applied to ensure the dataset's relevance to the study's objectives. Applying the criteria resulted in a dataset of 489 MSK radiographs relevant for analysis.

2.3. Ground Truth

Three experienced radiologists (E.H., A.H., and Z.T., with 6, 7, and 9 years of experience, respectively) independently evaluated the collected MSK radiographs using dedicated annotation software (Carebot Label App). The radiologists had access to standard image viewing tools, such as panning, zooming, and window level adjustments, to facilitate detailed assessments. No AI assistance was provided during the annotation process. For identifying fracture-positive cases, a majority vote (e.g., 2/3 or 3/3) was acquired to determine the ground truth. Conversely, for cases identified as fracture-negative, full agreement (3/3) on the evaluation was required. Radiographs with disagreements were excluded from the multi-reader study.

Out of 489 MSK radiographs, the ground truth was reached for 448 (91.6%) images (n_{GT}, Table 1), with 359 classified as fracture-negative (n_{NORMAL}) and 89 as fracture-positive ($n_{FRACTURE}$), resulting in the exclusion of 41 radiographs (8.4%) with no ground truth. The dataset reflected a diverse range of imaging modalities, with the majority of images obtained using the SIEMENS Fluorospot Compact FD system (n=427), followed by Philips Medical Systems PCR (n=16), Philips Medical Systems DigitalDiagnost (n=3), and DRGEM GXR-32SD (n=2).

Table 1. Distribution of fractures and normal cases across different body parts.

Body part	$n_{FRACTURE}$	n_{NORMAL}	n_{GT}
Foot/Ankle	26	87	113
Knee/Leg	7	113	120
Hand/Wrist	34	75	109
Elbow/Arm	9	14	23
Shoulder/Clavicle	7	29	36
Pelvic/Hip Joint	6	41	47
Total	**89**	**359**	**448**

2.4. Assessment

The study employed a multi-reader, multi-case (MRMC) design, involving six radiologists (RAD 1–RAD 6) with varying levels of experience and serving hospitals of different sizes (Table 2). The radiologists' experience in evaluating musculoskeletal radiographs ranged from 2 to 4 years. The evaluation was conducted in a blinded setting, with no access to clinical information and no imposed time limit. The reading sessions were conducted using a DICOM-compatible medical viewer. The diagnostic performance of the AI software was evaluated in comparison to the radiologists' independent assessments and the established ground truth.

Table 2. List of radiologists (RAD 1–RAD 6) participating in the multi-reader study alongside their respective experience levels and the size of the hospitals they are serving.

ID	Years of Experience	Hospital Size
RAD 1	2 years	Large
RAD 2	2 years	Large
RAD 3	3 years	Small
RAD 4	3 years	Large
RAD 5	4 years	Medium
RAD 6	4 years	Medium

2.5. Statistical Analysis

The evaluation of fracture detection in this study focused on key metrics, including sensitivity (Se), specificity (Sp), positive likelihood ratio (PLR), and negative likelihood ratio (NLR). A paired design was employed, wherein each radiograph was analyzed by both the AI software and all participating radiologists (RAD 1–RAD 6), with results compared to the established ground truth. Confidence intervals (CI) and p-values were calculated to assess the Se and Sp of the AI software relative to individual radiologists. Null hypothesis testing (H_0) was conducted using the Wald χ^2 test, testing H_0: $Se_1 = Se_2$ and $Sp_1 = Sp_2$ versus the alternative hypothesis H_1: $Se_1 \neq Se_2$ and/or $Sp_1 \neq Sp_2$. Upon rejection of H_0, subsequent pairwise comparisons were performed using McNemar's test with continuity correction for Se and Sp, while the Holm method was applied for adjusting multiple comparisons of PLR and NLR.

Results were visualized using forest plots, with 95% confidence intervals (CIs) calculated using the Wilson score interval for binomial proportions. Data processing and analysis were carried out using Python (version 3.9) and its libraries (SciPy, Scikit-learn, and Pandas). A two-sided significance level of $p < 0.05$ was applied to all secondary analyses.

3. Results

Out of a total of 89 images with confirmed fractures, the AI model correctly identified 82 of these as true positives (TP), resulting in a sensitivity (Se) of 0.921 (95% CI: 0.846–0.961, Table 3, Fig. 2). Additionally, the AI model incorrectly predicted 37 images as false positives (FP), leading to a specificity (Sp) of 0.897 (95% CI: 0.861–0.924). The AI software demonstrated significantly higher Se than RAD 1 (0.663, 95% CI: 0.560–0.753, $p < 0.05$), RAD 2 (0.798, 95% CI: 0.703–0.868, $p < 0.05$), and RAD 3 (0.719, 95% CI: 0.618–0.802, $p < 0.05$). The exception was with RAD 4 (0.910, 95% CI: 0.833–0.954, $p = 1.00$), RAD 5 (0.921, 95% CI: 0.846–0.961, $p = 1.00$), and RAD 6 (0.933, 95% CI: 0.861–0.969, $p = 1.00$), where the differences were not statistically significant. However, the Sp of the AI model (0.897, 95% CI: 0.861–0.924) was significantly lower than that of most radiologists, including RAD 1 (0.975, 95% CI: 0.953–0.987, $p < 0.05$), RAD 2 (0.983, 95% CI: 0.964–0.992, $p < 0.05$), RAD 3 (0.989, 95% CI: 0.972–0.996, $p < 0.05$), and RAD 4 (0.955, 95% CI: 0.929–0.972, $p < 0.05$). The differences in Sp between the AI model and RAD 5 (0.916, 95% CI: 0.883–0.941, $p = 0.44$) and RAD 6 (0.922, 95% CI: 0.890–0.945, $p = 0.30$) were not statistically significant.

Table 3. Performance of AI software (Carebot AI Bones 1.8.10; Carebot s.r.o.) and radiologists (RAD 1–RAD 6) in the multi-reader study, including sensitivity (Se) and specificity (Sp) with 95% confidence intervals (CI) and p-values.

ID	Se (95% CI)	p (Se)	Sp (95% CI)	p (Sp)
AI	**0.921 (0.846–0.961)**	**-**	**0.897 (0.861–0.924)**	**-**
RAD 1	0.663 (0.560–0.753)	< 0.05	0.975 (0.953–0.987)	< 0.05
RAD 2	0.798 (0.703–0.868)	< 0.05	0.983 (0.964–0.992)	< 0.05
RAD 3	0.719 (0.618–0.802)	< 0.05	0.989 (0.972–0.996)	< 0.05
RAD 4	0.910 (0.833–0.954)	1.00	0.955 (0.929–0.972)	< 0.05
RAD 5	0.921 (0.846–0.961)	1.00	0.916 (0.883–0.941)	0.44
RAD 6	0.933 (0.861–0.969)	1.00	0.922 (0.890–0.945)	0.30

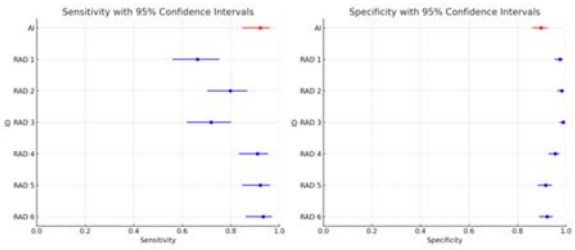


Fig. 2. Forest plots illustrating the performance comparison between the proposed AI software (Carebot AI Bones 1.8.10; Carebot s.r.o.) and assessed radiologists (RAD 1–RAD 6) in the multi-reader study in terms of sensitivity (Se) and specificity (Sp), including corresponding confidence intervals (CI).

The AI demonstrated superior sensitivity (Se) compared to radiologists across various body parts

(Table 4). For foot/ankle fractures, the AI software achieved a Se of 0.923 (95% CI: 0.759–0.979) and a specificity (Sp) of 0.874 (95% CI: 0.788–0.928), surpassing radiologists such as RAD 1 (Se 0.769, 95% CI: 0.579–0.890) and RAD 3 (Se 0.769, 95% CI: 0.579–0.890), although performing comparably to RAD 4 (Se 0.923, 95% CI: 0.759–0.979). In the knee/leg category, the AI model showed a Se of 0.714 (95% CI: 0.359–0.918) and an Sp of 0.894 (95% CI: 0.824–0.938), performing comparably to RAD 2 (Se 0.714, 95% CI: 0.359–0.918; Sp 1.000, 95% CI: 0.967–1.000). The AI model outperformed RAD 1 in sensitivity (Se 0.571, 95% CI: 0.250–0.842), but radiologists generally showed higher specificity in this category. For hand/wrist fractures, the AI model achieved a Se of 0.941 (95% CI: 0.809–0.984) and an Sp of 0.907 (95% CI: 0.820–0.954), outperforming most radiologists in sensitivity, such as RAD 1 (Se 0.647, 95% CI: 0.479–0.785) and RAD 2 (Se 0.794, 95% CI: 0.632–0.897). RAD 6 demonstrated the highest sensitivity (Se 0.971, 95% CI: 0.851–0.995), but with a slightly lower specificity (Sp 0.880, 95% CI: 0.787–0.936) compared to the AI. For elbow/arm fractures, the AI model achieved perfect sensitivity (Se 1.000, 95% CI: 0.701–1.000) but had a slightly lower specificity (Sp 0.929, 95% CI: 0.685–0.987) compared to RAD 4 (Se 1.000, 95% CI: 0.701–1.000; Sp 1.000, 95% CI: 0.785–1.000). In the shoulder/clavicle category, the AI model also achieved perfect sensitivity (Se 1.000, 95% CI: 0.646–1.000) and an Sp of 0.966 (95% CI: 0.828–0.994). The AI's performance in this category was comparable to that of RAD 4 (Se 1.000, 95% CI: 0.646–1.000; Sp 0.966, 95% CI: 0.828–0.994) and exceeded that of RAD 1 (Se 0.429, 95% CI: 0.158–0.750). For pelvis/hip joint fractures, the AI model achieved a high Se of 0.833 (95% CI: 0.436–0.970) and a moderate Sp of 0.878 (95% CI: 0.745–0.947). The AI outperformed RAD 1 in sensitivity (Se 0.500, 95% CI: 0.188–0.812), though radiologists generally had higher specificity in this category, with RAD 1 achieving an Sp of 1.000 (95% CI: 0.914–1.000).

Fig. 3. Radiographic images of the hand in two projections. An oblique projection was correctly classified as a True Positive (TP) with a detected fracture (FRA) in the proximal phalanx, while a posteroanterior (PA) projection was misclassified as a False Negative (FN).

Fig. 4. Radiographic images of multiple trauma patients, each presenting with a distinct fracture (FRA) across different anatomical regions. All images were correctly classified as True Positive (TP), as highlighted by the bounding box.

Fig. 5. Radiographic series of a trauma patient presenting with a shoulder fracture. The study includes four musculoskeletal (MSK) images: three projections were correctly identified as True Negative (TN) for the absence of a fracture, while the shoulder projection was correctly identified as True Positive (TP) with a fracture (FRA), as highlighted by the bounding box.

Table 4. Performance of AI software (Carebot AI Bones 1.8.10; Carebot s.r.o.) and radiologists (RAD 1–RAD 6) in the multi-reader study, including sensitivity (Se) and specificity (Sp) with 95% confidence intervals (CI) and p-values.

Body part	ID	Se (95% CI)	Sp (95% CI)
Foot/Ankle	**AI**	**0.923 (0.759–0.979)**	**0.874 (0.788–0.928)**
	RAD 1	0.769 (0.579–0.890)	0.977 (0.920–0.994)
	RAD 2	0.846 (0.665–0.938)	0.977 (0.920–0.994)
	RAD 3	0.769 (0.579–0.890)	0.977 (0.920–0.994)
	RAD 4	0.923 (0.759–0.979)	0.954 (0.888–0.982)
	RAD 5	0.923 (0.759–0.979)	0.897 (0.815–0.945)
	RAD 6	0.885 (0.710–0.960)	0.920 (0.843–0.960)
Knee/Leg	**AI**	**0.714 (0.359–0.918)**	**0.894 (0.824–0.938)**
	RAD 1	0.571 (0.250–0.842)	0.991 (0.952–0.998)
	RAD 2	0.714 (0.359–0.918)	1.000 (0.967–1.000)
	RAD 3	0.429 (0.158–0.750)	1.000 (0.967–1.000)
	RAD 4	0.857 (0.487–0.974)	0.991 (0.952–0.998)
	RAD 5	0.857 (0.487–0.974)	0.982 (0.938–0.995)
	RAD 6	0.857 (0.487–0.974)	0.982 (0.938–0.995)
Hand/Wrist	**AI**	**0.941 (0.809–0.984)**	**0.907 (0.820–0.954)**
	RAD 1	0.647 (0.479–0.785)	0.947 (0.871–0.979)
	RAD 2	0.794 (0.632–0.897)	0.960 (0.889–0.986)
	RAD 3	0.824 (0.665–0.917)	0.987 (0.928–0.998)
	RAD 4	0.882 (0.734–0.953)	0.880 (0.787–0.936)
	RAD 5	0.941 (0.809–0.984)	0.867 (0.772–0.926)
	RAD 6	0.971 (0.851–0.995)	0.880 (0.787–0.936)
Elbow/Arm	**AI**	**1.000 (0.701–1.000)**	**0.929 (0.685–0.987)**
	RAD 1	0.778 (0.453–0.937)	0.929 (0.685–0.987)
	RAD 2	0.889 (0.565–0.980)	1.000 (0.785–1.000)
	RAD 3	0.556 (0.267–0.811)	1.000 (0.785–1.000)
	RAD 4	1.000 (0.701–1.000)	1.000 (0.785–1.000)
	RAD 5	1.000 (0.701–1.000)	0.857 (0.601–0.960)
	RAD 6	1.000 (0.701–1.000)	0.857 (0.601–0.960)
Shoulder/Clavicle	**AI**	**1.000 (0.646–1.000)**	**0.966 (0.828–0.994)**
	RAD 1	0.429 (0.158–0.750)	0.966 (0.828–0.994)
	RAD 2	0.857 (0.487–0.974)	1.000 (0.883–1.000)
	RAD 3	0.857 (0.487–0.974)	1.000 (0.883–1.000)
	RAD 4	1.000 (0.646–1.000)	0.966 (0.828–0.994)
	RAD 5	1.000 (0.646–1.000)	0.828 (0.655–0.924)
	RAD 6	1.000 (0.646–1.000)	0.793 (0.616–0.902)
Pelvis/Hip joint	**AI**	**0.833 (0.436–0.970)**	**0.878 (0.745–0.947)**
	RAD 1	0.500 (0.188–0.812)	1.000 (0.914–1.000)
	RAD 2	0.500 (0.188–0.812)	0.976 (0.874–0.996)
	RAD 3	0.333 (0.097–0.700)	0.976 (0.874–0.996)
	RAD 4	0.833 (0.436–0.970)	0.976 (0.874–0.996)
	RAD 5	0.667 (0.300–0.903)	0.951 (0.839–0.987)
	RAD 6	0.833 (0.436–0.970)	0.951 (0.839–0.987)
	RAD 6	1.000 (0.646–1.000)	0.793 (0.616–0.902)

The AI software achieved a positive likelihood ratio (PLR) of 8.940 (95% CI: 6.550–12.202, Table 5, Fig. 6) and a negative likelihood ratio (NLR) of 0.088 (95% CI: 0.043–0.179). Compared to all assessed radiologists, the AI's PLR was significantly lower, indicating a limitation in its ability to rule in fractures as effectively as the radiologists. For instance, RAD 1 achieved a PLR of 26.443 (95% CI: 13.641–51.259, $p < 0.05$), RAD 2 reached 47.732 (95% CI: 21.441–106.264, $p < 0.05$), RAD 3 achieved 64.539 (95% CI: 24.146–172.503, $p < 0.05$), and RAD 4 achieved 20.421 (95% CI: 12.593–33.113, $p < 0.05$). Radiologists RAD 5 and RAD 6 demonstrated PLR values closer to the AI, with 11.025 (95% CI: 7.786–15.613, $p = 0.38$) and 11.957 (95% CI: 8.342–17.139, $p = 0.30$), respectively, but these differences were not statistically significant. However, the AI demonstrated a markedly lower NLR, highlighting its strong capability to rule out fractures when a negative result is obtained. The AI model's NLR of 0.088 (95% CI: 0.043–0.179) was significantly better than those of RAD 1 (0.346, 95% CI: 0.258–0.463, $p < 0.05$), RAD 2 (0.206, 95% CI: 0.136–0.311, $p < 0.05$), and RAD 3 (0.284, 95% CI: 0.204–0.396, $p < 0.05$). Notably, RAD 4 achieved an NLR of 0.094 (95% CI: 0.049–0.182, $p < 0.05$), which was comparable to the AI software. Radiologists RAD 5 and RAD 6 had slightly better NLR values of 0.086 (95% CI: 0.042–0.175, $p = 0.38$) and 0.073 (95% CI: 0.034–0.158, $p = 0.30$), respectively, but these differences were not statistically significant.

Table 5. Performance of AI software (Carebot AI Bones 1.8.10; Carebot s.r.o.) and radiologists (RAD 1–RAD 6) in the multi-reader study, including Positive Likelihood Ratio (PLR) and Negative Likelihood Ratio (NLR) with 95% confidence intervals (CI) and p-values.

ID	PLR (95% CI)	p (PLR)	PLR (95% CI)	p (PLR)
AI	**8.940 (6.550–12.202)**	**-**	**0.088 (0.043–0.179)**	**-**
RAD 1	26.443 (13.641–51.259)	< 0.05	0.346 (0.258–0.463)	< 0.05
RAD 2	47.732 (21.441–106.264)	< 0.05	0.206 (0.136–0.311)	< 0.05
RAD 3	64.539 (24.146–172.503)	< 0.05	0.284 (0.204–0.396)	< 0.05
RAD 4	20.421 (12.593–33.113)	< 0.05	0.094 (0.049–0.182)	< 0.05
RAD 5	11.025 (7.786–15.613)	0.38	0.086 (0.042–0.175)	0.38
RAD 6	11.957 (8.342–17.139)	0.30	0.073 (0.034–0.158)	0.30

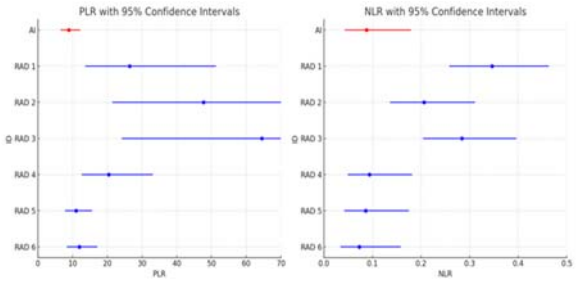


Fig. 6. Forest plots illustrating the performance comparison between the proposed AI software (Carebot AI Bones 1.8.10; Carebot s.r.o.) and assessed radiologists (RAD 1–RAD 6) in the multi-reader study in terms of Positive Likelihood Ratio (PLR) and Negative Likelihood Ratio (NLR), including corresponding confidence intervals (CI).

4. Discussion

The findings of this study highlight the diagnostic capabilities of the AI system evaluated for detecting musculoskeletal fractures, as assessed through a rigorous multi-reader, multi-case (MRMC) evaluation. Compared to other commercially-available AI systems described in the literature, this system demonstrated strong sensitivity (Se) of 0.921 (95% CI: 0.846–0.961) across various anatomical regions, although its specificity (Sp) of 0.897 (95% CI: 0.861–0.924) was relatively lower, consistent with trends observed in other AI-assisted diagnostic tools.

Several prior studies have evaluated the performance of similar AI systems, providing a basis for comparison. Duron et al. (2021) [12] reported that an AI system increased sensitivity (Se) by 8.7% and specificity (Sp) by 4.1%, while reducing the reading time by 15%. Similarly, Canoni-Meynet et al. (2022) [13] observed an increase in Se from 0.66 to 0.86 with a minimal improvement in Sp, alongside a reduction in reading times by up to 16 seconds per study. Guermazi et al. (2021) [14] reported that a leveraged AI solution improved Se from 0.648 to 0.752 and Sp from 0.906 to 0.956. The findings of this study also align with Regnard et al. (2022) [15], who noted that their AI solution reached a Se of 0.981 for fracture detection but had a Sp of 0.88, lower than radiologists' Sp of 1.000. Similarly, Dupuis et al. (2022) [16] tested a system that achieved a Se of 0.957 and Sp of 0.912, demonstrating comparable performance. Nguyen et al. (2022) [17] highlighted the role of AI in augmenting the performance of junior radiologists, reporting an increase in Se from 0.733 to 0.828, with the improvement most pronounced among less experienced clinicians. This observation is consistent with findings from this study, where the evaluated AI system outperformed less experienced radiologists in sensitivity across most anatomical regions. Cohen et al. (2022) [18] further demonstrated that a combined AI and human reading approach increased Se from 0.76 to 0.88, with standalone AI performance of 0.83, emphasizing the complementary role of AI in enhancing diagnostic accuracy.

From a clinical integration perspective, this study reinforces the potential of AI systems to reduce missed fractures, a significant source of diagnostic errors in emergency and routine care. However, minimizing false positives is crucial to prevent unnecessary follow-ups and interventions [19]. The lower specificity of the evaluated system, compared to radiologists, underscores the need for careful implementation strategies to mitigate these challenges.

4.1. Limitations

While this study provides valuable insights into the diagnostic performance of AI in musculoskeletal fracture detection, certain limitations should be acknowledged. The research was conducted in a single-center setting, which may limit its generalizability to other clinical environments with different patient populations and imaging practices. Although the dataset reflects real-world prevalence, the retrospective study design does not fully capture the impact of AI integration in a real-time clinical workflow. Future studies should focus on prospective, multi-center validation to assess the AI system's performance across diverse healthcare settings and evaluate its influence on diagnostic decision-making and efficiency. Another consideration is the AI system's higher sensitivity but slightly lower specificity compared to radiologists, leading to a higher rate of false positives. While this ensures fewer missed fractures, it could also increase the number of cases requiring secondary review. A crossover study would be necessary to determine how AI influences diagnostic accuracy, workflow efficiency, and reader confidence in clinical practice. As the field of AI in medical imaging is rapidly evolving, state-of-the-art models are emerging that may offer improvements in efficiency or diagnostic performance. However, as the AI system examined in this study is a medical device undergoing the MDR 2017/745 certification process, its design and development are finalized, with modifications only possible through recertification. The authors will continue to monitor advancements in fracture detection performance and assess their potential impact on clinical workflows in the future.

5. Conclusions

This study demonstrates the potential of the AI system as a complementary tool for musculoskeletal fracture detection, achieving high sensitivity across various anatomical regions, often exceeding radiologist performance. While the AI showed slightly lower specificity, its strong sensitivity highlights its value in minimizing missed fractures, particularly in resource-limited or high-workload environments. The results support the integration of AI as a second reader in clinical workflows, enhancing diagnostic accuracy and reducing human error. Future validation in larger, multicentric studies is necessary to confirm these findings and optimize AI implementation in real-world settings.

References

[1]. V. Arasu, H. Abujudeh, P. Biddinger, V. Noble, E. Halpern, J. Thrall, and R. Novelline, Diagnostic emergency imaging utilization at an academic trauma center from 1996 to 2012, *Journal of the American College of Radiology*, Vol. 12, 2015, pp. 467–474.

[2]. J. Willett, Imaging in trauma in limited-resource settings: a literature review, *African Journal of Emergency Medicine,* Vol. 9, 2019, pp. S21–S27.

[3]. H. Guly, Diagnostic errors in an accident and emergency department, *Emergency Medicine Journal*, Vol. 18, 2001, pp. 263–269.

[4]. A. Thabet, A. Adams, S. Jeon, J. Pisquiy, R. Gelhert, T. DeCoster, and A. Abdelgawad, Malpractice lawsuits in orthopedic trauma surgery: a meta-analysis of the literature, *OTA International*, Vol. 5, 2022, article e199.

[5]. D. Rosenthal and O. Pianykh, Efficient Radiology: How to Optimize Radiology Operations, *Springer Nature,* 2020.

[6]. J. Husarek, S. Hess, S. Razaeian, T. Ruder, S. Sehmisch, M. Müller, and E. Liodakis, Artificial intelligence in commercial fracture detection products: a systematic review and meta-analysis of diagnostic test accuracy, *Scientific Reports,* Vol. 14, 2024, article 23053.

[7]. N. Petrick, B. Sahiner, S. Armato III, et al., Evaluation of computer-aided detection and diagnosis systems, *Medical Physics*, Vol. 40, 2013, article 087001.

[8]. P. Kalmet, S. Sanduleanu, S. Primakov, et al., Deep learning in fracture detection: a narrative review, *Acta Orthopaedica,* Vol. 91, 2020, pp. 215–220.

[9]. A. Bhatnagar, A. Kekatpure, V. Velagala, and A. Kekatpure, A review on the use of artificial intelligence in fracture detection, *Cureus*, 2024, Vol. 16.

[10]. J. Zech, S. Santomartino, and P. Yi, Artificial intelligence (AI) for fracture diagnosis: an overview of current products and considerations for clinical adoption, *American Journal of Roentgenology*, Vol. 219, 2022, pp. 869–878.

[11]. J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, You only look once: Unified, real-time object detection, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 779–788.

[12]. L. Duron, A. Ducarouge, A. Gillibert, et al., Assessment of an AI aid in detection of adult appendicular skeletal fractures by emergency physicians and radiologists: a multicenter cross-sectional diagnostic study, *Radiology*, Vol. 300, 2021, pp. 120–129.

[13]. L. Canoni-Meynet, P. Verdot, A. Danner, P. Calame, and S. Aubry, Added value of an artificial intelligence solution for fracture detection in the radiologist's daily trauma emergencies workflow, *Diagnostic and Interventional Imaging,* Vol. 103, 2022, pp. 594–600.

[14]. A. Guermazi, C. Tannoury, A. Kompel, et al., Improving radiographic fracture recognition performance and efficiency using artificial intelligence, *Radiology*, Vol. 302, 2022, pp. 627–636.

[15]. N. Regnard, B. Lanseur, J. Ventre, et al., Assessment of performances of a deep learning algorithm for the detection of limbs and pelvic fractures, dislocations, focal bone lesions, and elbow effusions on trauma X-rays, *European Journal of Radiology*, Vol. 154, 2022, article 110447.

[16]. M. Dupuis, L. Delbos, R. Veil, and C. Adamsbaum, External validation of a commercially available deep learning algorithm for fracture detection in children, *Diagnostic and Interventional Imaging*, Vol. 103, 2022, pp. 151–159.

[17]. T. Nguyen, R. Maarek, A. Hermann, et al., Assessment of an artificial intelligence aid for the detection of appendicular skeletal fractures in children and young adults by senior and junior radiologists, *Pediatric Radiology,* Vol. 52, 2022, pp. 2215–2226.

[18]. M. Cohen, J. Puntonet, J. Sanchez, E. Kierszbaum, M. Crema, P. Soyer, and E. Dion, Artificial intelligence vs. radiologist: accuracy of wrist fracture detection on radiographs, *European Radiology*, Vol. 33, 2023, pp. 3974–3983.

[19]. V. Bousson, G. Attané, N. Benoist, et al., Artificial intelligence for detecting acute fractures in patients admitted to an emergency department: real-life performance of three commercial algorithms, *Academic Radiology,* Vol. 30, 2023, pp. 2118–2139.

(013)

Prototype Networks for Reliable Classification Decision based on Gene Expressions for Breast Cancer Detection Integrating Expert Knowledge

T. Villmann [1,2,3], J. Voigt [1,2] and M. Kaden [1,2]
[1] University of Applied Sciences Mittweida, Technikumplatz 17, 09648 Mittweida, Germany
[2] Saxon Institute for Computational Intelligence and Machine Learning, 09648 Mittweida, Germany
[3] Technical University Bergakademie Freiberg, 09599 Freiberg, Germany
Tel.: + 49 3727 5821124
E-mail: {villmann | voigt5 | kaden1}@hs-mittweida. de

Summary: We propose a shallow AI model for reliable detection of breast cancer subtypes based on gene expression data determining the survival probability. This prototype-based model provides easy interpretability of classification decisions and allows to integrate domain knowledge available from bio-medical databases. Decision robustness is obtained by the inherent large classification margin property of the model, which also can be used to establish a reject option for declining model suggestions in case of uncertain decision. Model interpretability is confirmed by the vector quantization approach for the prototypes together with the relevance learning extension, which optimizes the classification decision exploiting those correlation between data features contributing to better class separation. Additionally, the model can be used to mitigate or to remove suspected bias information – here the menopause information. Finally, this prototype-based model is easy to combine with deep learning approaches such that it is working in the resulting embedding space determined by the deep backbone. We demonstrate the model abilities for breast cancer subtypes identification.

Keywords: Cancer gene expression analysis, Robust classification, Interpretable prototype models, Knowledge integration, Bias mitigation, Federated learning.

1. Introduction

Class detection in medical data is still an important topic for AI-applications in diagnosis or for decision support systems. Thereby, the model prediction is required to be made with high certainty avoiding misclassification/incorrect predictions. Besides this, robustness and interpretability of the model decisions is crucial for acceptance, usefulness and reliability of the AI-decision making [1]. Other challenges being also of particular importance for AI-training in medicine and healthcare are data privacy, data merging and federated learning [2]. Further, bias detection and its removal from data as well as domain-knowledge integration in AI for improved inference are key ingredients for supporting personalized/precision medicine and, hence, in the focus of AI development.

Currently, AI applications in medicine are dominated by deep learning models [3], which frequently provide impressive performance but lack the above-mentioned properties and demands. Yet following [4], the development of less complex but better interpretable and robust models with high flexibility is more challenging. Here, we propose to utilize prototype-based approaches related to vector quantization as a promising alternative in case of class detection focusing on the model family of learning vector quantizers (LVQ) [5]. We explain how LVQ models contribute to better AI-decision understanding, acceptance and reliability while yielding robustness and flexibility. Moreover, we review useful extensions of the basic models which are particularly interesting in medical data analysis including feature relevance learning (weighting) as well as bias detection and related bias mitigation in the data.

We illustrate the faithful application of these concepts for the medical challenge of breast cancer subtype detection. Yet, in this contribution we do not emphasize and investigate the medical aspects and conclusions which we would leave to the medical experts. Instead, we will concentrate on how to use these approaches efficiently and what interpretation abilities are provided with those LVQ-models.

2. Learning Vector Quantization – An AI-Tool for Reliable Data Classification

LVQ models suppose vectorial data provided as feature vectors $\vec{x} \in \mathbb{R}^n$ or embedded data vectors obtained by a deep backbone of maybe patients data, which have to be classified with respect to predefined classes. For this purpose, a finite set $\Pi \subset \mathbb{R}^n$ of class responsible prototypes $\vec{p}_k \in \Pi$ is assumed in the data/embedding space such that each class is represented by at least one prototype. The class label of a prototype $\vec{p}_k$ is denoted by $c(\vec{p}_k)$.

The prototypes are optimized during training based on the nearest-prototype-principle (NPP) regarding a given dissimilarity measure d, usually the (squared) Euclidean distance

$$d_E^2(\vec{x}, \vec{p}_k) = (\vec{x} - \vec{p}_k)^T \cdot (\vec{x} - \vec{p}_k) \qquad (1)$$

and a training set $T = \{(\vec{x}, c(\vec{x}))\}$ where $c(\vec{x})$ is the class label of the training data vector $\vec{x}$. Respective loss

functions are the overall misclassification loss or sum of local cross-entropy-losses. After training, a new data vector $\vec{x}$ is assigned to the class $c(\vec{p}_{k^*})$, where $\vec{p}_{k^*}$ is the closest prototype to $\vec{x}$, i.e. we set $c(\vec{x}) = c(\vec{p}_{k^*})$. This scheme is known as nearest prototype classification (NPC), see Fig. 1.

The Generalized LVQ (GLVQ) is a mathematically verified LVQ variant defining the state-of-the-art in interpretable classification learning [5]. In GLVQ training, the prototypes $\vec{p}_k \in \Pi$ are adjusted by stochastic gradient descent learning (SGDL) regarding the local loss $l(\vec{x}, \Pi)$ determined by the overall loss. If the class representing prototype-property is required, the classification loss must be super-imposed by a representation loss based on the NPP. GLVQ is proven to be robust against noise and perturbations in data (large hypothesis margin approach) and reflects class dependent data properties after training [6]. The margin information can be used to implement reject rules for uncertain classification decisions in case of class borderline samples [7].

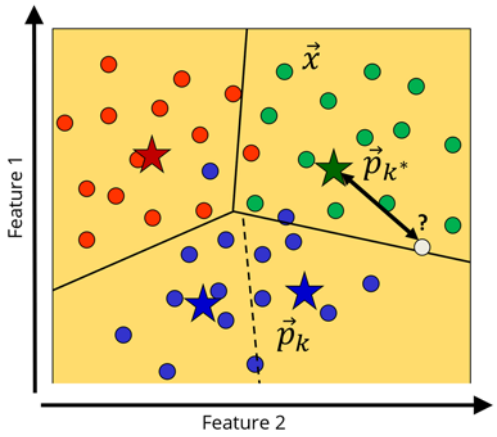


Fig. 1. Illustration of the LVQ model and the nearest prototype classification decision for two-dimensional data: The colored circles stand for training data vectors $\vec{x}$, the stars for prototypes $\vec{p}_k$ and the color indicates their class label. The white circle is an unknown sample $\vec{x}$ assigned via the nearest prototype $\vec{p}_{k^*}$ to the corresponding (green) class, i.e. $c(\vec{x}) = c(\vec{p}_{k^*})$. The black lines indicate the class decision borders for unknown samples whereas the dashed line indicates the border of the decision regions of the blue-class prototypes and, hence, is not a class decision border.

In the following we briefly highlight modifications of the basic GLVQ for better performance, interpretability, bias mitigation and federated learning.

2.1. Relevance Learning in GLVQ

Relevance Learning in GLVQ determines data features contributing to an improved class separability. For this purpose, data-are transformed by

$$\vec{v} = \Omega\vec{x} + \vec{s} \tag{2}$$

where $\Omega \in \mathbb{R}^{m\times n}$ is a matrix realizing a linear map with $m \leq n$ and $\vec{s} \in \mathbb{R}^m$ is a data shift. Now, the prototypes are in the mapping space, i.e. $\vec{p}_k \in \mathbb{R}^m$ and the squared parametrized Euclidean distance reads as

$$d^2_\Omega(\vec{x}, \vec{p}_k) = (\Omega\vec{x} + \vec{s} - \vec{p}_k)^T \cdot (\Omega\vec{x} + \vec{s} - \vec{p}_k) \tag{3}$$

with the mapping matrix Ω and the shift $\vec{s}$ which has to be adjusted during training – additionally to the prototypes $\vec{p}_k$ but again by SGDL. The resulting Classification-Correlation-matrix (CCM)

$$\Lambda = \Omega^T\Omega \tag{4}$$

reflects those data feature correlations by non-zero matrix entries, which support the classification decision [6]. The CCM-profile vector $\vec{\lambda}$ with entries

$$\lambda_i = \sum_j |\Lambda_{i,j}| \tag{5}$$

contains the respective feature relevancies and is denoted as Λ-classification-influence-profile (Λ-CIP). Otherwise, the matrix

$$\Upsilon = \Omega\Omega^T \tag{6}$$

yields the classification-influence-profile vector $\vec{\upsilon}$ (Υ-CIP) with entries

$$\upsilon_i = \sum_j |\Upsilon_{i,j}| \tag{7}$$

giving this information for the mapping space $\mathbb{R}^m$. This GLVQ variant is known as Matrix-relevance GLVQ (GMLVQ) [9].

Often, the mapping matrix $\Omega \in \mathbb{R}^{m\times n}$ contains a huge number of entries if m and n are large as it is usually the case for medical imaging [10] or gene-expression data analysis in cancer research [11,12]. If *external domain-knowledge* is available, e.g. hierarchical feature dependencies, this information can be used to decompose the mapping matrix Ω into a product of sparse matrices representing this knowledge – for example gene associations – provided by gene-knowledge databases like KEGG or Reactome in bio-medical context. Adaptation of these sparse matrices in GMLVQ becomes easier and the adjusted sparse factor matrices frequently allow causal conclusion regarding the influence of features their dependencies [13].

Further, the resulting sparsity for the knowledge-informed mapping matrix Ω contributes to an inherent regularization of the learning process for the matrix adaptation in GMLVQ.

2.2. Data Bias Mitigation and Federated Learning

A suspected bias implicitly contained in the data can be detected by a GMLVQ, if the bias information $b \in B$ is available for training data [15]. The bias is detected if a GMLVQ model is trained using the bias as class information and Ω_B as the corresponding mapping matrix to be adjusted during training. If the resulting GMLVQ achieves a performance better than random guess, bias-caused data deviations are detected. Taking the projector P_{Ω_B} to map the data into the null-space of Ω_B, i.e. the projection $P_{\Omega_B}\vec{x}$ delivers an unbiased data representation and, hence, the bias information is removed from the original data [16].
For example, this strategy can be used to realize federated learning of data coming from different institutions with privacy constraints: In this case, privacy (institute) information is taken as bias such that the bias-corrected data do not contain privacy information regarding the data source (institution) anymore [10]. This might be of interest when several clinical institutes and hospitals want to share their data but have to ensure privacy.

3. GLVQ – Breast Cancer Subtype Detection

As an illustrative application of the GMLVQ based classification learning and classification analysis we demonstrate the proposed approach for breast cancer subtype detection by means of gene expression analysis for METABRIC-data [14].

The respective gene expression data set contains 1700 expression profiles $\vec{x} \in \mathbb{R}^{652}$ of 488 genes for 1700 patients and is obtained from the Kaggle-database. Four cancer sub-types form the class information. Hierarchical gene knowledge and related pathway/biological process information is taken from the KEGG database resulting in the matrix decomposition $\Omega_{KEGG} = \Omega \cdot A_G$ with $\Omega = K_B \cdot K_P$ being the adjustable matrix for GMLVQ and where the matrix wit $A_G \in \{0,1\}^{488\times652}$ describes the fixed assignment of the profile entries to the genes and, hence cannot be adjusted. The structured matrix $K_P \in \mathbb{R}^{430\times488}$ contains information about which genes possibly can contribute to the 430 involved pathways, and the matrix $K_B \in \mathbb{R}^{57\times430}$ represents possible path correlations contributing to 57 biological processes related to the breast cancer subtypes.

Hence, taking the information from the KEGG-database into account, we have for the adjustable matrix $\Omega \in \mathbb{R}^{57\times488}$ less than 10 % non-zero entries, which can be adapted during the classification learning.

GMLVQ classification learning for both prototypes and mapping matrix Ω, with the described knowledge integration achieves 85.0% test accuracy using ten-fold cross validation and three prototypes for each cancer subtype (class). The resulting CCM Λ is from eq. (4) depicted in Fig. 2, whereas Fig. 3 represents the matrix Υ from eq. (6).

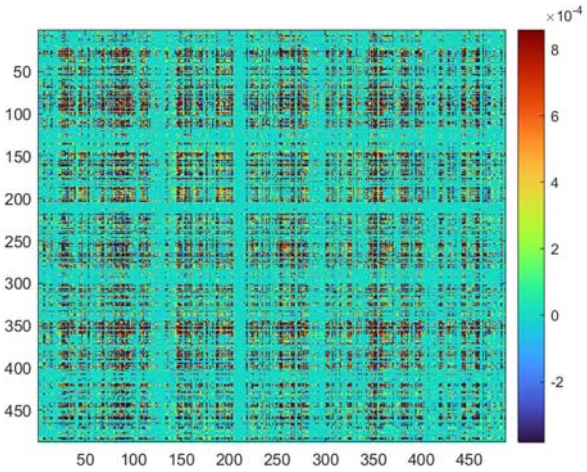


Fig. 2. Color coded visualization of the CCM $\Lambda = \Omega^T\Omega$. The entries of this matrix reflect gene correlations contributing to the cancer subtype classification. The horizontal and the vertical axes represent the considered genes.

The evaluation of these matrices has to be done by the medical and bio-chemic experts, which may require further in-depth clinical investigations.

Further, the pre-post-menopause information available for all profiles is a suspected bias possibly affecting the original data. The respective GMLVQ-based bias detection using the mapping matrix Ω_B results in 86.2% prediction accuracy revealing strong implicit bias to the data, which could have influenced the original GMLVQ performance in cancer subtype classification.

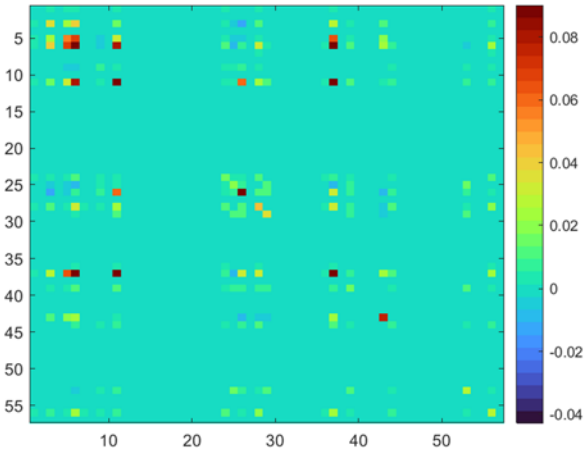


Fig. 3. Color coded visualization of the matrix $\Upsilon = \Omega\Omega^T$ from eq.(5). The horizontal and the vertical axes represent the biological processes involved. The entries of this matrix indicate those correlations between the biological processes, which contribute to the cancer subtype classification.

To prove this hypothesis, we trained another GMLVQ model again with three prototypes per subtype but now using the bias-mitigated data $P_{\Omega_B}\vec{x}$ which yields a small classification accuracy improvement of 1.0% compared to the GMLVQ for

biased data. Thus, the detected bias contributes only marginally to breast cancer subtype detection, and we are allowed to continue to evaluate the former GMLVQ model trained with the original data.

The resulting Λ-CIP vector $\vec{\lambda}$ of the 488 involved genes obtained from the CCM Λ according to eq. (5) is depicted in Fig. 4 reflecting the importance of the considered genes for class discrimination.

Analogously, the ϒ-CIP vector $\vec{\upsilon}$ resulted by eq. (7) gives the equivalent information for the biological processes, see Fig. 5.

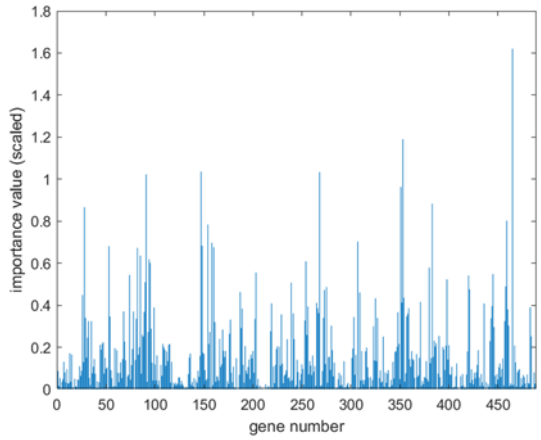


Fig. 4. Classification-influence-profile Λ-CIP for the 488 genes considered for breast cancer subtype classification. The higher the bar, the higher is the evidence for contribution to class discrimination.

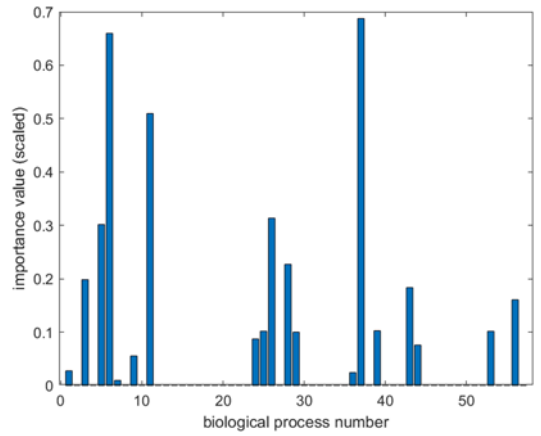


Fig. 4. Classification-influence-profile ϒ- CIP for the 57 biological (pathway) processes possibly contributing to cancer subtype detection. The higher the bar, the higher is the evidence for contribution to class discrimination.

4. Conclusions

In this contribution we demonstrate the ability of shallow interpretable LVQ approaches with easy to implement extensions for reliable medical data analysis for diagnosis support. In particular, we focus on relevance analysis of the contributing data features for improved classification abilities and model interpretation. Further, we demonstrate possibilities to deal with high-dimensional feature vector by knowledge integration the user might have about the data features. Thereby, this knowledge obviously leads to an inherent regularization of the model training process.

Additionally, we have shown, how this GMLVQ method can be used for the detection of a suspected bias. Respective GMLVQ learning also provides a mathematical concept to mitigate this bias in the data.

The methodical explanations about classifier model and the respective extensions are demonstrated for the challenge of breast cancer subtype detection. Yet, we emphasize here that the investigations only have illustrative character from a machine learning perspective and should be verified for real clinical finding by medical experts.

Finally, it should be noted that GLVQ can easily be combined with a deep model as a backbone. Yet, in this case interpretability is limited due to the general difficulty interpreting the deep model appropriately. However, taking the deep backbone as a somewhat intelligent pre-processing of the data, a GLVQ prototype layer instead of the usual (dense) perceptron layer in deep networks leads to a large margin classifier in the mapping space increasing the robustness of the classification decision due to the robustness properties of GLVQ [17,18].

References

[1]. P. Lisboa, S. Saralajew, et al., The coming of age of interpretable and explainable machine learning models, *Neurocomputing*, Vol. 535, 2023, pp. 25-39.

[2]. A. Vellido, The importance of interpretability and visualization in machine learning for applications in medicine and health care, *Neural Computing Applications*, Vol. 32, Issue 24, 2020, pp. 8069-18083.

[3]. D. Bacciu, P. Lisboa, et al. Deep Learning in Biology and Medicine, *World Scientific*, 2022.

[4]. L. Semenova, C. Rudin, et al., On the existence of simpler machine learning models, in *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAccT'22)*, Seoul, Republic of Korea, 21-24 June 2022, pp. 1827-1858.

[5]. M. Biehl, B. Hammer, et al., Prototype-based models in machine learning, *Wiley Interdisciplinary Reviews: Cognitive Science*, Vol. 7, Issue 2, 2016, pp. 92-111.

[6]. K. Crammer, R. Gilad-Bachrach, et al., Margin analysis of the LVQ algorithm, in Advances of Neural Information Processing Vol. 15 (S. Becker, S. Thrun, K. Obermayer, Eds.), *MIT Press*, 2003, pp. 462-469.

[7]. L. Fischer, B. Hammer, et al., Efficient rejection strategies for prototype-based classification, *Neurocomputing*, Vol. 169, 2015, pp. 334-342.

[8]. T. Villmann, M. Kaden et al., Self-Adjusting Reject Options in Prototype Based Classification, in Advances in Self-Organizing Maps and Learning Vector Quantization: Proceedings of 11th International Workshop (WSOM 2016), E. Merényi, M. J. Mendenhall, P. O'Driscoll, P. (Eds.), Houston, Vol. 428, Advances in Intelligent Systems and Computing, *Springer*, 2017, pp. 269-279.

[9]. P. Schneider, B. Hammer, et al., Adaptive Relevance Matrices in Learning Vector Quantization, *Neural Computation*, Vol. 21, 2009, pp. 3532-3561.

[10]. R. van Veen, N.R. Bari Tambolia, et al., Subspace Corrected relevance learning with application in Neuroimaging, *Artificial intelligence in Medicine*, Vol. 149, 2024, No. 102786.

[11]. H. Elmarakeby, J., Hwang, et al., Biologically informed deep neural network for prostate cancer discovery, *Nature*, Vol. 598, 2021, pp. 348-352.

[12]. J. Oldenburg, J. Wagner et al., XModNN: Explainable Modular Neural Network to Identify Clinical Parameters and Disease Biomarkers in Transcriptomic Datasets, *Biomolecules*, Vol. 14, 2024, No. 1501.

[13]. J. Voigt, S. Saralajew, et al., Biologically-informed shallow classification learning integrating pathway knowledge, in *Proceedings of the 17th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC'2024)*, Rome, Italy, 21-23 February 2024, Vol. 1, pp. 357-367.

[14]. C. Curtis, S.P. Shah, et al., The genomic and transcriptomic architecture of 2,000 breast tumors reveals novel subgroups, *Nature*, Vol. 486, 2012, Issue 7403, pp. 346-352.

[15]. F. Störck, F. Hinder, J. Brinkrolf, B. Paaßen, V. Vaquet, and B. Hammer. FairGLVQ: Fairness in partition-based classification. in T. Villmann, M. Kaden, T. Geweniger, and F.-M. Schleif, editors, in *Proceedings of the 15th Workshop on Self-Organizing Maps, Learning Vector Quantization and Beyond (WSOM+ '2024)*, 2024, pp. 141–151.

[16]. M. Kaden, A. Engelsberger, et al., Mitigating the bias in data for fairness using an advanced generalized learning vector quantization approach. In M. Verleysen, editor, in *Proceedings of the 33rd European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN'2025), Bruges (Belgium),* Louvain-La-Neuve, Belgium, in press.

[17]. K. Crammer, R. Gilad-Bachrach et al., Margin analysis of the LVQ algorithm, in Advances in Neural Information Processing (Proc. NIPS 2002), S. Becker, S. Thrun, K. Obermayer (Eds.), Vol. 15, 2003, pp. 462-469.

[18]. S. Saralajew, L. Holdijk, T. Villmann, Fast Adversarial Robustness Certification of Nearest Prototype Classifiers for Arbitrary Seminorms. In H. Larochelle, M. Ranzatoet al. (Eds.) *Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS 2020)*, Vol. 33, 2020, pp. 13635-13650.

(015)

Explaining Classifications of Raman Spectral Data based on Quotient Games

M. Piazza [1], M. Passacantando [2], M. Bedoni[3] and E. Messina [1]
[1] University of Milano-Bicocca, Department of Informatics, Systems and Communication,
Viale Sarca 336, 20125 Milan, Italy
[2] University of Milano-Bicocca, Department of Business and Law,
Via Bicocca degli Arcimboldi 8, 20126, Milan, Italy
[3] IRCCS Fondazione Don Carlo Gnocchi ONLUS, Via Capecelatro 66, Milan, 20148, Italy
E-mail: enza.messina@unimib.it

Summary: Recent advancements in healthcare technologies have enabled the collection of large biological datasets. The combination of Raman Spectroscopy of biological samples with machine learning has demonstrated potential for early-stage, fast, and cost-effective diagnosis. However, due to the critical nature of these applications, it is essential to provide tools for medical professionals to interpret model decisions. EXplainable Artificial Intelligence (XAI) aims to solve this problem by justifying model predictions. One of the main problems is to explain models with high-dimensional input data, such as Raman spectroscopy. This work introduces a novel XAI method inspired by SHAP, leveraging the concept of cooperative games with a partition structure to address this issue. The efficacy of the proposed method is evaluated through a comparative analysis with traditional SHAP in a real case study using saliva Raman spectra, demonstrating improved interpretability and applicability.

Keywords: Raman spectroscopy, XAI, Diagnosis, Game theory, SHAP.

1. Introduction

Raman spectroscopy (RS) is a technique based on the inelastic scattering of monochromatic light, allowing the extraction of a "fingerprint" that provides information about a sample's chemical composition. This information can be exploited for various applications, including disease diagnosis, by analyzing the spectral information of biological samples such as blood or serum [1, 2]. However, blood collection is an invasive procedure, whereas saliva collection is non-invasive. Despite this advantage, interpreting the spectral fingerprint of saliva remains challenging, even for domain experts, due to the complex interactions between its molecular components.

Recent research has explored the application of Machine Learning (ML) and Deep Learning (DL) for the automated classification of spectral fingerprints [3, 4]. However, these techniques are often considered "*black-box*" models, as their internal decision-making processes are not easily comprehensible to human users. While this lack of transparency is not always problematic, in high-stakes domains such as healthcare, interpretability is crucial for the real-world deployment. Moreover, following the AI Act and GDPR, interpretability is now a legal requirement in the European Union.

The field of eXplainable Artificial Intelligence has emerged over the past decade to provide explanations for the predictions of ML and DL models. However, most explainability approaches are designed and optimized for traditional data structures, such as tabular data and images. Adapting these methods to non-traditional data, such as Raman spectra, is not always straightforward.

A recent survey [5] indicates that the most frequently employed XAI method for RS applications is SHAP [6], but its high computational complexity can limit the applicability in the field, where typical problems are characterized by thousands of features. To address this challenge, most of the SHAP-based approaches implement a feature selection or reduction strategy before classification and explanation, limiting the scope of the results [7, 8]. There are also approaches specific to Neural Networks, with the two most important being grad-CAM [9, 10] and Integrated Gradients [11]. However, these approaches require the use of a specific DL architecture, limiting the scope for possible comparisons. Moreover, many of the reviewed studies do not take into account the unique characteristics of spectral data. Specifically, since feature positions correspond to particular Raman shifts, which in turn correspond to molecular fingerprints, XAI attributions should reflect this continuous and structured nature of the spectrum, rather than just treating features as independent values.

In this work, we propose the use of a novel formulation of SHAP based on cooperative games with partition structures [12], specifically designed to account for the spectral structure of data. The proposed approach has two main objectives: first, it enables the efficient application of SHAP-based methods directly to raw spectra, eliminating the need for feature selection or dimensionality reduction that could disrupt their inherent structure. Second, it produces interpretability results at the level of Raman-shift intervals, reducing the impact of noise in spectral data. This is particularly important since molecular components and spectral features are often associated

with Raman-shift intervals rather than individual features, such as fixed-position peaks.

Since evaluating XAI approaches remains an open problem [13], we also propose an evaluation framework that integrates quantitative metrics with a qualitative analysis of the results based on domain expert knowledge. As metrics we used *Faithfulness* [14] and the *Remove-On-And-Retrain* (ROAR) paradigm [15]. The proposed method has been validated on a real-world dataset for the Covid-19 [16] diagnosis from saliva Raman spectra and compared with traditional SHAP. The computational experiments suggest that the Quotient Game is more effective than the baseline in identifying sensitive regions and assigning them a reliable importance score.

2. Related Works

Explainable Artificial Intelligence is a rapidly evolving research field, with numerous challenges and open questions still emerging. Similarly, the application of ML models for the automatic classification of Raman spectra is relatively new, with only a few studies [4, 16] thoroughly investigating this combined approach. Due to the novelty of both fields, the extensive application and evaluation of XAI methods for Raman spectral data remain limited, as also highlighted in recent literature [5].

Moreover, Raman spectral data has unique characteristics that complicate the direct adaptation of XAI methods developed for other data types. Each spectral point corresponds to a specific position on the Raman shift, and groups of points collectively form molecular signatures. Consequently, analyzing individual points in isolation is insufficient for capturing the true influence of molecular structures on classification outcomes. Furthermore, Raman spectra are inherently high-dimensional, leading to significant computational complexity for traditional XAI methods, necessitating the use of approximation techniques.

The most commonly used approach for explaining ML models in Raman Spectroscopy is SHAP [5]. However, its computational complexity grows exponentially, with a complexity of 2^N (corresponding to the number of possible coalitions), where N represents the number of features. In a typical Raman spectral problem, N is approximately 1000, making direct computation infeasible. Surveyed studies suggest two main strategies to address this issue: approximating the Shapley Value [17] or limiting the analysis to a subset of features obtained through feature selection or dimensionality reduction techniques [18].

Each approach presents challenges. In the case of approximation, if the number of sampled coalitions is significantly smaller than the total possible set, the quality of the approximation becomes very poor. On the other hand, applying feature selection or reduction modifies the spectral structure of the data, potentially compromising the reliability of the final results.

The second major family of approaches applied to Raman spectroscopy consists of Class Activation Mapping (CAM) [19] and Grad-CAM [9, 10]. These methods analyze the gradient flow in convolutional layers of a Convolutional Neural Network (CNN) to estimate an importance score for each input feature. Their results are typically visualized as heatmaps, enabling users to quickly identify the most sensitive spectral regions. However, these methods have a significant limitation: they require a specific model type (Neural Networks) and, more specifically, a CNN architecture. Closely related to these methods are gradient-based approaches [21, 22], which analyze gradient information across the entire neural network. While these methods share the advantages of activation-based techniques—such as heatmap-based visualization—they also inherit the same constraints, as they are only applicable to neural network-based classifiers.

It is worth mentioning LIME [23], one of the most widely used methods in XAI. Some applications for spectral data have been reported in [24, 25]. However, its reliability is limited when applied to high-dimensional problems, as it relies on a linear surrogate model, which may struggle to accurately approximate complex relationships in the data.

We also include in this literature review existing methods designed to natively apply SHAP to high-dimensional data. The two most common approaches involve a hierarchical framework [26] or the use of the "Group-Shapley" concept [27]. However, these methods have primarily been applied to image and financial data and additionally they do not relate their approaches to the cooperative Game Theory literature.

To the best of our knowledge, this work represents the first attempt to propose the use of a novel SHAP-based formulation—still based on Game Theory concepts—specifically tailored for high-dimensional spectral data in the healthcare domain.

3. Methodology

3.1. SHAP

SHAP is a local, post-hoc XAI technique based on Game Theory. The term 'local' refers to methods that provide an individual explanation for each specific instance, while 'post-hoc' indicates that explanations are generated only after the model has been fully trained and considered as fixed. The core idea is to model the features of a ML problem as players in a Transferable Utility (TU) cooperative game and calculate their Shapley values [28] as sensitivity scores for feature importance. A cooperative game is a type of game in which coalitions of players (features) work together to maximize a collective reward represented by the marginal contribution of the model with respect to a random classifier. SHAP fairly divides this reward among the players based on their impact. For a detailed description of the method refer to [6]. Since the number of possible coalitions increases exponentially

with the number of features (2^N where N is the number of players/features) and the computational complexity is directly proportional to this number, this approach becomes infeasible for high-dimensional data. To address this problem, we formulate the game as a cooperative game with a partition structure and solve it using the Quotient Game [12].

3.2. Cooperative Game with Partitions

The Quotient Game applies to transferable utility cooperative games with a partition structure, i.e., where certain players must cooperate and cannot be assigned to different coalitions. In such settings, conventional cooperative game theory approaches are not directly applicable. Instead, it is necessary to employ methods that explicitly consider the partition constraint, such as the Owen Value [29]. In this work, we adopt the Quotient Game approach, where partitions are elevated to the level of players.

Formally, given a TU cooperative game (N, v) with the set of players partitioned into disjoint subsets $N_1, \dots, N_m$, the quotient game is defined as (M, v^P), where $M = 1, \dots, m$ represents the partitions and the characteristic function is defined as $v^P(S) = v(\cup_{i \in S} N_i)$ for any $S \subseteq M$. This formulation enables the application of standard solution concepts; in this work, we focus on the Shapley Value:

$$\phi(v) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!\,(|N| - |S| - 1)!}{|N|!} [v(S \cup i) - v(S)] \quad (1)$$

Partition structures can be defined in various ways, including hierarchical clustering, mutual information, or tree-based methods. Given the challenge of automatically identifying molecular fingerprints in spectra, we adopt a straightforward strategy: dividing the input spectrum into equal-sized intervals, with the aim of balancing computational complexity with interpretability. The number of partitions was controlled by a parameter π, initially set to 8 to balance result fidelity with the computational feasibility of the Shapley Value. We also examined how varying this parameter affected the final results and reported this analysis in Section 4.3.

3.3. Dataset

The dataset consists of Raman spectra extracted from saliva samples. Each spectrum contains 991 measurement points corresponding to intensity values (Y-axis in Fig. 1) obtained through the Raman procedure. These intensity values are measured at specific Raman shift positions (X-axis in Fig. 1) in units of cm^{-1}. The total number of subjects involved in the study was 101, composed as follows: 30 patients affected by COVID-19, 37 subjects negative to the SARS-CoV-2 test with a past episode of COVID-19, and 34 age and sex-correlated healthy subjects. To reduce the noise related to specific acquisitions about 25 spectra have been acquired for each patient, resulting in a total of 2408 spectra. More information regarding the acquisition protocol and the patients' selections is available in [16].

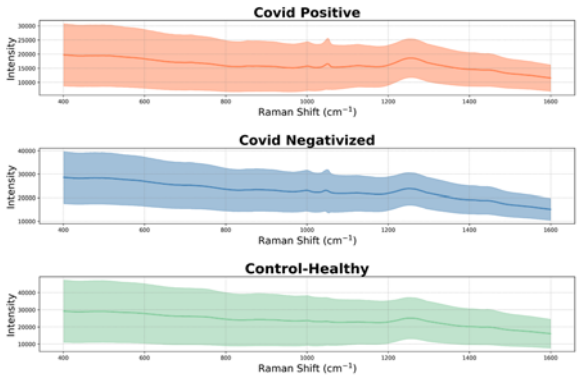


Fig. 1. Comparison between average and standard deviation of spectra belonging to different classes.

3.4. Evaluation Metrics

The definition and formalization of Explainable AI fundamental concepts is still evolving [30, 31]. Consequently, there is no universally accepted framework for evaluating and comparing different methodologies. As highlighted in recent work [13], many novel approaches heavily rely on manual result analysis, introducing risks of confirmation bias and an inclination toward favoring the proposed method or selectively choosing examples that reinforce its effectiveness. This lack of standardization complicates the objective assessment of explainability techniques, making it essential to establish rigorous evaluation methodologies.

To address this issue, the authors of [13] propose a taxonomy of twelve properties under which explanations can be assessed. Among these, in this work, we focus on evaluating the correctness of explanations, specifically determining whether the features identified as most influential by the model truly contribute to its decision-making process. To this end, we incorporate Faithfulness [14], one of the most widely adopted metrics due to its computational efficiency, in combination with ROAR [15], a more computationally intensive but robust approach.

Faithfulness [14] assesses the reliability of feature attributions by measuring the change in model predictions when an individual feature (or feature subsets) are removed. Since the model is fixed after training, feature removal is typically simulated using approaches such as baseline substitution, marginal distribution sampling, or noise injection [31]. The Pearson correlation is commonly used to quantify the relationship between these variations and the feature importance scores assigned by the XAI method. Due to its low computational complexity, Faithfulness serves as an efficient preliminary evaluation of explanation reliability.

In contrast, the ROAR [15] paradigm provides a more rigorous assessment of explanation quality by analyzing the impact of systematically removing the most important features. Features are ranked by importance, and the model is retrained iteratively, omitting entire spectral regions in decreasing order of importance. This methodology ensures that the absence of key features leads to a measurable decline in model performance, thereby validating the significance of identified features. To enhance robustness, we apply a 10-fold cross-validation approach and monitor the F1-score to quantify the impact of feature removal on classification performance.

4. Results

4.1. Experimental Settings

The experimental setup was designed as follows. Each dataset was split into training and testing sets using a *Leave-One-Patient-Out Cross-Validation* (LOPO-CV) approach. In this study, LOPO-CV operates as a 101-fold cross-validation process, where each iteration assigns all but one patient to the training set. Once the classifier is trained, explanations are generated for the single patient in the test set. This process is repeated iteratively for all patients to ensure comprehensive evaluation. For model training and analysis, we employed a CNN previously proposed in the literature for similar tasks [16, 32].

4.2. Main Results

As already introduced, Faithfulness is used to assess the correctness of explanations by correlating feature importance scores with the drop in classification probability upon feature removal, using Pearson correlation [14]. Its values range from -1 to +1. Average results are reported in the first column of Table 1, while the distribution of the faithfulness value is reported in the boxplots in Fig. 2.

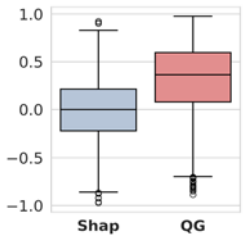


Fig. 2. Boxplots representing the distribution of Faithfulness value for SHAP and Quotient Game.

The QG approach demonstrated higher average value, as well as a distribution that is mostly in the positive direction (Fig. 2). On the other hand, SHAP has a value near zero, which suggests no correlation between SHAP importance score and model's behaviour represented by classification probability.

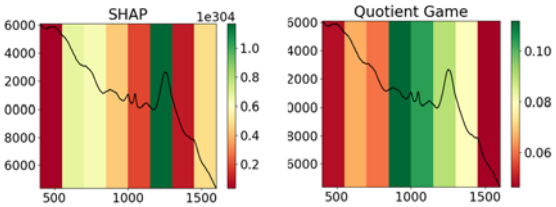
In contrast to Faithfulness, the ROAR methodology assesses the change in model classification performance by retraining the ML model from scratch after removing one or more features [15]. In this study, we monitored the F1-score, focusing the analysis on the two intervals with the highest average importance. The results for this metric are reported in the second and third columns of Table 1. For the first interval, traditional SHAP slightly outperforms QG, although the difference is minimal. Conversely, QG performs better when considering the second interval.

Table 1. Faithfulness and ROAR results. For ROAR the results related to the two best intervals are shown. Note that for Faithfulness' the greater is the better, while for ROAR the lower is better.

	Faithfulness	**ROAR**	
		Best interval	Second best interval
SHAP	0.00 (±0.09)	**0.84 (±0.16)**	0.82 (±0.13)
QG	**0.33 (±0.13)**	0.85 (±0.11)	**0.79 (±0.14)**

Given the lack of standardized evaluation procedures, we complemented our analysis with a qualitative assessment by comparing the findings with established patterns and domain knowledge from the scientific literature. According to domain experts [16], critical spectral peaks for Covid-19 are expected at approximately 1048 $cm^{-1}$and 1136 cm^{-1}, which can be associated with molecules that can possibly represent Covid-19 biomarkers. For more information it is possible to refer to [16].

The heatmap corresponding to the average importance score for both methods is reported in Fig. 3. SHAP identified the most significant region between ~1015 and ~1138cm^{-1}, closely aligned with the critical interval reported in the literature. On the other hand, QG highlighted an area below 1000 cm^{-1} as the most critical but also identified the second and third most relevant intervals between 892cm^{-1} and 1138cm^{-1}, encompassing the expected sensitive regions.



(a) (b)

Fig. 3. Heatmaps of the average importance score computed with SHAP (a) and Quotient Game (b) for each spectral interval. The average Raman Spectrum is represented by the black line.

4.3. Sensitivity Analysis for Interval Size (π)

To enhance and conclude the analysis and assess the impact of different spectral windowing strategies, we conducted a sensitivity analysis on the parameter π, which represents the number of partitions on which divide the spectrum. We tested partition cardinalities ranging from 6 to 10 and monitored the faithfulness metric for both methods. This parameter is expected to have a greater effect on QG, as it natively incorporates the partition structure, whereas SHAP results at the partition level are obtained by aggregating fine-grained outputs. We excluded the ROAR metric from this analysis due to its requirement for full model retraining, which is too computationally intensive. In Fig. 4 we reported the comparison of average faithfulness value between QG and SHAP for different values of parameter π.

As it is possible to observe, SHAP values remain relatively constant, indicating that this parameter has a small effect on the final results. Similarly, QG values are mostly stable, although a slight increase is observed when using nine spectral windows. To gain deeper insight into this effect, in Fig. 5 we reported the results for the Quotient Game alone.

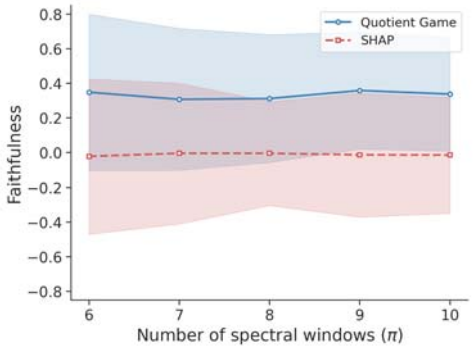


Fig. 4. Comparison of average faithfulness value between QG and SHAP for different values of parameter π.

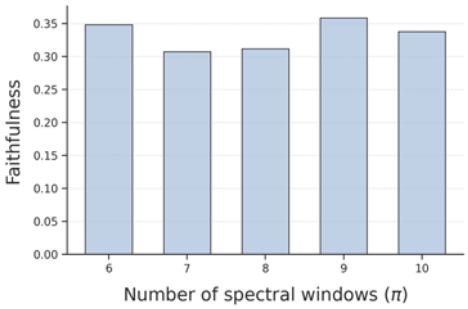


Fig. 5. Average faithfulness value for different π values for Quotient Game approach.

The best results are achieved when the parameter π is equal to 9, although differences between window sizes are not significant. The choice of window size may primarily influence how end users perceives the results, suggesting that future analysis should involve the system's final users

5. Conclusions

In this work, we presented a novel approach based on Game Theory for XAI applied to Raman Spectroscopy. Unlike traditional SHAP, the proposed method has lower computational complexity that allows the computation of the exact Shapley Value also when facing problems with a high number of features. For this reason, it can be directly applied to raw spectra, facilitating the diffusion of transparent classification algorithms in the healthcare domain and providing domain experts with a tool for understanding and debugging ML decisions. Additionally, as demonstrated by the computational experiments, the proposed method is more robust and reliable than traditional SHAP when applied to RS data both from a quantitative and qualitative perspective.

Future work includes applying refined optimization-based approaches to identify intervals in Raman spectra and exploring solution concepts from game theory literature for games with partition structure, such as the Owen Value [29]. Additionally, to validate the proposed method in high-dimensional contexts, it is essential to extend its application to 2D data.

Acknowledgements

This work has been supported by Fondazione Regionale per la Ricerca Biomedica, project CORSAI ID 383 - JTC 2021 ERA PerMed, GA 779282 - CUP: H45E22000030006.

References

[1]. S. Deepaisarn, C. Vong, and M. Perera, Exploring Machine Learning Pipelines for Raman Spectral Classification of COVID-19 Samples, in *Proceedings of the 14th International Conference on Knowledge and Smart Technology (KST' 2022),* Chon Buri, Thailand 26-29 January 2022, pp. 51–56.

[2]. S. Chen et al., Identifying non-muscle-invasive and muscle-invasive bladder cancer based on blood serum surface-enhanced Raman spectroscopy, *Biomedical Optics Express*, Vol. 10, No. 7, 2019, p. 3533.

[3]. Carlomagno et al., Identification of the Raman Salivary Fingerprint of Parkinson's Disease Through the Spectroscopic– Computational Combinatory Approach, *Frontiers in Neuroscience*, 26, 15, 2021, 704963.

[4]. C.-S. Ho et al., Rapid identification of pathogenic bacteria using Raman spectroscopy and deep learning, *Nature Communications,* Vol. 10, No. 1, 2019, 4927.

[5]. J. Contrearas, T. Bocklitz, Explainable artificial intelligence for spectroscopy data: a review, *Pflügers Archiv - European Journal of Physiology*, 2024.

[6]. S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in *Proceedings of the 31st International Conference on Neural Information*

Processing Systems (NIPS' 17), Long Beach, California 04-09 December 2017, pp. 4768–4777.

[7]. J. Chi, X. Bu, X. Zhang, L. Wang, and N. Zhang, Insights into cottonseed cultivar identification using Raman spectroscopy and explainable machine learning, *Agriculture*, Vol. 13, No. 4, 2023, p. 768.

[8]. C. J. Shibu, S. Sreedharan, K. M. Arun, C. Kesavadas, and R. Sitaram, Explainable artificial intelligence model to predict brain states from fNIRS signals, *Front. Hum. Neurosci.,* Vol. 16, 2022, p. 1029784.

[9]. R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization, in *Proceedings of the IEEE International Conference on Computer Vision (ICCV' 2017),* Venice, Italy 22-29 October 2017, pp. 618–626.

[10]. C.-Y. Wang, T.-S. Ko, and C.-C. Hsu, Interpreting convolutional neural network for real-time volatile organic compounds detection and classification using optical emission spectroscopy of plasma, *Analytica Chimica Acta,* Vol. 1179, 2021, p. 338822.

[11]. M. Sundararajan, A. Taly, and Q. Yan, Axiomatic attribution for deep networks, in *Proceedings of the 34th International Conference on Machine Learning (ICML, 2017)*, Sidney, Australia 7-9 August 2017, pp. 3319–3328.

[12] R. J. Aumann and J.H. Dreze, Cooperative games with coalition structures, *Int. J. Game Theory*, Vol. 3, No. 4, 1974, pp. 217–237.

[13]. M. Nauta et al., From Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable AI, *ACM Computing Surveys,* Vol. 55, No. 13s, 2023, pp. 1–42.

[14]. D. Alvarez-Melis, Towards robust interpretability with self-explaining neural networks, in *Proceedings of the 32nd International Conference on Neural Information Processing Systems (NIPS' 18)*, Montreal, Canada 02-08 December 2018, pp. 7786–7795.

[15]. S. Hooker, A benchmark for interpretability methods in deep neural networks, in *Proceedings of the 33rd International Conference on Neural Information Processing Systems (NIPS' 19)*, Vancouver, Canada 08-14 December 2019, pp. 9737–9742.

[16]. C. Carlomagno, et al, COVID-19 salivary Raman fingerprint: innovative approach for the detection of current and past SARS-CoV-2 infections, *Scientific Reports*, Vol. 1, 2021, pp. 4943.

[17]. A. Nakanishi, H. Fukunishi, R. Matsumoto, and F. Eguchi, Development of a Prediction Method of Cell Density in Autotrophic/Heterotrophic Microorganism Mixtures by Machine Learning Using Absorbance Spectrum Data, *BioTech,* Vol. 11, No. 4, 2022, p. 46.

[18]. L. Bellantuono et al., An eXplainable Artificial Intelligence analysis of Raman spectra for thyroid cancer diagnosis, *Scientific Reports*, Vol. 13, No. 1, 2023, 16590.

[19]. B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, Learning Deep Features for Discriminative Localization, in *Proceedings of the IEEEE Conference on Computer Vision and Pattern Recognition (CVPR),* Las Vegas, Nevada, 27-30 June 2016, pp. 2921-2929.

[20]. X. Zhang et al., Understanding the learning mechanism of convolutional neural networks in spectral analysis, *Analytica Chimica Acta*, Vol. 1119, 2020, pp. 41–51.

[21]. B. Panos, L. Kleint, and J. Zbinden, Identifying preflare spectral features using explainable artificial intelligence, *Astronomy & Astrophysics,* Vol. 671, 2023, p. A73.

[22]. Y. Zhang et al., Explainable liver tumor delineation in surgical specimens using hyperspectral imaging and deep learning, *Biomedical Optics Express*, Vol. 12, No. 7, 2021, p. 4510.

[23]. M. T. Ribeiro, S. Singh, and C. Guestrin, 'Why Should I Trust You?': Explaining the Predictions of Any Classifier, in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 13 -17 August 2016, San Francisco, California, pp. 1135–1144.

[24]. C. Li et al., Combining Raman spectroscopy and machine learning to assist early diagnosis of gastric cancer, *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy,* Vol. 287, 2023, p. 122049.

[25]. F. Akulich, H. Anahideh, M. Sheyyab, and D. Ambre, Explainable predictive modeling for limited spectral data, *Chemometrics and Intelligent Laboratory Systems,* Vol. 225, 2022, p. 104572.

[26]. J. Teneggi, A. Luster, and J. Sulam, Fast Hierarchical Games for Image Explanations, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 45, No. 4, 2023, pp. 4494–4503.

[27]. K. Lin and Y. Gao, Model interpretability of financial fraud detection by group SHAP, *Expert Systems with Applications,* Vol. 210, 2022, p. 118354.

[28]. L. S. Shapley, 17. A Value for n-Person Games, in Contributions to the Theory of Games (AM-28), Volume II, H. W. Kuhn and A. W. Tucker (Eds.), *Princeton University Press*, Princeton, 1953, pp. 307–318.

[29]. G. Owen, Values of Games with a Priori Unions, in Lecture Notes in Economics and Mathematical Systems, Berlin, *Springer Berlin Heidelberg*, Heidelberg, 1977, pp. 76–88.

[29]. R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, A Survey of Methods for Explaining Black Box Models, *ACM Computing Surveys,* Vol. 51, No. 5, 2018, pp. 1–42.

[30]. L. Longo et al., Explainable Artificial Intelligence (XAI) 2.0: A manifesto of open challenges and interdisciplinary research directions, *Information Fusion,* Vol. 106, 2024, p. 102301.

[31]. H. Chen, I. C. Covert, S. M. Lundberg, and S.-I. Lee, Algorithms to estimate Shapley value feature attributions, *Nature Machine Intelligence*, Vol. 5, No. 6, 2023, pp. 590–601.

[32]. D. Bertazioli, M. Piazza, C. Carlomagno, A. Gualerzi, M. Bedoni, and E. Messina, An integrated computational pipeline for machine learning-driven diagnosis based on Raman spectra of saliva samples, *Computers in Biology and Medicine,* Vol. 171, 2024, p. 108028.

(016)

BLIP-Eye: Vision-Language Pre-training Classification Model for Eye Diseases Using OCT Scans

Khalid Ayed Alharthi [1], Saja A Alshahrani [1], Rasha O Alshahrani [1], Rahaf A Alshahrani [1], Ali Alshahrani [1] and Zhaorui Zhang [2]
[1] Department of Computer Science, College of Computing, University of Bisha,
Bisha 61922, P.O. Box 551, Saudi Arabia
[2] The Hong Kong Polytechnic University, Hong Kong
Tel.: + 00966 536444911
E-mail: kharthi@ub.edu.sa

Summary: Vision-Language Pre-training (VLP) has significantly revolutionized the performance across numerous vision-language tasks various fields, including medicine. In modern ophthalmology, deep learning technologies play a vital role in diagnosing retinal diseases, where early diagnosis is critical for timely intervention and treatment, ultimately preventing disease progression. Despite recent advancements, the current methods still require substantial improvements particularly regarding accuracy, which is crucial for the reliable classification of eye diseases. To address this issue, inspired by Bootstrapping Language-Image Pre-training (BLIP), we propose a novel VLP-based model, called BLIP-Eye, to identify three retinal diseases through features extracted from optical coherence tomography (OCT) scan images. The classification process categorizes retinal conditions into four classes: normal retina, Diabetic Macular Edema (DME), Choroidal Neovascular Membrane (CNM), and Age-related Macular Degeneration (AMD). We evaluate BLIP-Eye on two real-world (OCT) scan datasets and compare it against a state-of-the-art CNN-based approach. The experimental results show that our model significantly outperforms the baseline in terms of accuracy, precision, recall, and F1-score in distinguishing between the three disease types and a normal retina, highlighting its value as a reliable diagnostic tool in ophthalmology. On average, across the two datasets, our proposed model achieved an F1-score of 91.02 %, while baseline model got only 48.97 % in F1-score.

Keywords: Deep learning, BLIP, CNN, Pre-trained Vision Language VLP, Eye Diseases, OCT Images.

1. Introduction

Artificial intelligence (AI) technology has emerged as a foundational element with far-reaching impacts across numerous domains, particularly within the medical field. Among the myriad applications of AI, deep learning stands out as a pivotal innovation, demonstrating remarkable efficacy in emulating cognitive functions through the provision of sophisticated methods for automation and decision-making. This technological advancement has catalyzed significant progress, exemplified by the development of transformers and visual modeling vision technologies. These innovations have notably enhanced diagnostic and therapeutic processes, especially in the realm of ophthalmology. In particular, deep learning has been instrumental in advancing the diagnosis and management of eye diseases, offering unprecedented precision and accuracy. For instance, it has played a crucial role in the diagnosis of retinal conditions through the analysis of optical coherence tomography (OCT) images. By leveraging deep learning algorithms, clinicians can now detect and monitor retinal abnormalities with greater accuracy and efficiency, thereby improving patient outcomes and advancing the field of eye care. Various existing studies on the integration of AI in medical imaging not only augment the capabilities of healthcare professionals but also pave the way for more personalized and effective treatment strategies.

Several studies have utilized Convolutional Neural Networks (CNN) for various classification purposes and tasks in the field of eye diseases (e.g., [1-7]). Mohamed and Marwa presented a study [8] on the classification of eye diseases using CNN to analyze OCT images. The model was able to detect certain diseases early, including Diabetic Macular Edema (DME), Choroidal Neovascular Membrane (CNM), and Age-related Macular Degeneration (AMD). Fig. 1 [9] and Fig. 2 [10] present the original photograph and the OCT images, respectively.

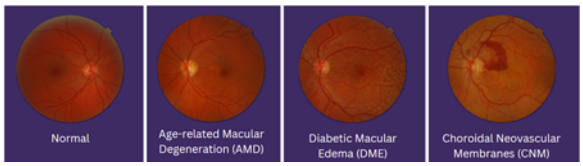


Fig. 1. Comparative View of Retinal Diseases Through Original Photography.

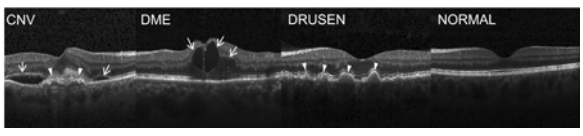


Fig. 2. OCT-Based Comparison of Retinal Diseases and Normal Retina.

This study was conducted by improving the quality of the input images with publicly available data. The most critical drawback includes the high requirements for computer resources and the need for large amounts of data, long training time, in addition to the possibility of bias in the data. A group of researchers from the Chinese University of Hong Kong and King's College London have developed a study on VisionFM [11], an AI model trained on 3.4 million eye images to diagnose 12 common diseases such as diabetic retinopathy and glaucoma. The model demonstrates a high accuracy and a strong ability to generalize across multiple imaging devices and diseases. However, it has certain drawbacks, including potential bias, relatively slow performance due to its size, and occasional inaccuracies in diagnostic descriptions, particularly in rare cases.

- Due to the lack of accurate and fast technologies in the field of ophthalmology for classification tasks, and the challenges that CNN models face in this area, we propose a novel Vision-Language Pre-training (VLP)-based model, called BLIP-Eye, which combines rapid classification of retinal diseases with high accuracy even with limited datasets. Our solution was trained on two real-world OCT scan datasets, making it an effective diagnostic tool in the medical field. The model demonstrates a high capability to classify diseases with high accuracy and speed, enhancing the applicability of VLP-based model in medical contexts as an effective diagnostic tool. This research aims to diagnose retinal disorders and detect the disease early, which is important for many reasons: (1) early treatment before the disease worsens, (2) ensuring patients receive an accurate diagnosis, and (3) accelerating the diagnosis process, which provides greater comfort to the patient. Accordingly, we adapt BLIP to categorize OCT images into four classes: normal retina, Diabetic Macular Edema (DME), Choroidal Neovascular Membrane (CNM), and Age-related Macular Degeneration (AMD). We list our contributions as follows:
- We propose a novel VLP-based model, called BLIP-Eye, which combines rapid classification of retinal diseases with a high accuracy even with limited datasets.
- Our solution was trained on two real- world OCT scan datasets making it an effective diagnostic tool in the medical field.
- Extensive experimental results show that our model significantly outperforms the baseline [8] (i.e., CNN-based model) in accuracy, precision, recall, and F1-score across four different classes, emphasizing its efficiency as a reliable medical image classification model in ophthalmology. On two different datasets, BLIP-Eye achieves an F1-score of 91.97% and 90.06%, representing a significant improvement over the baseline model's F1-scores of 48.12% and 49.83%.
- To the best of our knowledge, we are the first to utilize BLIP for medical image classification in ophthalmology, though multiple studies (e.g., [12–19]) have employed BLIP [20] for different classification tasks and applications.

The rest of this paper is organized as follows: Section 2 describes our Motivation. Section 3 discusses related work. Section 4 presents our methodology. Section 5 provides the experimental results for both models, and we finally summarize our work in Section 6.

2. Motivation

Medical diagnosis in the field of ophthalmology is a major challenge for doctors, as medical decisions often depend on the doctor's experience and the results of medical examinations, which may lead to human errors and delays in diagnosis. Remote areas suffer from a shortage of specialized doctors, which makes immediate diagnosis difficult and affects the quality of healthcare. With the rapid progress of deep learning technologies, it has become possible to use artificial intelligence models to analyze medical images. In this research, we employ BLIP model to analyze OCT images and then classify them with an accuracy rate of up to 92%. This research aims to present the BLIP-Eye model for medical diagnosis, which is able to address current challenges and enhance the efficiency of diagnosis in modern healthcare systems. Recently, many medical technologies have relied on artificial intelligence in diagnosis, which solves the problem of the shortage of specialized doctors in the field of diagnosing eye diseases. The high accuracy and speed of the BLIP-Eye model in classifying retinal diseases helps in early detection of the disease. Not only does this research improve the accuracy of diagnosis, but it can also reduce the time cost and improves medical efficiency, allowing doctors to focus on more complex cases.

3. Related Work

With the rapid advancement of artificial intelligence, its applications in healthcare have expanded significantly. One of the areas that has heavily used AI technologies is the medical field, where AI models help in early diagnosis of disease and improve accuracy. Various studies have explored these developments, highlighting their potential and challenges. In NLP for Early Disease Detection was utilized by Waudby et al. (2011) and Peissig et al. (2012) in NLP and text mining to detect cataracts, achieving high accuracy but facing challenges related to data quality, variations in doctors' writing styles, and computational demands [21][22]. In the field of Machine Learning for Heart Disease, Farooq & Sattar

(2015) applied machine learning algorithms for heart disease prediction, demonstrating effectiveness but highlighting data quality as a key limitation [23]. For eye diseases, statistical analysis by Al-Muhtaseb et al. (2021) used independent Component Analysis (ICA) and Pearson correlations for dry eye disease diagnosis, finding weak correlations among symptom [24]. The studies by Mohamed & ElFatatry (2015) [25], Ting et al. (2018) [26], and Zhao et al. (2020) [27] focus on the ability of CNN techniques to detect non-linear patterns in OCT images for early detection of retinal diseases like DME and AMD, emphasizing the importance of using large and balanced datasets. Meanwhile, studies by Burlina et al. (2017) [28] and Moccia et al. (2018) [29] focus on improving model accuracy through preprocessing techniques like macula and vessel segmentation, while studies by Prasad et al. (2019) [30] and Wang et al. (2021) [31] highlight the importance of high-resolution images and data augmentation techniques to overcome challenges like noise and computational time in clinical environments. In the field of pathology, multiple machine learning and deep learning techniques have been used to detect retinal diseases using optical coherence tomography and fundus photography. For example, Srinivasan et al. [32] applied machine learning with HOG feature extraction and SVM classification to detect AMD and DME. Their approach contributed to early automation in retinal disease detection and provided a publicly available OCT dataset. Similarly, Wang et al. [33] leveraged a public OCT dataset to classify images using LCP features and the SMO algorithm, which enhanced the integration of local and global image information. However, in the study SVM [32], SMO [33] the reliance on handcrafted features limited the adaptability, and SMO, while efficient for small datasets, faced scalability issues with larger datasets. Several CNN architectures have been evaluated for medical image classification, including VGG19, GoogLeNet, ResNet50, and DeNet. Studies, such as Juan et al. [34], have highlighted differences in performance based on dataset quality and preprocessing methods. Additionally, Cheng et al. [35] proposed a deep hashing algorithm based on ResNet50 to optimize feature extraction and image retrieval speed, though challenges remain in handling highly complex or noisy medical images. Aggarwal et al. [14] applied (BLIP) transformer to generate captions for chest X-ray images. Their approach achieved an accuracy of 92.31%, demonstrating the potential of pretrained transformers in improving diagnostic efficiency. We used the same model in our research but our model classifies diseases based on OCT images.

In comparison to the above related work, our study aims to apply BLIP technology in the medical field to classify retinal diseases using OCT images, where textual descriptions will be combined with images to improve classification accuracy using artificial intelligence, filling a significant gap. This achieves more efficient and accurate classification.

4. Methodology

The complexity and variability inherent in medical imaging data necessitate the development of advanced and automated artificial intelligence (AI) models to achieve precise accuracy in the detection, classification, and analysis of such data, including medical scans. These AI models are crucial for enhancing diagnostic tools and minimizing human error in medical practice. In fields such as medicine, where data availability may be limited, AI tools are particularly valuable for integrating multimodal information, such as combining medical images with textual data, to improve decision-making processes and provide comprehensive insights. Drawing inspiration from recent advancements in computer vision and natural language processing (NLP) tasks BLIP [20], we propose a novel vision-language pre-training (VLP) model, termed BLIP-Eye, designed to identify three retinal diseases using optical coherence tomography (OCT) scan images. Our approach builds upon the multimodal BLIP [20] framework, originally developed for image captioning, which we have adapted for the classification of eye diseases, specifically distinguishing between a normal retina and retinas affected by diabetic macular edema (DME), choroidal neovascular membrane (CNM), or age-related macular degeneration (AMD). The objective is to leverage BLIP's capabilities in adapting visual and textual representations while refining its architecture and training procedures to achieve accurate disease classification in ophthalmology.

This approach employs a combination of contrastive learning, pseudo-label bootstrapping, and multimodal alignment to effectively address the challenges associated with diagnosing eye diseases. These challenges encompass the processing of medical data, enhancement of data quality, design of robust classification algorithms, and development of precise diagnostic mechanisms. By integrating these advanced techniques, the approach aims to improve the accuracy and reliability of eye disease diagnosis, thereby contributing to more effective clinical decision-making and patient care.

The methodology begins with a vision encoder that processes high-resolution retinal images into hierarchical feature maps, effectively capturing both local details, such as fluid accumulation in the central macula, and global patterns, including vascular boundaries and color variations. Concurrently, a parallel text encoder processes textual classifications pertinent to retinal diseases, facilitating semantic alignment between visual and textual features. This multimodal alignment is crucial for linking optical coherence tomography (OCT) scans with their corresponding textual data in the classification of eye diseases. The multimodal fusion unit plays a pivotal role by calculating mutual attention between image and text symbols, concentrating on discriminative areas such as blood vessel boundaries, fluid shapes, and thickness variations related to the nerve or retina. This focused attention aids in accurately

differentiating between various categories of eye diseases.

This study uses a structured, multi-stage approach to adapt the Bootstrapping Language-Image Pre-training (BLIP) model [20] for diagnoses retinal diseases into four categories based on eye disease images. The process is divided into three key stages: data preprocessing, model selection and adaptation, and training and fine-tuning.

i. In the first stage, data preprocessing involves collecting a diverse set of retinal images representing the four target classes. To improve model generalization and prevent overfitting, standard data augmentation techniques such as random cropping, permutation, and normalization are applied. In addition, datasets balancing is performed to ensure equal representation across all classes, reducing the risk of class imbalance affecting model performance. Image label pairs are then formatted to conform to the BLIP framework, ensuring that textual descriptions complement class labels in a way that enhances multimodal learning.

ii. The second stage, model selection and adaptation, focuses on adapting the BLIP framework to image classification rather than its traditional application in image labeling. The vision encoder in the model, which is typically based on the vision transformer (ViT) architecture, is used to extract meaningful visual features from retinal images. Instead of generating textual descriptions, these features are mapped to pre-defined class embeddings corresponding to eye disease categories. To this end, the contrastive learning and multimodal alignment mechanisms within BLIP are tuned so that the model learns distinct visual representations directly associated with class labels rather than textual descriptions. Additionally, the text encoder, typically responsible for text-image alignment, is reused as a class prototype matcher, enabling efficient alignment between images and their respective classes. The classification head is then fine-tuned using Softmax activation and cross-entropy loss to enhance class separability.

iii. The final stage, training and fine-tuning, involves optimizing the adapted BLIP model through transfer learning and extensive hyperparameter tuning. Initially, the pre-trained BLIP model is fine-tuned on the two OCT datasets while keeping the early layers of the vision encoder frozen to preserve general feature extraction capabilities. A gradual unfreezing approach is used, with the deeper layers gradually fine-tuned to adapt to the unique visual patterns of the datasets. The training process uses the AdamW optimizer along with a carefully selected learning rate scheduler to balance convergence speed and stability.

Further improvements in classification performance are achieved through techniques such as label smoothing and focus loss, especially for dealing with visual similarities between classes. Model performance is rigorously evaluated using key evaluation metrics such as precision, accuracy, recall, and F1-score across all four classes, ensuring that the fine-tuned BLIP classifier delivers optimal classification accuracy.

5. Experimental Results

Evaluation Method: We trained our solution and the baseline (e.g., the CNN model [8]) on two publicly available datasets [36],[37] where the dataset from 2018 consists of 3,608 images and the largest dataset from 2021 consists 12,000 images evenly distributed across four classes. The datasets were split into 90% reserved for training and 10% for testing. Both models were optimized using the cross-entropy loss function and the Adam optimizer with a learning rate of 0.001. Training was conducted on Google Colab.

Results Analysis: As shown in (Table 1), the results clearly demonstrate that our solution, BLIP-Eye significantly outperforms the baseline, achieving an impressive accuracy of 91.94% in OCT Dataset (2018) [36], compared to only 54.17% for baseline, and 90.07% in OCT Dataset (2021) [37], compared to only 53.14% for baseline. Additionally, BLIP-Eye maintains consistently high precision and recall values (approximately 92%), demonstrating its superior classification performance and reliability.

Table 1. The accuracy comparison between our proposed method and baseline.

		BLIP-Eye	CNN
OCT Dataset (2018)	**Accuracy**	91.94	54.17
	Precision	92.38	53.29
	Recall	91.94	54.17
	F1	91.97	48.12
OCT Dataset (2021)	**Accuracy**	90.07	53.14
	Precision	90.66	59.53
	Recall	90.07	53.14
	F1	90.06	49.83

According to the evaluation results that shown in the Fig. 3, 4, 5 and 6, which show a comparison between the baseline model and our model, it is clear that the results of the CNN model showed in various metrics, whether Test Accuracy, Precision, Recall, or the F1 Score metric, which was at a rate of (48.12%), (49.83%) low accuracy. On the other hand, our proposed model achieved higher accuracy in various metrics F1 Score metric, which was (91.97 %), (90.06%). This indicates that large vision and language models are more accurate because they rely on Transformer techniques that are characterized by a

higher ability to understand complex patterns in images. Also, it is due to its ability to benefit from prior learning and analyze visual data and text together, which enhances its understanding of diseases.

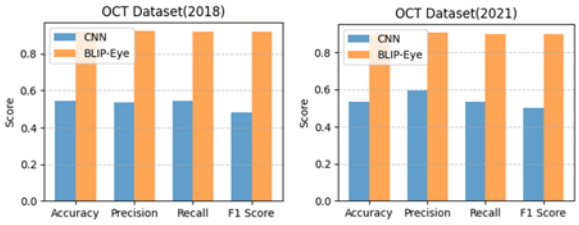


Fig. 3. Performance Comparison: BLIP-Eye vs.CNN [8].

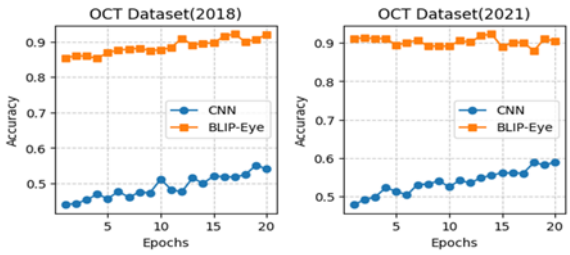


Fig. 4. Training Accuracy Progression.

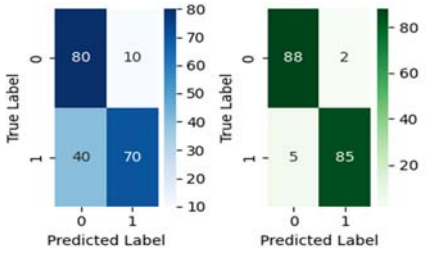


Fig. 5. CNN and BLIP-Eye Confusion Matrix in OCT Dataset (2018).

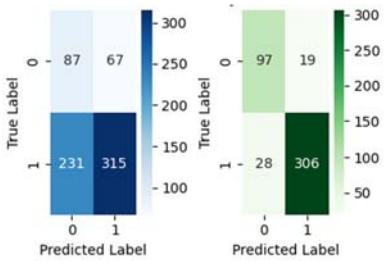


Fig. 6. CNN and BLIP-Eye Confusion Matrix in OCT Dataset (2021).

6. Conclusion

In this paper, we propose a novel VLP-based model, called BLIP-Eye, which combines rapid classification of retinal diseases with high accuracy, even with limited datasets. Our solution was trained on two real-world OCT scan-datasets. Experiment results demonstrates that BLIP-Eye outperforms the baseline in predicting various eye diseases, achieving a 43.88 % and 49.83% higher accuracy and maintaining balanced performance across all classes (i.e., normal retina, DME, NVM, and AMD) compared to the baseline. The results indicate that Vision-Language Pre-training based architecture, such as our model, enhances feature extraction and generalization, leading to better detection in comparison with CNN-based model. As future work, we will demonstrate our model performance on different types of retina disease and explore various of architectural optimizations for further improve model performance.

References

[1]. M. A. Hossain and M. S. A. Sajib, Classification of image using convolutional neural network (CNN), *Global Journal of Computer Science and Technology*, Vol. 19, No. 2, 2019, pp. 13-14.

[2]. Q. Li, W. Cai, X. Wang, Y. Zhou, D. D. Feng, and M. Chen, Medical image classification with convolutional neural network, in *Proceedings of the 13th IEEE International Conference on Control Automation Robotics & Vision (ICARCV)*, Dec. 2014, pp. 844-848.

[3]. J. Lu, L. Tan, and H. Jiang, Review on convolutional neural network (CNN) applied to plant leaf disease classification, *Agriculture*, Vol. 11, No. 8, 2021, p. 707.

[4]. S. S. Yadav and S. M. Jadhav, Deep convolutional neural network-based medical image classification for disease diagnosis, *Journal of Big Data*, Vol. 6, No. 1, 2019, pp. 1-18.

[5]. B. Kayalibay, G. Jensen, and P. van der Smagt, CNN-based segmentation of medical imaging data, *arXiv preprint arXiv:1701.03056*, 2017.

[6]. A. Kumar, J. Kim, D. Lyndon, M. Fulham, and D. Feng, An ensemble of fine-tuned convolutional neural networks for medical image classification, *IEEE Journal of Biomedical and Health Informatics*, Vol. 21, No. 1, 2016, pp. 31-40.

[7]. A. A. Reshi, F. Rustam, A. Mehmood, A. Alhossan, Z. Alrabiah, A. Ahmad, and G. S. Choi, An efficient CNN model for COVID - 19 disease detection based on X-ray image classification, *Complexity*, Vol. 2021, 2021, article ID 6621607.

[8]. M. Elkholy and M. A. Marzouk, Deep learning-based classification of eye diseases using convolutional neural network for OCT images, *Frontiers in Computer Science*, Vol. 5, 2024.

[9] Burlington Eye Docs, *Retinal Image Gallery*, 2024. [Online]. Available: https://www.burlingtoneyedocs.ca/retinal-image-gallery/

[10]. D. S. Kermany, M. Goldbaum, W. Cai, C. C. Valentim, H. Liang, S. L. Baxter, et al., Identifying medical diagnoses and treatable diseases by image-based deep learning, *Cell*, Vol. 172, No. 5, 2018, pp. 1122-1131.

[11]. J. Qiu, Z. Liu, X. Li, Z. Zhang, Y. Zhu, X. Chen, and J. Liu, VisionFM: A multi-modal multi-task vision foundation model for generalist ophthalmic artificial intelligence, *arXiv preprint arXiv:2310.04992*, 2023.

[12]. K. Zhang, F. Wu, G. Zhang, J. Liu, and M. Li, BVA-Transformer: Image-text multimodal classification and dialogue model architecture based on BLIP and visual attention mechanism, *Displays*, Vol. 83, 2024, art. ID 102710.

[13]. Y. Xiao, Y. Dou, and S. Yang, PointBLIP: Zero-training point cloud classification network based on BLIP-2 model, *Remote Sensing*, Vol. 16, No. 13, 2024, article ID 2453.

[14]. Aggarwal, A. K., Revealing AI-Driven Chest X-Ray Image Captioning Using Blip Transformer, 2023.

[15]. U. Naseem, S. Thapa, and A. Masood, Advancing accuracy in multimodal medical tasks through bootstrapped language-image pretraining (BioMedBLIP): Performance evaluation study, *JMIR Medical Informatics*, Vol. 12, No. 1, 2024, article ID e56627.

[16]. Z. Zhang, B. Wang, W. Liang, Y. Li, X. Guo, G. Wang, et al., SAM-guided enhanced fine-grained encoding with mixed semantic learning for medical image captioning, in *Proceedings of the* IEEE *International Conference on Acoustics, Speech and Signal Processing (ICASSP' 2024)*, Apr. 2024, pp. 1731-1735.

[17]. Q. Chen and Y. Hong, MedBLIP: Bootstrapping language-image pre-training from 3D medical images and texts, in *Proceedings of the Asian Conference on Computer Vision*, 2024, pp. 2404-2420.

[18]. W. Zhou, Z. Ye, Y. Yang, S. Wang, H. Huang, R. Wang, and D. Yang, Transferring pre-trained large language-image model for medical image captioning, in *CLEF (Working Notes)*, 2023, pp. 1776-1784.

[19]. M. Aono, T. Asakawa, K. Shimizu, and K. Nomura, Medical image captioning using CUI-based classification and feature similarity, in CLEF2024 Working Notes, CEUR Workshop Proceedings, *CEUR-WS.org*, Vol. 3740, 2024, paper 137.

[20]. J. Li, D. Li, C. Xiong, and S. Hoi, BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation, in *Proceedings of the International Conference on Machine Learning (ICML)*, PMLR, June 2022, pp. 12888-12900.

[21]. R. Waudby, S. Cohen, L. MacLeod, and J. P. M. Craven, Using natural language processing to detect cataract disease from electronic medical records, *Journal of Medical Informatics*, Vol. 58, No. 5, 2011, pp. 452-458.

[22]. P. Peissig, C. Friedman, V. Huser, G. W. Anderson, and K. M. Altman, Applying natural language processing techniques to electronic health records for detection of cataract disease and other ocular conditions, *Journal of Health Informatics*, Vol. 59, No. 3, 2012, pp. 303-312.

[23]. M. Farooq and S. Sattar, Data analysis techniques for early diagnosis of heart diseases using machine learning algorithms, *Journal of Medical Informatics*, Vol. 48, No. 2, 2015, pp. 105-112.

[24]. Z. Al-Mohtaseb, S. Schachter, B. Shen Lee, J. Garlich, and W. Trattler, The relationship between dry eye disease and digital screen use, *Clinical Ophthalmology*, Vol. 15, 2021, pp. 3811-3820.

[25]. A. Mohamed and S. ElFatatry, Deep learning for retinal disease detection using CNNs and OCT imaging, *Journal of Medical Imaging*, Vol. 22, No. 4, 2015, pp. 1123-1132.

[26]. D. S. W. Ting, J. J. Foo, and G. Y. Lim, Exploring CNN's ability to classify retinal images for early disease detection, *Journal of Ophthalmology*, Vol. 45, No. 7, 2018, pp. 734-741.

[27]. H. Zhao, T. Zhang, and X. Li, Using deep learning to detect non-linear patterns in retinal images for diagnosing diseases using OCT, *Medical Image Analysis*, Vol. 43, 2020, pp. 66-77.

[28]. P. Burlina, Y. Liu, and A. Sharma, Achieving high diagnostic accuracy in OCT retinal disease detection using pre-segmentation and CNNs, *Journal of Retina and Vitreous*, Vol. 39, No. 5, 2017, pp. 234-241.

[29]. S. Moccia, M. Ferro, and F. Santoro, Machine learning for vessel segmentation in retinal images to improve diagnostic accuracy, *IEEE Transactions on Medical Imaging*, Vol. 37, No. 2, 2018, pp. 140-148.

[30]. P. R. Prasad, R. Gupta, and S. Sharma, Deep neural networks for early detection of diabetic retinopathy and glaucoma, *Journal of AI in Healthcare*, Vol. 18, No. 3, 2019, pp. 98-105.

[31]. Y. Wang, Z. Li, and T. Chen, Improving retinal disease detection using high-resolution OCT and CNNs, *Ophthalmology Research Journal*, Vol. 44, No. 4, 2021, pp. 122-129.

[32]. P. P. Srinivasan, R. Kim, S. L. Mettu, J. A. Cousins, and S. Farsiu, Fully automated detection of diabetic macular edema and dry age-related macular degeneration from optical coherence tomography images, *Biomedical Optics Express*, Vol. 5, 2014, pp. 3568–3577.

[33]. Y. Wang, R. Kim, S. L. Mettu, and S. Farsiu, Machine learning-based detection of age-related macular degeneration (AMD) and diabetic macular edema (DME) from optical coherence tomography (OCT) images, *Biomedical Optics Express*, Vol. 7, 2016, pp. 4928–4940.

[34]. J. J. Gómez-Valverde, R. Pérez-Rovira, M. J. Ramírez, and A. F. De La Rosa, Automatic glaucoma classification using color fundus images based on convolutional neural networks and transfer learning, *Biomedical Optics Express*, Vol. 10, 2019, pp. 892–913.

[35]. S. Cheng, L. Wang, and A. Du, Histopathological image retrieval based on asymmetric residual hash and DNA coding, *IEEE Access*, Vol. 7, 2019, pp. 101388–101400.

[36]. D. Kermany, K. Zhang, and M. Goldbaum, Labeled Optical Coherence Tomography (OCT) and Chest X-Ray Images for Classification, *Mendeley Data*, V2, 2018.

[37]. O. S. Naren, *Retinal OCT Image Classification - C8 [Data set]*, Kaggle, 2021.

(018)

From Digital Disruption to AI Revolution: The Evolution of Healthcare Transformation

Shafaq Khan [1], Munir Majdalawieh [2], Dhruvi Kharadi [1], Tarun Verma[1], Tasnim Farhin [1] and Aditya Kumar [1]
[1] School of Computer Science, University of Windsor, Windsor, ON, Canada
E-mail: shafaq.khan@uwindsor.ca
[2] College of Technological Innovation, Zayed University, Dubai
E-mail: munir.majdalawieh@zu.ac.ae

Summary: Each technological bubble has uniquely influenced industries, producing varying levels of disruption and diverse outcomes that have reshaped economic and innovation landscapes. This report analyzes the impact of technological advancements on the U.S. healthcare industry across three major tech eras: the Dot-Com Boom, Web 2.0, and the AI-driven COVID-19 era. It explores how each phase drove innovation through mechanisms like Venture Capitalist (VC) funding, telemedicine, and improvements in health indicators such as life expectancy, while exposing vulnerabilities like speculative investments. Using machine learning models and time series forecasting, the report predicts no imminent tech bubble burst forthcoming, supported by trends such as steady GDP growth, increased VC funding, and rising healthcare expenditure. However, unemployment remains a critical risk factor to monitor. The findings emphasize the need for proactive strategies by policymakers, investors, and industry leaders to sustain innovation, mitigate potential risks, and promote equitable growth in healthcare.

Keywords: Dot-com boom, Web 2.0, Venture capitalist (VC) funding, Telemedicine, Time series forecasting, GDP growth, Healthcare expenditure, Innovation sustainability, Proactive strategies, Equitable growth.

1. Introduction

This report examines the transformative impact of technological advancements, often referred to as "tech bubbles," on the U.S. healthcare industry over the past few decades. It focuses on three major technological waves: the dot-com boom of the late 1990s, the emergence of Web 2.0 technologies in the early 2000s, and the AI-driven innovations accelerated during the COVID-19 era. Each phase brought groundbreaking changes to healthcare, such as the introduction of electronic health records (EHR), widespread adoption of telemedicine, and the integration of AI-powered diagnostics, significantly improving healthcare accessibility, efficiency, and outcomes. However, these advancements were not without challenges, as speculative investments and market volatility exposed vulnerabilities within the system. The report leverages economic indicators, including GDP growth, venture capital (VC) funding trends, and healthcare expenditures, alongside technological metrics such as the adoption rates of telemedicine, FDA drug approvals, and R&D spending, to provide a comprehensive analysis of the sector's evolution. Advanced machine learning models like logistic regression, random forests, and neural networks were employed to predict the likelihood of a tech bubble burst. This report underscores the need for policymakers, investors, and industry leaders to foster sustainable growth, address inherent risks, and ensure equitable benefits across the healthcare landscape in this rapidly evolving technological era.

2. Tech Bubble and Healthcare Industry

Healthcare has been impacted by tech bubbles, which are defined by speculatively high investments in technological areas [1]. Changes in technology have always had an impact on the healthcare sector; new avenues for growth and challenges are presented by every technological wave. This evaluation includes an analysis of the impact of tech bubbles on healthcare from version 1.0 to version 3.0, as well as a look at potential future effects from Tech Bubble 4.0

2.1. Tech Bubble 1: The Dot-Com Boom and Healthcare

The dot-com bubble of the late 1990s and early 2000s marked a pivotal moment in technological innovation, influencing multiple industries, including healthcare. Characterized by excessive investment in internet- based companies, the bubble saw healthcare organizations cautiously adopting digital tools like electronic health records (EHRs) and early telemedicine systems [2]. While healthcare lagged in reaping immediate benefits due to regulatory complexities and high capital requirements, the bubble set the stage for integrating technology into health services, despite significant financial disruptions caused by the crash [3].

2.2. Tech Bubble 2: Web 2.0 and the Healthcare Sector

During the mid-2000s, the rise of Web 2.0 technologies fueled the second tech bubble, marked by the proliferation of social media, cloud computing, and data-driven innovation. This period ushered in advanced analytics and digital platforms, enhancing patient engagement and operational efficiencies in healthcare [4]. For example, social media enabled real-time communication between patients and providers, while cloud- based solutions supported scalable healthcare operations. However, the bubble's eventual deflation underscored the challenges of over-reliance on speculative investments in technology-driven healthcare solutions [5].

2.3. Tech Bubble 3: The Digital Transformation in Healthcare

The COVID-19 pandemic triggered the third tech bubble, driven by investments in digital health technologies like telemedicine, AI, and wearables, reshaping healthcare delivery with a focus on accessibility and personalization [3]. However, the rapid influx of venture capital has raised concerns about overvaluation and sustainability, possibly leading to a burst like previous tech boom [6]. The rapid adoption of AI-driven diagnostics, robotic-assisted surgeries, and AI-powered personalized healthcare solutions has further strengthened digital transformation in healthcare [23][26]. These technologies not only enhance precision and efficiency but also support real-time patient monitoring and clinical decision-making.

3. Impact of Technological Advancements

3.1. Tech Bubble 1

During the 1990s dot-com boom, the U.S. healthcare industry faced economic and technological shifts. Healthcare spending as a share of GDP increased, driven by medical technology and service expansion [6]. However, employment growth remained stable, emphasizing infrastructure over workforce expansion [7]. Venture capital for healthcare startups was minimal, slowing innovation, while telemedicine languished due to weak infrastructure [8]. FDA approvals mainly focused on traditional pharmaceuticals, with tech-driven breakthroughs largely untapped. While life expectancy and health outcomes improved slightly, disparities remained, especially among low-income groups. These incremental gains set the stage for future healthcare innovations.

3.2. Tech Bubble 2

The 2000s ushered in significant changes for the U.S. healthcare industry, shaped by Web 2.0 and mobile technology. Healthcare spending as a share of GDP rose, driven by electronic medical records (EMRs) and tech solutions for data management, further fueled by an aging population and increased healthcare demand [11]. Although venture capital funding for healthcare startups rose, investments remained overshadowed by social media and mobile apps [6]. Telemedicine usage gradually expanded, facilitated by mobile devices and better internet connectivity [12], yet regulatory hurdles and reimbursement policies stifled broader adoption. The FDA approved more drugs, spurred by biotech companies leveraging computational biology [13]. Pharmaceutical research and development surged, focusing on biologics and advanced treatments. Despite the broader tech sector’s emphasis on social media and mobile platforms, healthcare benefited from incremental spillovers. Life expectancy continued to rise, although obesity and chronic diseases tempered the rate of improvement. In 2010, the Affordable Care Act (ACA) expanded insurance coverage, enhancing access to care [10], but inequalities persisted [14]. Overall, the industry’s growing integration with technology laid a solid foundation for further transformations in healthcare.

3.3. Tech Bubble 3

The 2010s marked a transformative shift in healthcare, driven by AI, biotechnology, and digital health. Healthcare spending increased as a share of GDP due to rising care costs and tech innovations [6]. Employment in the sector rose, driven by personalized medicine and telehealth growth [13]. Venture capital funding surged for AI diagnostics, biotech, and digital health platforms, signaling tech-healthcare convergence. Telemedicine grew with AI, better broadband, and user-friendly platforms, while innovations in data analytics and wearable devices offered new ways to personalize care. Regulatory changes and patient acceptance, particularly during COVID-19, solidified telehealth's role [12]. FDA approvals rose with expedited pathways and AI-driven discovery, and pharmaceutical R&D focused on precision medicine, gene therapies, and AI development, connecting healthcare to the broader tech evolution [15]. Despite these innovations, life expectancy growth slowed due to the opioid epidemic. However, the Human Development Index (HDI) improved, aided by digital health tools that expanded care access [16]. The 2010s underscored technology's potential in reshaping healthcare systems into more equitable, efficient, and patient-centered models for 8 seconds. While AI’s expansion in healthcare has significantly improved clinical outcomes, concerns regarding algorithmic bias and fairness persist. Studies highlight the necessity of ensuring diverse training datasets and ethical AI implementation to prevent disparities in healthcare delivery [25].

Now we will see the response to each indicators.

4. Economical Indicators

4.1 Healthcare Expenditure as a Percentage of GDP

This reflects how much the U.S. spends on healthcare relative to its total economy. As healthcare grows more dependent on technology, a burst could slow growth or decrease this percentage. Monitoring expenditure shifts during past tech bubbles can reveal potential financial stress on healthcare during Tech Bubble 3.0's burst. [17]

Fig. 1 illustrates the trends in total national health expenditures in the United States from 1920 to 2020, presented in both nominal terms (green line) and adjusted for inflation (blue line, constant 2022 dollars). The inflation-adjusted expenditures reveal a steady increase in healthcare spending over the decades, reflecting economic growth, rising healthcare costs, and advancements in medical technology. By examining these trends, it becomes evident that healthcare spending has been growing consistently, but significant shifts during periods of economic or technological disruption, such as tech bubbles, could alter this trajectory. This historical perspective underscores the importance of monitoring healthcare expenditure during Tech Bubble 3.0, as any significant shifts could indicate financial stress within the healthcare sector.

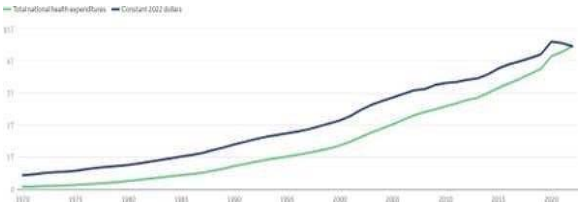


Fig. 1. Illustrates how GDP (in Billions) has increased over years.

4.2 Employment in Healthcare

Employment in healthcare includes jobs directly impacted by technological innovation, such as telemedicine or health data analysis [18]. A bubble burst might lead to job losses in tech- driven roles, making this a key indicator of economic stability within the sector.

Fig. 2 compares cumulative employment growth in the health sector (pink line) and non-health sectors (blue line) from January 1990 to February 2024.

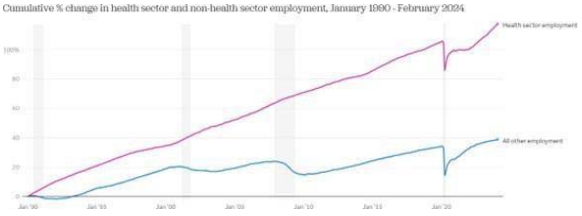


Fig. 2. Cumulative percentage change in employment in health sector [18].

The health sector shows consistent growth, significantly outpacing non-health sectors, even during economic downturns. This highlights the resilience and expanding importance of healthcare in the U.S. economy.

4.3. Venture Capitalist Funding for Healthcare Startups

VC funding boosts innovation by supporting startups developing new health technologies. It is highly sensitive to economic downturns. A Tech Bubble 3.0 burst may reduce investment, limiting advancements in areas like AI-driven diagnostics and health tech startups [19].

Fig. 3 depicts a fluctuating trend in venture funding and deal activity from 2014 to H1 2024. Total venture funding peaked in 2021, followed by a decline in 2022 and H1 2024. Similarly, the number of deals increased until 2021, then decreased in the subsequent period. The average deal size grew over the years, reaching its highest point in 2021. This suggests that while the overall number of deals declined, the size of individual deals increased. These fluctuations likely reflect broader economic trends and investor sentiment. The surge in activity until 2021 may have been driven by a strong economic climate and optimistic investor outlook. The subsequent decline could be attributed to factors such as rising interest rates, inflation, and geopolitical uncertainties. Despite the surge in venture funding for AI-driven healthcare startups, challenges such as regulatory compliance, institutional readiness, and economic constraints continue to shape AI's adoption trajectory [24].

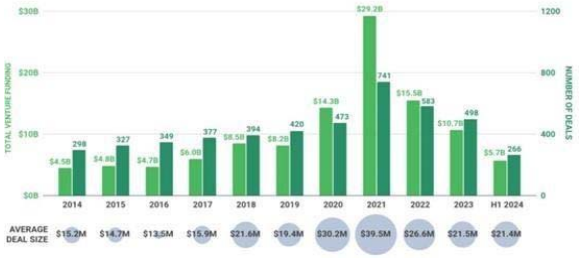


Fig. 3. Venture capitalist funding (in Billions) in health industry over the years.

5. Technological Indicators

5.1 Telemedicine Usage and Future

Telemedicine represents a key innovation that expanded significantly due to technological progress. A slowdown in tech investments could reduce access and advancements in telemedicine, limiting healthcare access in remote areas. [20]

Fig. 4 provides insights into the utilization of telehealth services. It shows that telehealth usage among both adults and children experienced a

significant surge in 2021, particularly during the early months of the year. While usage declined slightly in the later part of 2021 and continued to decrease in 2022, it remains higher than pre-pandemic levels. This trend is mirrored in the growth of the global telemedicine market, which is projected to expand significantly from 2022 to 2032, driven by increasing demand for remote healthcare services.

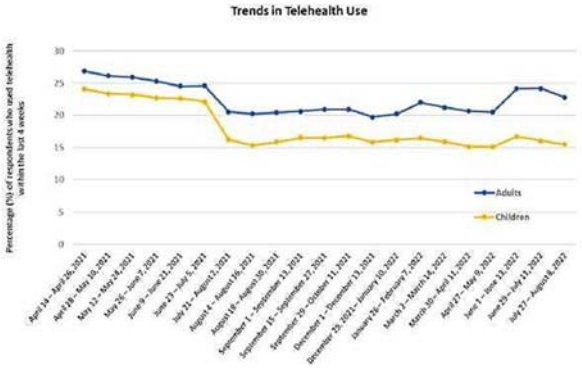


Fig. 4 (a). Trend of telemedicine usage from 2014.

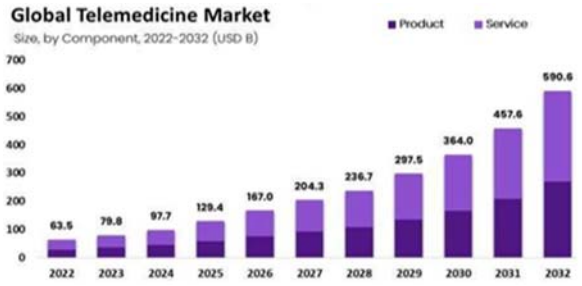


Fig. 4 (b). Future market value for telemedicine [20].

5.2. FDA Drug Approvals

The FDA has increasingly fast-tracked drug approvals with the aid of advanced tech. A tech bubble burst might lead to delays in drug approvals, hampering patient access to new treatments and reducing industry innovation. [21].

Fig. 5 illustrates the adoption of telehealth services, showing a significant surge in usage among both adults and children in 2021, particularly during the early months. While usage has declined since then, it remains higher than pre-pandemic levels. This trend is mirrored in the growth of the global telemedicine market, which is projected to expand significantly from 2022 to 2032, driven by increasing demand for remote healthcare services.

5.3. R&D Expenditure in the Pharmaceuticals Industry

R&D in pharmaceuticals relies on new technologies to develop drugs more efficiently. A reduction in R&D spending due to economic fallout from a bubble burst could slow down new drug discovery and development [22].

Fig. 6 illustrates the research and development (R&D) expenditure of the U.S. pharmaceutical industry from 1995 to 2023. It shows that R&D spending has increased significantly over this period, with the total pharmaceutical industry investing more than $112 billion in 2023. Notably, the contribution of PhRMA member companies to this expenditure has also grown steadily.

Fig. 5. Number of new drugs approved by FDA in USA over the years [21].

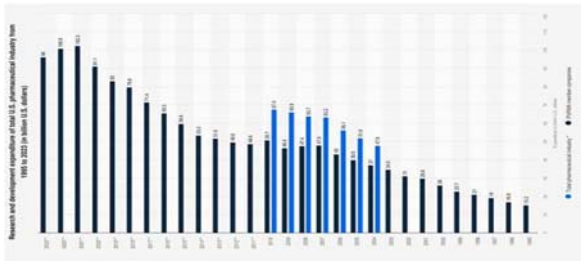


Fig. 6. R&D expenditure (in Billions) in pharmaceuticals industry in USA [22].

6. Healthcare Indicators

6.1 Life Expectancy Rate

Life expectancy is influenced by healthcare quality and access to advanced medical technologies. A bubble burst that impacts health tech innovation may result in stagnation or a decline in life expectancy, particularly if new treatments are delayed. [10].

Fig. 7 illustrates the research and development (R&D) expenditure of the U.S. pharmaceutical industry from 1995 to 2023. It shows a significant increase in R&D spending over this period, with the total pharmaceutical industry investing more than $112 billion in 2023. Notably, the contribution of PhRMA member companies to this expenditure has also grown steadily, representing a substantial portion of the total R&D investment. This trend reflects the industry's commitment to innovation and the development of new treatments and therapies.

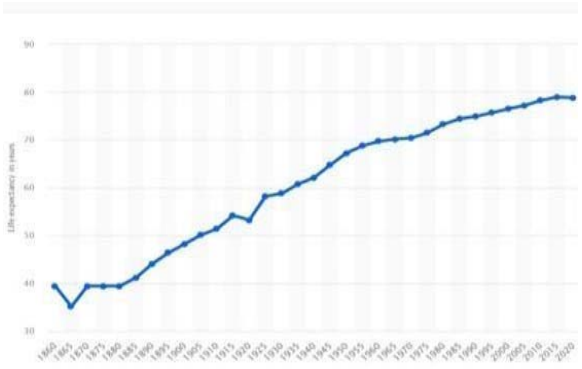


Fig. 7. Illustrates increase in Life Expectancy over years in the USA [10].

6.2. Access to Healthcare

Tech-driven improvements, such as telemedicine, have expanded healthcare access. Reduced tech funding could limit innovations, leading to disparities in healthcare access, especially in underserved communities. [14]

6.3. Human Development Index (HDI)

Fig. 8 illustrates the upward trend in life expectancy in the United States from 1840 to 2020. Life expectancy has increased significantly over this period, from around 38 years in 1840 to approximately 79 years in 2020. This notable increase reflects advancements in public health, medicine, and overall living standards.

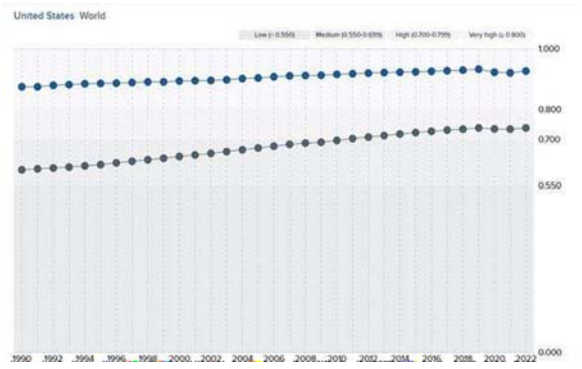


Fig. 8. Score of USA in HDI over the years.

7. Predicting Tech Bubble

The purpose here is to investigate the potential for a tech bubble burst in 2024. By utilizing a dataset containing various economic and health-related features, this analysis aims to identify key indicators that may signal a significant shift in the tech industry.

7.1 Dataset Understanding

The provided dataset appears to be a collection of economic and social indicators for the United States from 1990 to 2023. The data includes variables such as: Economic Indicators: GDP (Gross Domestic Product), GDP per capita, Venture Capital Funding, Venture Deal Count, R&D Expenditure, Unemployment Rate, Inflation Rate.

Health and Social Indicators: Infant Mortality Rate, National Health Expenditure, Human Development Index (HDI), Life Expectancy, Health Employment Count.

8. Models and Methodology

8.1. Exploratory Data Analysis (EDA)

This study began with an in-depth Exploratory Data Analysis (EDA) to understand the dataset's structure, identify key patterns, and explore relationships among features. First, numerical and categorical feature distributions were plotted to detect skewness and outliers, aiding in data quality checks and guiding feature selection. Temporal trends in indicators like GDP and unemployment were examined for long-term patterns or anomalies.

Next, a correlation matrix—visualized through a heatmap—revealed positive and negative correlations, crucial for identifying multicollinearity and important predictors. By focusing on variables with strong relationships to the target, this analysis streamlined model training and improved predictive performance.

Overall, this EDA ensured the data was well-understood, prepared, and optimized for subsequent machine learning tasks, reducing noise and enhancing model accuracy.

8.2. Correlation Analysis

To identify the most influential features for predicting the likelihood of a tech bubble burst, a correlation analysis according to Fig. 9, was conducted as part of the exploratory data analysis. This involved constructing a correlation matrix to measure the strength and direction of relationships between various features in the dataset. The results were visualized using a heatmap, which provided an intuitive way to highlight interdependencies and significant correlations.

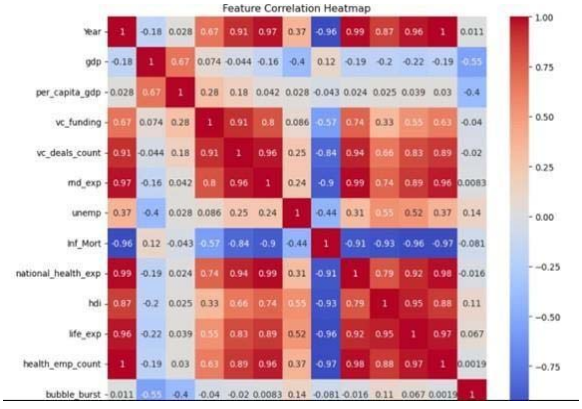


Fig. 9. Correlation between various features of the dataset.

8.3. Key Findings from Correlation Analysis

A heatmap (Fig. 9) identified several strongly correlated features with the target variable:

1) VC Funding: Exhibited a strong positive correlation, suggesting that higher venture capital influx might signal market saturation or speculative risks.
2) Unemployment Rate: Showed a negative correlation, reflecting how rising unemployment often precedes market downturns.
3) GDP: Displayed a positive correlation, indicating that rapid GDP growth could point to overheating markets.
4) VC Deals Count: Strong positive correlation, implying that increased deal flow may contribute to market exuberance.
5) National Health Expenditure: Unexpectedly positively correlated, hinting at broader economic or sectoral interconnections.

8.4. Feature Selection for Modeling

Based on correlation analysis, features with the strongest relationships to the target were chosen to streamline model training, reduce overfitting, and prioritize high-impact variables. Excluding weakly correlated features enhanced performance and interpretability, mitigating multicollinearity and avoiding redundancy. Overall, this analysis was instrumental in refining the dataset, ensuring meaningful predictive value, and aligning with economic theory and domain knowledge to bolster model accuracy and reliability.

8.5. Target Variables

The target variable for this study was designed as a binary classification to represent the likelihood of a tech bubble burst, where: 1: Indicates the presence or likelihood of a bubble burst; 0: Indicates no likelihood of a bubble burst.

1. Rationale for Target Variable Creation:

The development of the target variable was a critical step in aligning the study's objectives with its predictive modeling approach. This binary format was chosen for its simplicity and effectiveness in capturing the essence of the research question—whether the indicators analyzed could signal an imminent tech bubble burst.

2. Informed by Domain Knowledge:

The definition and thresholds for the target variable were established based on:

a) Economic Theories and Historical Trends: Previous occurrences of tech bubble bursts, such as the dot-com bubble of the early 2000s, were analyzed to understand key economic indicators and their thresholds. These insights provided a basis for mapping the predictors to the binary classification.

b) Correlation Analysis Findings: The strong relationships identified between features (e.g., VC funding, GDP, unemployment rate) and the target variable informed the decision to include these as key predictors in creating the target variable. This ensured that the binary classification reflected meaningful and data-driven relationships rather than arbitrary thresholds.

c) Creation Process: The binary target variable was generated by assessing a combination of statistical thresholds and trends derived from the data. Features such as a sharp increase in VC funding, abnormal GDP growth, or significant unemployment rate changes were used as flags to identify instances of potential instability. Cases that met predefined criteria suggesting market instability were classified as 1 (bubble burst), while all others were classified as 0 (no bubble burst)

8.6. Model Training

To accurately predict the likelihood of a tech bubble burst, various machine learning classification models were trained. These models were chosen to represent a diverse set of methodologies, including statistical, ensemble, and neural approaches. The selection ensured a comprehensive exploration of prediction strategies, each leveraging different strengths to address the problem effectively.

8.7. Selected Models and Their Roles

Here are the models that we will be using in this process below:

1) Logistic Regression: Serves as a baseline statistical model for binary classification, offering simplicity, interpretability, and insights into feature importance.
2) Random Forest Classifier: Builds multiple decision trees and averages their outputs, capturing non-linear relationships and providing robust predictions alongside feature importance insights.
3) XGBoost Classifier: A gradient boosting algorithm noted for high predictive accuracy, efficient loss function optimization, and resilience to missing data. It excels in handling correlated features.
4) K-Neighbors Classifier (KNN): Classifies data points based on similarity to neighboring points, offering a proximity-based perspective on decision boundaries.
5) Neural Network: Captures complex, non-linear interactions within the data, making it suitable for diverse economic indicators and providing cutting-edge pattern recognition.
6) Training Process: All models were trained on features identified through correlation analysis. Preprocessing steps (normalization, feature scaling) and careful splitting into training and testing sets ensured data integrity. Each model underwent hyperparameter tuning to optimize performance and mitigate overfitting.

This multi-model strategy provides varied methodological perspectives, enhancing the robustness and accuracy of the predictions regarding potential tech bubble bursts.

9. Results

9.1. Model Performance

The performance of each machine learning model was evaluated using accuracy scores, providing a comparative assessment of their predictive capabilities. The results are summarized as follows: Logistic Regression: 1.0 (100% accuracy); Random Forest Classifier: 0.85 (85% accuracy); XGBoost Classifier: 0.71 (71% accuracy); K-Neighbors Classifier: 0.86 (86% accuracy); Neural Network: 0.86 (86% accuracy).

All models consistently predicted that there would be no tech bubble burst forthcomming. Logistic Regression emerged as the most accurate model, achieving perfect classification; however, this may also indicate potential overfitting or an over-reliance on specific features. The ensemble and neural approaches provided robust results, reinforcing the consensus among the models.

9.2. Insights from Classification Models

The analysis based on the predictions of these models suggests that the economic indicators analyzed do not exhibit patterns associated with an imminent tech bubble burst in near future. The results highlight the stability of the current economic landscape while underscoring the importance of continuous monitoring of these variables.

9.3. Time Series Forecasting

A SARIMAX time series analysis (2024–2028) projected key economic and market indicators to offer insight into future trends:

1. Predicted Trends GDP: Continues rising, reflecting sustained economic growth.
2. VC Funding & Deals: Show steady increases, indicating robust tech investment.
3. National Health Expenditure: Grows over five years, signaling higher healthcare spending.
4. Unemployment Rate: Remains steady or slightly decreases, potentially stabilizing tech markets.

Results Interpretation: Overall, forecasts suggest positive economic momentum with minimal near-term risk of a tech bubble burst. However, monitoring unemployment remains essential, as employment fluctuations may impact market dynamics.

9.4. Combined Analysis

The combined results of the classification models and the time series forecasts provide a comprehensive assessment of the economic and tech market outlook. Both analytical approaches consistently indicate no imminent tech bubble burst forthcomming; increasing trends in GDP, VC funding, and deal count reinforce the perception of market stability and growth potential, and the unemployment rate emerges as a key indicator to monitor, as any adverse changes could impact the broader market.

10. Informing Stakeholders and Decision-Makers

The analysis predicts that a tech bubble burst is unlikely in near future.

Key indicators such as venture capital (VC) funding, gross domestic product (GDP), VC deal counts, and national health expenditure demonstrate steady growth trends. However, the unemployment rate, which remains stagnant or shows slight declines, is a critical variable to monitor. Historically, rising unemployment has correlated with economic instability. For instance, during the 2008 financial crisis, job losses eroded consumer confidence and triggered widespread market corrections. While the current trends do not signal an immediate risk, stakeholders must remain vigilant and prepared for potential shocks that could disrupt this balance.

10.1. Suggested Actions

1. For Policymakers

Policymakers play a pivotal role in safeguarding the economy against potential risks. To ensure equitable AI adoption, regulatory bodies must establish transparent AI governance frameworks that mitigate algorithmic bias, improve data diversity, and enforce accountability in healthcare AI models [25] Suggested measures include:

a) Job Creation and Workforce Resilience: Investing in programs that boost employment, particularly in sectors poised for growth like healthcare. For example, expanding funding for healthcare job training programs can help ensure that employment keeps pace with rising national health expenditures.

b) Stimulus Packages: During the COVID-19 pandemic, targeted stimulus packages stabilized the labor market. Policymakers could design similar interventions to mitigate risks in the future.

c) Digital Health Infrastructure: Investments in digital health, such as improving broadband access in underserved areas, can drive job creation and enhance healthcare accessibility, supporting overall economic stability.

2. For Investors

Investors should adopt diversified strategies to safeguard against market fluctuations:

a) Portfolio Diversification: While sectors such as healthcare and technology exhibit promising growth, diversification into stable yet emerging areas is essential. For instance, during the tech boom of the 2010s, investments in biotech and digital health startups surged. A similar focus on AI-driven healthcare platforms or sustainable technologies could yield high returns while mitigating risks.

b) Focus on Resilient Sectors: Shifting investment strategies toward industries with consistent growth

potential, like renewable energy and digital health, can help manage risks associated with overinvestment in high-growth sectors.

3. For Industry Leaders

Corporate strategies should align with stable economic trends to ensure long-term growth:

a) Expansion into Emerging Markets: Healthcare organizations could expand into telemedicine or personalized medicine, leveraging the increasing national health expenditure and VC funding. Training healthcare professionals in AI-assisted diagnostics and clinical decision-making will be crucial for seamless integration. Institutions should invest in AI literacy programs to enhance workforce readiness and optimize AI-powered healthcare delivery [24].

b) Monitoring Employment Trends: Companies must remain vigilant about workforce stability, as employment trends directly impact operational sustainability. During previous tech bubbles, companies that diversified into emerging but stable markets, such as cloud computing in the late 2000s, demonstrated resilience against broader economic challenges.

References

[1]. OECD Digital Economy Outlook 2015. Report, *Organisation for Economic Co-operation and Development*, 2015.

[2]. Mahajan, K., Pal, A., & Desai, A., Revolutionizing fan engagement: adopting trends and technologies in the vibrant Indian sports landscape, *International Journal of Management Thinking,* 1, 2, 2023, pp. 116–135.

[3]. Festa, G., Kolte, A., Carli, M. R., & Rossi, M., Envisioning the challenges of the pharmaceutical sector in the Indian health-care industry: a scenario analysis, *Journal of Business and Industrial Marketing*, 37, 8, 2021, pp. 1662–1674.

[4]. Here's how technology has changed the world since 2000, *World Economic Forum*, Nov 18, 2020. https://www.weforum.org/stories/2020/11/heres-how-technology-has-changed-and-changed-us-over-the-past-20-years/

[5]. Historical Trends in Healthcare Expenditure and Innovation: Explores healthcare's gradual adoption of technology and spending growth during the dot-com era, *BMJ*, 2024. Available at: https://www.bmj.com.

[6]. M. Young, The Technical Writer's Handbook. *University Science*, Mill Valley, CA, 1989.

[7]. Economic Impact of the 1990s Dot-Com Bubble: Analyzes the limited venture capital focus on healthcare during this tech-driven economic expansion, *The Balance*, 2024, Available at: https://www.thebalance.com

[8]. A Digital Prescription for Pharma Companies: Insights on pharmaceutical R&D and real-world evidence, *McKinsey & Company*, 2023, Available at: https://www.mckinsey.com

[9]. Franzoni, C., & Stephan, P., Uncertainty and risk-taking in science: Meaning, measurement and management in peer review of research proposals, *Research Policy,* 52, 3, 2022, 104706.

[10]. M'Gonigle, R. M., Jamieson, T. L., McAllister, M. K., & Peterman, R. M., Taking uncertainty seriously: From permissive regulation to preventative design in environmental decision making, *Osgoode Hall Law Journal*, 32, 1, 1994, pp. 99-169.

[11]. Healthcare and Life Sciences Outlook: Covers innovations such as EMRs and the slow integration of mobile tech in healthcare during the Web 2.0 era. *Deloitte*, 2024, https://www2.deloitte.com.

[12]. Bhugra, D., Smith, A., Ventriglio, A., Hermans, M. H., et al., World Psychiatric Association-Asian Journal of Psychiatry Commission on Psychiatric Education in the 21st century, *Asian Journal of Psychiatry*, 88, 2023, 103739.

[13]. Agarwal, R., Telehealth Adoption Accelerates Amid COVID-19 Pandemic, *Harvard Health Publishing*, 2021.

[14]. DeSalvo, K. B., Wang, Y. C., Harris, A., Auerbach, J., Koo, D., & O'Carroll, P., Public Health 3.0: A call to action for public health to meet the challenges of the 21st century. *Preventing Chronic Disease*, 14, 2017. 170017.

[15]. AI and Biotech Revolution in Healthcare: The 2010s Boom, Industry Report, *Khosla Ventures*, 2022.

[16]. Betran, A. P., Ye, J., Moller, A., Souza, J. P. and Zhang, J., Trends and projections of caesarean section rates: global and regional estimates. *BMJ Global Health*, 6, 6, 2021, e005671.

[17]. Coopersmith, J., The Electrification of Russia, 1880–1926, *Cornell University Press*, 1992.

[18]. Baade, R. A., & Matheson, V. A., Going for the gold: The economics of the Olympics, *The Journal of Economic Perspectives,* 30, 2, 2016, pp. 201–218.

[19]. Kehdinga George Fomunyam (Ed.), Theorizing STEM education in the 21st century, *IntechOpen*, 2020.

[20]. Telemedicine Statistics, *Market.us*, 2024. Available at: https://market.us

[21]. Wheeler, D. L., Database resources of the National Center for Biotechnology Information: update, *Nucleic Acids Research,* 32, 90001, 2003, 35D–40.

[22]. Meigs, L., Smirnova, L., Rovida, C., Leist, M., & Hartung, T., Animal testing and its alternatives – the most important omics is economics, *ALTEX*, 2018, pp. 275–305.

[23]. Bekbolatova, M., Mayer, J., Ong, C. W., & Toma, M., Transformative potential of AI in healthcare: definitions, applications, and navigating the ethical landscape and public perspectives, *Healthcare*, Vol. 12, No. 2, 2024, 125.

[24]. Roppelt, J. S., Kanbach, D. K., & Kraus, S., Artificial intelligence in healthcare institutions: A systematic literature review on influencing factors, *Technology in Society,* 76, 2024, 102443.

[25]. Ueda, D., Kakinuma, T., Fujita, S., Kamagata, K., et al., Fairness of artificial intelligence in healthcare: review and recommendations, *Japanese Journal of Radiology*, 42 1, 2024, pp. 3-15.

[26]. Li, Y. H., Li, Y. L., Wei, M. Y., & Li, G. Y., Innovation and challenges of artificial intelligence technology in personalized healthcare, *Scientific reports,* 14, 1, 2024, 18994.

(019)

Machine Learning-Based Sarcopenia Classification Using Multimodal Patient Data

A. Abdulnazar [1], J. Traub [2], N. Feldbacher [1,2], L. Celcer [1,2], V. Stadlbauer [1,2] and L. Herbsthofer [1]
[1] CBmed GmbH – Center for Biomarker Research in Medicine, Graz, Austria
[2] Division of Internal Medicine, Department of Gastroenterology and Hepatology, Medical University of Graz, Austria.
E-mail: akhilanaz.kuppasseryabdulnazar@cbmed.at

Summary: Machine learning (ML) offers a powerful approach for classifying sarcopenia, a condition marked by progressive muscle decline. This study applies ML models to predict sarcopenia as defined by the 2010 and 2019 European Working Group on Sarcopenia in Older People (EWGSOP) criteria, which shifted from emphasizing muscle mass (2010) to prioritizing muscle strength (2019). A dataset of 159 patients, refined from an initial 182, was analyzed after excluding 23 patients lacking annotations for both criteria. Five ML models—logistic regression (LR), support vector machines (SVM), k-nearest neighbors (k-NN), random forest (RF), and neural networks (NN)—were trained using demographic, clinical, and laboratory features. Balanced accuracy was used to evaluate classification performance across different data modality combinations. Results indicate that the 2019 criteria enable more accurate ML-based classification, particularly when incorporating clinical data. SVM and LR achieved the highest performance, with balanced accuracy exceeding 0.85 when integrating demographic and clinical features. These findings underscore ML's potential in sarcopenia diagnosis, though dataset limitations necessitate further patient recruitment to improve model generalizability.

Keywords: Sarcopenia, Machine learning, Diagnostic criteria, Muscle strength.

1. Introduction

Sarcopenia, a condition characterized by the progressive loss of muscle mass, strength, and function, is highly prevalent in older age and patients with chronic diseases. Using physical, laboratory, and clinical data, machine learning (ML) has significantly improved sarcopenia detection. Models based on physical characteristics and activity levels, such as LightGBM and deep neural networks, achieve over 80% accuracy, emphasizing BMI and grip strength as relevant features [1]. Approaches based on electronic health records (EHRs) integrate structured clinical data for early detection [2], while feature selection methods like CatBoost with LIME enhance interpretability [3]. Deep learning models further refine predictions, identifying key physiological markers [4]. Although imaging-based AI offers high accuracy, particularly in CT scans, non-imaging methods provide broader applicability, underscoring ML's role in advancing sarcopenia diagnosis and risk assessment.

Accurate and consistent diagnosis is essential for optimizing patient care and predicting health outcomes. The European Working Group on Sarcopenia in Older People (EWGSOP) guidelines, widely adopted for diagnosing sarcopenia, were revised in 2019 to prioritize muscle strength as the principal criterion, diverging from the 2010 framework, which emphasized muscle mass [5]. Our previous work demonstrated significant differences in the rate of sarcopenia diagnosis depending on which of the two criteria used, underscoring the need for further evaluation to determine the optimal diagnostic approach [5]. This study extends that analysis by leveraging ML models to classify sarcopenic status under both frameworks. Using multimodal data, we investigate the performance of ML models to predict sarcopenia of patients under both the 2010 and 2019 guidelines while exploring the relative contributions of different data modalities, including demographic, clinical, and laboratory features.

2. Methods

Dataset: The dataset was derived from a clinical cohort study from Medical University of Graz and initially included 182 patient records containing (a) demographic (age, sex, weight, height, body mass index (BMI), and disease information), (b) clinical (handgrip strength, gait speed, chair rise performance, mid-arm muscle circumference, and the Charlson Comorbidity Index (Charlson CI)), and (c) laboratory features (30 biochemical and hematological markers, including albumin, creatinine, glucose, and hemoglobin levels). Clinical features are used to classify patients as sarcopenic according to the 2010 and 2019 definitions and act only as positive controls for our study. Specific inclusion criteria were applied to ensure the quality and consistency of the analysis. Additionally, annotations were unavailable for some patients corresponding to 2010 and 2019 definitions, excluding 23 records. The final dataset comprised 159 patients. The study was approved by the research ethics committee of the Medical University of Graz (29-280 ex 16/17) and is registered at clinicaltrials.gov (NCT03080129). The study was conducted after

informed consent according to the principles of the Declaration of Helsinki. Further characteristics of the study cohort are published in [5, 6].

Tabular Data Filtering and Imputation: We employed imputation techniques [7] addressed missing values within the data. Specifically, a simple imputer using the mean strategy was applied to replace missing values of the respective features. This ensured that the dataset was complete and ready for model training. Subsequently, the dataset underwent 10 different random splits into training and test sets to evaluate the robustness of the models, and the average performance metrics from each test set were calculated. Model performance is reported as the average test set balanced accuracy of each of the 10 random splits.

Machine Learning Models: Five ML models were employed to evaluate the data: logistic regression (LR), support vector machine (SVM), k-nearest neighbors (k-NN), random forest (RF), and neural network (NN). These models were chosen due to their widespread use in classification tasks and their ability to handle varying data characteristics. Prior to model training, feature scaling via a standard scaler was applied to the data for all models, and missing values were imputed to minimize potential biases and leakage. To ensure robust evaluation across data splits, we implemented stratified k-fold cross-validation (N=5) and maintained a separate test set (20%) for unbiased final validation. Each model was trained and evaluated using stratified datasets to ensure fair comparisons and prevent biases introduced by class imbalance.

To address potential overfitting, we implemented several strategies within our methodology. Regularization was applied in models such as LR and SVM, where L2 regularization was controlled via hyperparameter tuning of the C parameter through GridSearchCV. In RF, constraints on tree depth, minimum samples per split, and the number of estimators were optimized to prevent excessive complexity. The k-NN model was tuned for the number of neighbors and weighting schemes to avoid high variance. The NN architecture incorporated dropout layers (50%) and 'Adam' optimization, which inherently applies L2 regularization.

Feature Importance Extraction: Feature importance was extracted using model-specific techniques. For tree-based models like RF, importance was derived from the *'feature importances'* attribute[1], reflecting the contribution of each feature to model decisions. For LR and SVM, feature importance was calculated from the absolute values of the model coefficients, indicating the strength of each feature's influence. For NN and k-NN, SHAP[2] analysis was used to determine feature impact by calculating SHAP values, which quantify the contribution of each feature to the predictions. All feature importance values were aggregated, sorted, and compared across models.

3. Results and Discussion

This study investigated the predictive performance of ML models for sarcopenia classification using various combinations of demographic, clinical, and laboratory data. Model performance was evaluated on two annotation guidelines (AG): 2010 and 2019 of EWGSOP, using mean balanced accuracy (MBA), upper and lower bounds of 95% confidence intervals (LCL, UCL), and top balanced accuracy (TBA) (see Tables 1-7). Results are summarized in Fig. 1.

Demographic Data Classification: Utilizing demographic features, the NN exhibited the highest MBA (0.771) for predicting sarcopenia using the 2010 criteria, followed by RF and LR. For the 2019 criteria, a notable decline in accuracy was observed across all models, with RF and k-NN achieving 0.538 MBA. This decrease may be attributed to the 2019 definition being less correlated with demographic features.

Table 1. Comparison of model performance for demographic data. Best results per column highlighted in bold.

Model	AG	LCL	MBA	UCL	TBA
NN	2010	**0.727**	**0.771**	**0.815**	**0.889**
	2019	0.479	0.497	0.515	0.551
RF	2010	0.687	0.738	0.789	0.845
	2019	0.498	0.538	0.608	0.625
LR	2010	0.679	0.737	0.794	0.867
	2019	0.475	0.507	0.538	0.571
k-NN	2010	0.683	0.723	0.763	0.824
	2019	0.484	0.538	0.592	0.614
SVM	2010	0.649	0.704	0.759	0.833
	2019	0.456	0.515	0.574	0.708

Clinical Data Classification: For the 2019 diagnostic criteria, LR and SVM demonstrated superior performance with an MBA of 0.768 and 0.761, respectively, while RF performed best for the 2010 criteria (0.666). This shift in performance is attributed to the EWGSOP criteria's change from emphasizing muscle mass (2010) to prioritizing muscle strength (2019), allowing for more clinically relevant assessments.

Table 2. Comparison of model performance for clinical data. The best results per column are highlighted in bold.

Model	AG	LCL	MBA	UCL	TBA
RF	2010	0.607	0.666	0.726	0.732
	2019	0.699	0.760	0.820	0.857
k-NN	2010	0.622	0.659	0.695	0.782
	2019	0.584	0.668	0.752	**0.966**
SVM	2010	0.617	0.653	0.689	0.746
	2019	0.683	0.761	**0.839**	0.944
LR	2010	0.596	0.636	0.677	0.756
	2019	**0.705**	**0.768**	0.830	0.944
NN	2010	0.584	0.609	0.633	0.645
	2019	0.653	0.706	0.758	0.792

[1] https://skforecast.org/0.9.0/user_guides/feature-importances

[2] https://shap.readthedocs.io/en/latest/

The increased accuracy in the 2019 cohort likely results from the greater reliance on clinical features like muscle strength, compared to the 2010 criteria, which focused on muscle mass and relied more on CT imaging. This clinical data set acts as our positive control, containing features directly used for the EWGSOP diagnosis of our samples.

Laboratory Data Classification: ML models performed worse on laboratory data than demographic and clinical datasets. SVM achieved the highest MBA (0.632) for the 2010 criteria, while LR (0.592) barely improved from random guessing for the 2019 criteria.

Table 3. Comparison of model performance for laboratory data. The best results per column are highlighted in bold.

Model	AG	LCL	MBA	UCL	TBA
SVM	2010	**0.585**	**0.632**	**0.679**	**0.777**
	2019	0.493	0.551	0.608	0.646
LR	2010	0.567	0.620	0.673	0.800
	2019	0.523	0.592	0.661	0.729
RF	2010	0.514	0.562	0.609	0.659
	2019	0.494	0.545	0.595	0.714
NN	2010	0.495	0.531	0.567	0.636
	2019	0.471	0.492	0.512	0.564
k-NN	2010	0.400	0.475	0.551	0.659
	2019	0.465	0.525	0.585	0.694

Demographic + Clinical Data: Combining these datasets boosted predictive performance. For the 2019 criteria, SVM achieved the highest MBA (0.978), followed by LR. Some individual models achieved 1.0 TBA, i.e. an error-free prediction on the test set, which is expected since this data set combination contains all information for the 2019 AG, whereas this data set again acts as a positive control for our experiment. For the 2010 criteria, k-NN demonstrated the highest performance (0.782).

Table 4. Comparison of model performance for a combination of demographic and clinical data. The best results per column are highlighted in bold.

Model	AG	LCL	MBA	UCL	TBA
k-NN	2010	0.760	0.782	0.803	0.825
	2019	0.538	0.588	0.637	0.746
NN	2010	0.734	0.763	0.792	0.833
	2019	0.595	0.674	0.753	0.854
RF	2010	0.691	0.730	0.770	0.833
	2019	0.793	0.842	0.891	0.929
LR	2010	0.664	0.716	0.769	0.833
	2019	0.915	0.959	**1.000**	**1.000**
SVM	2010	0.664	0.715	0.765	0.845
	2019	**0.945**	**0.978**	**1.000**	**1.000**

Demographic + Laboratory Data: The inclusion of demographic data in this model decreased predictive performance compared to other approaches. LR yielded moderate MBA values of 0.711 (2010 criteria) and 0.597 (2019 criteria).

Table 5. Comparison of model performance for a combination of demographic and laboratory data. The best results per column are highlighted in bold.

Model	AG	LCL	MBA	UCL	TBA
LR	**2010**	0.642	**0.711**	**0.780**	**0.825**
	2019	0.549	0.597	0.645	0.688
NN	2010	**0.647**	0.700	0.753	0.812
	2019	0.470	0.483	0.496	0.500
SVM	2010	0.641	0.699	0.756	0.802
	2019	0.508	0.555	0.603	0.643
RF	2010	0.624	0.674	0.724	0.800
	2019	0.516	0.539	0.562	0.575
k-NN	2010	0.545	0.611	0.678	0.755
	2019	0.477	0.519	0.562	0.604

Clinical + Laboratory Data: This combination improved performance, with LR (0.848) and SVM (0.824) leading for 2019 and SVM (0.665) performing best for 2010, reflecting diagnostic differences.

Table 6. Comparison of model performance for a combination of clinical and laboratory data. The best results per column are highlighted in bold.

Model	AG	LCL	MBA	UCL	TBA
SVM	2010	0.595	0.665	0.735	0.759
	2019	0.762	0.824	0.887	**0.966**
LR	2010	0.608	0.660	0.713	0.824
	2019	**0.803**	**0.848**	**0.893**	**0.966**
NN	2010	0.548	0.594	0.640	0.755
	2019	0.496	0.560	0.624	0.714
k-NN	2010	0.536	0.592	0.648	0.732
	2019	0.494	0.570	0.646	0.766
RF	2010	0.515	0.570	0.626	0.679
	2019	0.600	0.660	0.719	0.833

Demographic + Clinical + Laboratory Data: Integrating all three datasets yielded the highest predictive performance across all models. LR (0.867 in 2019 and 0.750 in 2010) consistently demonstrated superior performance in both cohorts, achieving the highest mean balanced accuracies. SVM (0.856) also exhibited excellent performance for the 2019 criteria. Future research may integrate imaging or other biomarkers to refine these models and expand their clinical use.

Table 7. Comparison of model performance for demographic, clinical, and laboratory data. The best results per column are highlighted in bold.

Model	AG	LCL	MBA	UCL	TBA
LR	2010	0.691	0.750	0.809	0.905
	2019	**0.814**	**0.867**	**0.920**	0.962
RF	2010	0.647	0.693	0.739	0.845
	2019	0.665	0.727	0.789	0.857
NN	2010	0.643	0.689	0.736	0.783
	2019	0.515	0.553	0.591	0.646
SVM	2010	0.615	0.675	0.734	0.802
	2019	0.794	0.856	0.919	**0.980**
k-NN	2010	0.570	0.642	0.714	0.818
	2019	0.495	0.577	0.660	0.786

Overall Performance: Our results indicate variability in model performance across different data combinations and annotation guidelines. LR and SVM consistently achieved strong and reliable performance, frequently attaining the highest or near-highest mean balanced accuracy. In contrast, NN and k-NN exhibited more variability, with some data combinations yielding high accuracy while others showed lower performance. Notably, clinical data, especially for the 2019 cohort criteria, significantly enhanced predictive performance (see Fig. 1 for detailed results), and the combination of demographic and clinical data resulted in the highest overall accuracy.

Feature Importance Extraction: The interpretability of ML models across different datasets reveals varied feature prioritization, with certain models emphasizing specific variables while others account for complex feature interactions. In the demographic dataset, LR and SVM consistently highlighted BMI and sex as crucial predictors of sarcopenia, with both models assigning high importance to BMI (SVM: 0.785, LR: 0.736). Conversely, RF emphasized continuous features like weight and age, reflecting its ability to model non-linear relationships. Height was deemed less critical across models, suggesting its secondary role in prediction. This is to be expected, as muscle mass and height are not firmly related, whereas age and weight are, thus further validating the plausibility of our findings. In contrast, the sarcopenia dataset focused on physical strength and mobility metrics, particularly in LR. RF, however, captured a broader set of features, including chair rises and the Charlson CI, highlighting its strength in identifying feature interactions. SVM struggled to identify significant predictors, assigning low importance to most features. Similarly, in the laboratory dataset, LR and SVM identified laboratory features thrombocytes, lymphocytes, and creatinine as key predictors. At the same time, RF results were more dispersed, emphasizing the importance of biochemical markers like gamma-glutamyl-transferase and glomerular filtration rate (according to the Chronic Kidney Disease Epidemiology Collaboration equation[3]).

These findings underscore the varying interpretability across models. Simpler algorithms like LR excel in linear relationships, while RF offers a more comprehensive view by modeling complex, non-linear dependencies.

Limitations and Ongoing Work: The primary limitations of this study are the small sample size (159 patients) and data imbalance. In 2010, there were 112 non-sarcopenic and 47 sarcopenic patients, while in 2019, the imbalance increased to 135 non-sarcopenic and 24 sarcopenic patients. This imbalance may bias model predictions, as the classifier might be skewed toward the majority class, affecting overall sensitivity. To mitigate this, we employed stratified cross-validation and balanced accuracy as the primary evaluation metric. However, the limited sample size restricts generalizability, and larger, more balanced datasets are needed for a more robust analysis. Additionally, this study lacks external validation on an independent dataset, which is crucial for assessing the model's real-world applicability. Future research should include external validation across diverse populations to ensure broader applicability and model reliability.

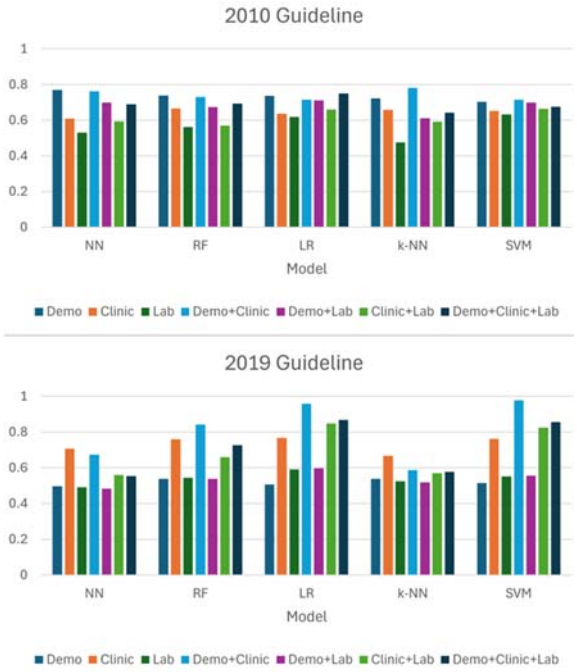


Fig. 1. Comparison of model performance (mean balanced accuracy) across different machine learning models (NN, RF, LR, k-NN, SVM) using 2010 and 2019 guideline data, categorized by feature set (Demo, Clinic, Lab, and combinations thereof).

4. Conclusions

By integrating demographic, clinical, and laboratory data, this study highlights the potential of ML algorithms for classifying sarcopenia based on EWGSOP criteria. The 2019 guidelines, emphasizing muscle strength, enabled superior model performance compared to the 2010 framework. Our findings underscore the importance of comprehensive features for improving model accuracy, with LR and SVM emerging as the most reliable models across data set combinations. Despite the limitations of a relatively small dataset, these results suggest ML algorithms could supplement sarcopenia diagnosis in clinical practice. Future research should focus on dataset expansion, external validation across diverse populations, and incorporating additional predictive features such as imaging and genetic markers.

[3]https://www.kidney.org/ckd-epi-creatinine-equation-2021

References

[1]. Kim J hee. Machine-learning classifier models for predicting sarcopenia in the elderly based on physical factors, *Geriatr Gerontol Int.,* 24, 6, 2024, pp. 595–602.

[2]. Luo X., Ding H., Broyles A., Warden S. J., Moorthi R. N., Imel E. A., Using machine learning to detect sarcopenia from electronic health records, *Digit Health*, 9, 2023, 20552076231197098.

[3]. Tukhtaev A, Turimov D, Kim J, Kim W., Feature Selection and Machine Learning Approaches for Detecting Sarcopenia Through Predictive Modeling, *Mathematics,* 13, 1, 2025, 98.

[4]. Bae J. H., Seo J. Won, Kim D.Y., Deep-learning model for predicting physical fitness in possible sarcopenia: analysis of the Korean physical fitness award from 2010 to 2023, *Front Public Health*, 11, 2023. https://www.frontiersin.org/journals/public-health/articles/10.3389/fpubh.2023.1241388/full

[5]. Traub J, Bergheim I, Eibisberger M, Stadlbauer V., Sarcopenia and Liver Cirrhosis—Comparison of the European Working Group on Sarcopenia Criteria 2010 and 2019, *Nutrients*, 12, 2, 2020, 547.

[6]. Aliwa B, Horvath A, Traub J, Feldbacher N, Habisch H, Fauler G, et al., Altered gut microbiome, bile acid composition and metabolome in sarcopenia in liver cirrhosis, *J. Cachexia Sarcopenia Muscle*, 14, 6, 2023, pp. 2676–2691.

[7]. Little R. J. A., Rubin D. B., Statistical Analysis with Missing Data, *John Wiley & Sons*, 2019, 462 p.

(020)

Digital Twins: Synthesizing Patient's 3D Anatomical Model from a CT Scan

V. Paneta [1], V. Eleftheriadis [1], G.C. Kagadis [2] and P. Papadimitroulas [1]
[1] BIOEMTECH, Kleisthenous 271, 15344, Athens, Greece
[2] 3dmi Research Group, Department of Medical Physics, University of Patras, 26504 Rion, Greece
Tel.: +30 2106548192
E-mail: vpaneta@bioemtech.com

Summary: Research on developing high performing AI models for medical imaging and oncology has recently been extended towards personalized care, among other fields. The present work aims to utilize state-of-the art AI models to provide tools for personalized medicine in the direction of digital twins. More specifically, a methodology for providing a 3D anatomical map for each patient undergoing a CT scan is developed and presented here, together with results for the respective proof-of-concept Generative AI task. An AI model developed using this methodology can be implemented as a software tool in hospitals to generate a digital representation of patients undergoing CT scans, integrating it into their medical records to support personalized medical protocols.

Keywords: CT, AI, Digital twin, Segmentation, Generative AI.

1. Introduction

Recent advancements in artificial intelligence (AI) for medical imaging and oncology have increasingly focused on personalized care [1]. In this context, the concept of synthetic patients—artificially generated patient models that mimic really anatomical and physiological characteristics—has gained traction as a means to enhance model training, validation, and simulation in clinical applications [2]. This study aims to leverage state-of-the-art AI models to enhance precision medicine towards personalization through the development of anatomical digital twins [3]. Specifically, we present a methodology for generating a three-dimensional anatomical map for individual patients undergoing computed tomography (CT) scans. The proposed approach, along with proof-of-concept results from a Generative AI model, is introduced and evaluated. This method has the potential to be implemented as a software tool in clinical settings, enabling the integration of digital patient representations into medical records to support personalized treatment protocols.

2. Materials and Methods

The developed methodology consists of two distinct steps, namely the raw data processing and the AI model development. Clinical CT data must be harmonized, segmented and prepared for model training, for generating an extended 3D anatomical map from partial CT images, as described in the following sections.

2.1. Image Processing

A set of clinical CT images was obtained from the University Hospital of Patras after following appropriate privacy measures. The raw dataset that was used in this study consisted of 591 anonymized CT images. Raw dicom images were converted to 3D nifti format and resized to 128 x 128 x 160 using linear interpolation. TotalSegmentator [4], a Deep Learning model for automatic CT segmentation based on nnU-Net [5], was applied on the resized 3D images and generated segmented images of 117 classes (organs), that were further grouped to 20 structures-organs of interest (heart, kidneys, spleen, bones, lungs, adrenals, pancreas, spleen, etc.). The body outline was additionally extracted from the resized images to be included as soft tissue in the 'rest of body' class. Fig. 1 presents an example of a raw and processed CT image at a central slice.

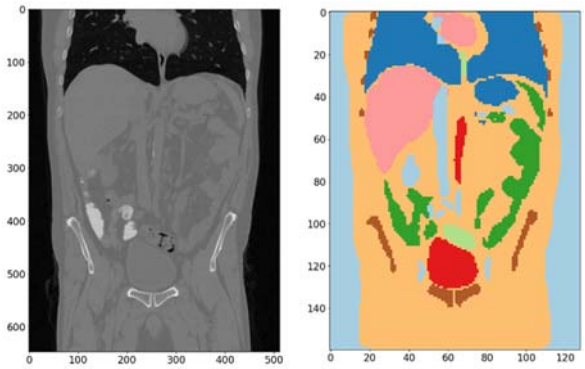


Fig. 1. Raw and processed CT image example as a central slice.

2.2. Generative AI Model

The processed dataset was used to train a vox2vox model [6], a conditional Generative Adversarial Network (cGAN) model for voxel-based Image

Translation tasks, to generate additional slices of a 3D anatomical map. The dataset was split to training (80%) and test (20%) datasets. For a proof-of-concept task, we trained a vox2vox model on partially cropped 3D segmented images (as input), using the corresponding whole 3D segmented images as ground truth (target). To this purpose a rolling window of 120 slices was applied on each segmented image with a step of 10 slices, resulting in 5 cropped input images with the same corresponding ground truth image, as seen in Fig. 2. This procedure resulted in the generation of a training set of 2360 images and 595 images for testing.

The default training losses of the vox2vox cGAN implementation, mean squared error (MSE) and generalized dice loss (GDL), were used for the training of the generator and discriminator. Performance evaluation of the vox2vox model in the present dataset was performed using three image quality metrics: (a) mean squared error (MSE), (b) peak signal-to-noise ratio (PSNR) and (c) structural similarity index measure (SSIM). The model was trained for 168 epochs on 4 nodes (4 GPUs per node) of the Meluxina HPC supercomputer [7] for about 10 hours.

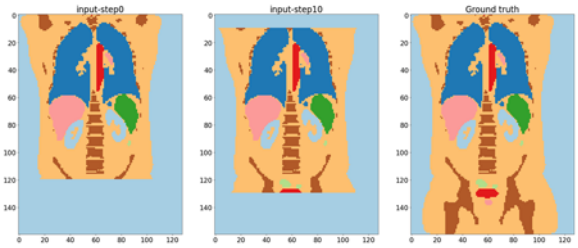


Fig. 2. Training dataset: cropped input images for the first 2 out of 5 steps with the corresponding ground truth image.

3. Results and Discussion

Fig. 3 presents the training progress of a vox2vox model over epochs in terms of training loss and SSIM. It is observed that although the network seems to be reaching some plateau state at early epochs, a continuous learning curve is indeed following. Training loss continues to get decreased values and correspondingly SSIM ones increase with more epochs of learning until the last epoch of training.

The continuous improvement was additionally evidenced through the visual inspection of results at several epochs and therefore we eventually selected the model after the last epoch (168) of training. An example of generated image is seen in Fig. 4 in a case where input CT did not include the biggest part of lungs. Generated lungs look very similar to the ones of ground truth image, although the smallest structures around heart are not well defined.

The used dataset in this study was strictly selected for torso imaging and average size of CT scans, while several organs and structures were at this stage grouped to simplify the task. A wider set of clinical CT scans (e.g. including head and neck CTs), coupled with further image processing tools need to be explored for better generalization of the model and to learn more complicated tasks.

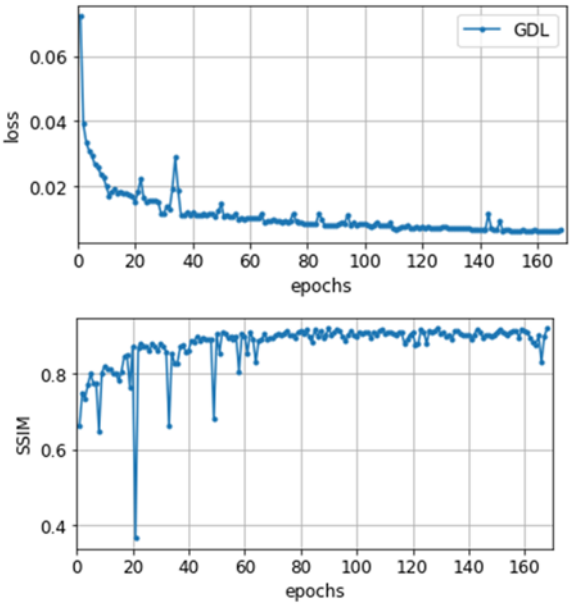


Fig. 3. Training loss and SSIM over epochs.

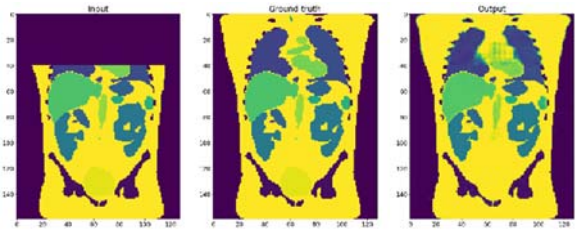


Fig. 4. Indicative model result predicting the upper torso area.

4. Conclusions

Clinical CT images were used to develop a methodology for creating AI-based synthetic patients, generating a personalized 3D anatomical map for each patient undergoing a CT scan. A trained AI model based on the developed methodology can provide a digital representation of these patients to be added to their medical record, when applied in hospitals as a SW tool. Moreover, the proposed methodology can be applied to other imaging modalities (such as MRI) and further contribute to the development of digital twins.

Further steps include the expansion of the dataset to include greater regions of the patients' CT scan and propose a clinical validation study.

Future work includes expanding the dataset to encompass larger regions of patients' CT scans and conducting a clinical validation study to assess the model's efficacy and reliability on actual patient outcomes.

Acknowledgements

We acknowledge EuroHPC Joint Undertaking for awarding us access to MeluXina at LuxProvide, Luxembourg and the support of EuroCC@Greece in this project.

References

[1]. K. B. Johnson, W. Q. Wei, et al., Precision Medicine, AI, and the Future of Personalized Health Care, *Clinical and Translational Science*, Vol. 14, Issue 1, 2021, pp. 86–93.

[2]. M. Giuffrè, D.L. Shung, Harnessing the power of synthetic data in healthcare: innovation, application, and privacy, *NPJ Digital Medicine*, Vol. 6, Issue 1, 2023, pp. 186.

[3]. K. Papachristou, P. F. Katsakiori, et al., Digital Twins' Advancements and Applications in Healthcare, Towards Precision Medicine, *Journal of Personalized Medicine*, Vol. 14*,* Issue 11, 2024, pp. 1101.

[4]. J. Wasserthal, H. C. Breit, et al., TotalSegmentator: Robust Segmentation of 104 Anatomic Structures in CT Images, *Radiology: Artificial Intelligence,* Vol. 5, Issue 5, 2023, e230024.

[5]. F. Isensee, P. F Jaeger, et al., nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation, *Nature Methods*, Vol. 18, Issue 2, 2021, pp. 203–211.

[6]. M. D. Cirillo, D. Abramian, el al., Vox2Vox: 3D GAN for Brain Tumour Segmentation, *arXiv:2003.13653v3,* 2020.

[7]. Meluxina HPC website, https://www.luxprovide.lu/meluxina/

(021)

A Platform for Machine Learning-Enhanced Decision Support in Molecular Tumor Boards

A. Abdulnazar, S. Stanzer and L. Herbsthofer
CBmed GmbH – Center for Biomarker Research in Medicine, Graz, Austria
E-mail: laurin.herbsthofer@cbmed.at

Summary: Making machine learning (ML) results accessible to end-users through visualizations is crucial for building trust and enabling their adoption in clinical practice. We present a platform designed to integrate ML results into molecular tumor boards, enhancing precision oncology decision-making. The platform supports the visualization and exploration of diverse data, including flow cytometry, metabolomics, and digital pathology. Key features include FlowSOM for automated gating, an ML model for evaluating sample quality, and ML-based tissue and cell segmentation for IHC images. These tools streamline the analysis of high-dimensional datasets, offering actionable insights to clinicians and researchers. Users can visualize spatial pathology data, interactively analyze metabolite ratios, and compare manually gated flow cytometry populations to those identified by the ML model. A particular focus on pre-analytical sample quality ensures reliability for downstream analyses. This platform provides a robust solution for navigating complex patient data by bridging ML-derived results with clinical usability. Future work involves usability evaluations and clinical impact assessments to enhance integration into routine oncology practice.

Keywords: Molecular tumor boards, Machine learning, Data visualization, Precision oncology.

1. Introduction

Precision oncology is about transforming cancer treatment based on the patient's unique features. Such personalization can be achieved only by integrating and interpreting diverse datasets involving genomic, metabolomic, proteomic, flow cytometry, and digital pathology data [1, 2]. These high-dimensional data are necessary for understanding biological mechanisms in cancer and the response of treatments in cancer patients, though their complexity often prevents effective clinical application [3].

Molecular tumor boards – a joint meeting of clinical experts – are tasked with using these diverse data types to make informed treatment decisions about patients but face significant hurdles in navigating, interpreting, and visualizing this complex information [4]. Several machine learning (ML) methods have emerged as transformative tools that can offer capabilities for data analysis, pattern recognition, and outcome predictions. However, their translation to clinical workflows has been limited because of usability challenges or the lack of accessible, user-centric platforms [5].

In a previous study [6], we introduced a data visualization tool that integrated clinical, molecular, and multi-omics data to support precision oncology decision-making. While it enabled intuitive navigation of complex datasets and personalized treatment strategies, further research was needed to assess its real-world impact. This study advances our earlier work by integrating machine learning (ML) to enhance automation, interpretability, and multi-omics data integration within molecular tumor boards and lays out a standardized framework for reproducing our work for other use cases. In particular, this work presents:

(i) **A template for the multi-omics database:** We detail how the database architecture behind our platform is designed to store all raw data and ML results to enable reproducibility of our work.

(ii) **ML-Driven Automated Analysis**: Automated flow cytometry gating using FlowSOM, sample quality assessment via metabolomic data, and tissue and cell segmentation of fluorescent multiplex immunohistochemistry (fm-IHC) images.

(iii) **Enhanced Interactive Visualization**: Dynamic exploration of ML-derived outputs, such as spatial pathology data and metabolite ratios to improve clinical interpretability.

The aim of this work is to provide a comprehensive and actionable framework for bringing ML results to the end-user to enable precision oncology decision-making.

2. Methods

In this study, we developed an interactive web platform that allows clinicians to analyze ML-based results for oncology patients. The platform integrates multimodal data, with a focus on multi-omics laboratory data, including digital pathology, flow cytometry, metabolomics, and sample quality assessments, all processed using ML methods. Data are securely stored in a structured database on on-premise servers, with access restricted to authorized users. The application allows users to select individual patients, compare them to the entire cohort, and view relevant clinical data. A screenshot of the platform is shown in Fig. 1.

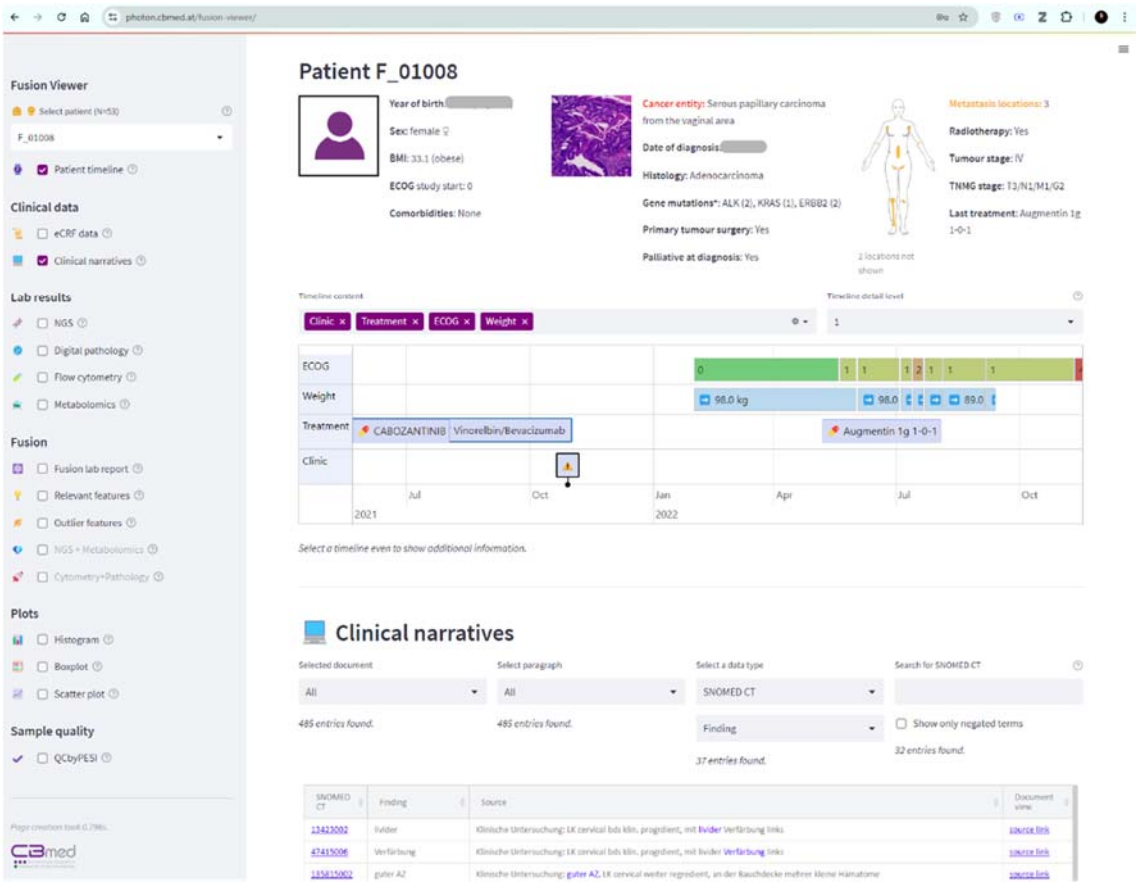


Fig. 1. Screenshot of the platform integrating multimodal data to support ML-based decision-making.

Multi-omics database template: The relational database behind our tool uses a modular *snowflake* schema (see Fig. 2). It consists of: (a) a central wide-format table (`tidyclin`) that holds basic demographic information of the patients and uses a `sample_ID` column as primary key that other tables connect to; (b) long-format tables for additional clinical data (e.g. comorbitities, treatments); (c) a module for each laboratory (`LAB`) consisting of three table templates for (i) wide-format data (`tidyLAB`), (ii) long-format data (`longLAB`), and (iii) general objects and files stored as binary objects in a long-format table (`byteLAB`); and (d) a hierarchical long-format table structure that decomposes electronic health records (EHR) into long-format tables starting from documents, paragraphs and identified narratives and predefined terms. Additionally, the central table is connected logically to a cloud-based laboratory information management system (LIMS) using the `sample_ID` as the primary key to link to additional data about the biological samples. This database template simplifies downstream database queries and visualization in the platform, and at the same time is flexible enough to incorporate new multi-omics lab data, regardless of their data structure.

Lab Module 1 (Flow Cytometry): FlowSOM facilitates the automated gating of flow cytometry data, significantly reducing manual effort and variability [6]. It clusters populations based on shared phenotypic characteristics using a self-organizing map (SOM) and generates heatmaps (see Fig. 3) and PCA plots for comparative analysis. In our platform, users can also combine manually gated populations with ML-driven clusters for deeper insights, enabling validation and refinement of results [7].

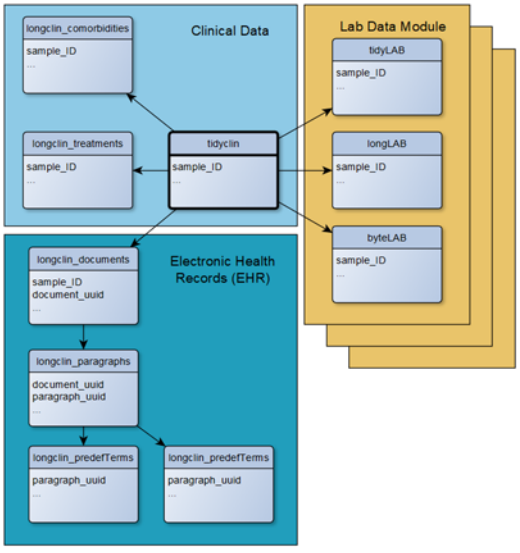


Fig. 2. Database schema of the platform consisting of clinical data tables (light blue), electronic health record tables (dark blue), and a module for each laboratory data set consisting of three tables per lab (yellow).

Lab Module 2 (Digital Pathology): This module allows users to explore spatial features obtained from H&E and fluorescent multiplex immunohistochemistry (fm-IHC) images, processed with ML models for tissue and cell segmentation. Analysis endpoints include phenotype percentages, tissue segment cell counts, cells within a specified radius around other cells, touching cells of two given phenotypes, and mean nearest neighbor distances (see Fig. 4). Users can inspect raw images with adjustable views, enhancing the interpretability of fm-IHC data [8].

Lab Module 3 (Metabolomics): This module maps metabolite features to standardized databases, providing associated compound IDs and visual representations. It includes interactive calculation of metabolite ratios, which is instrumental for biomarker identification. This supports clinical and research applications, enabling correlation analysis and hypothesis generation.

Lab Module 4 (Sample Quality Assessment): This module shows the biological sample quality as predicted by a random forest (RF) classifier trained on EDTA-plasma samples under various storage conditions and analyzed with probe electron ionization (PESI) [9]. With an accuracy of 88%, the RF model identifies optimal storage conditions and flags suboptimal samples, mitigating the risk of errors in downstream analyses.

3. Results and Discussion

The platform demonstrated the effective integration of ML-driven analyses with clinician-focused visualization tools. Key outcomes include:

Flow Cytometry: The implemented FlowSOM algorithm objectively identified cell populations, reducing subjective bias and variability from manual gating. Comparative heatmaps (Fig. 3) and PCA plots allowed clinicians to cross-validate ML-derived populations with manually gated results, enhancing confidence in the findings.

Digital Pathology: The interactive viewer of spatial analysis results enabled users to investigate tissue-level insights, such as cell distribution and neighborhood relationships, essential for understanding tumor microenvironments (Fig. 4). Access to raw images through the platform ensured transparency and facilitated data interpretation.

Metabolomics: The platform linked metabolomic data to standardized identifiers, supporting biomarker discovery. Users reported the ability to generate metabolite ratios (Fig. 5) as helpful in identifying clinically significant patterns.

Sample Quality Assessment: The platform automatically flagged suboptimal liquid sample quality, addressing a critical source of error in clinical laboratory workflows (Fig. 6). Its inclusion ensured that data used in ML models was reliable, improving downstream analyses.

These functionalities collectively streamline data interpretation of ML results obtained from multi-omics data, mitigate workflow inefficiencies, and enhance decision support in molecular tumor boards. The platform also enables cohort-level analyses for researchers, facilitating the identification of trends and correlations across patient groups.

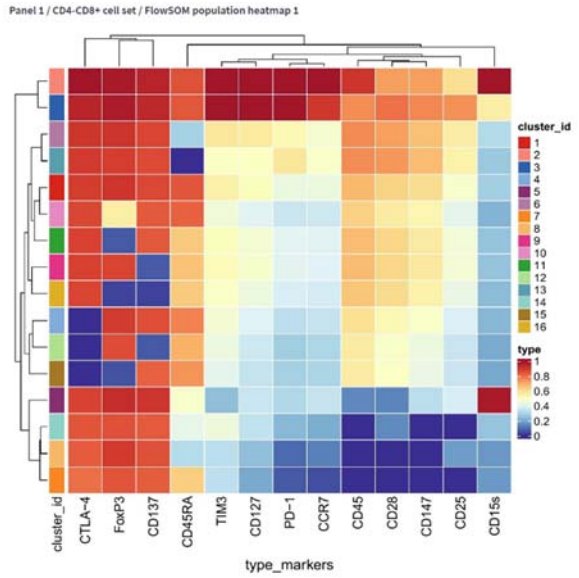


Fig. 3. Example FlowSOM heatmap of CD4-CD8+ T cell subsets for a selected patient using flow cytometry data.

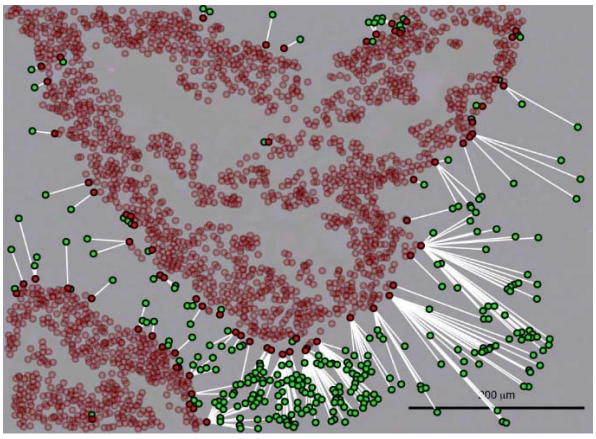


Fig. 4. Example spatial analysis of digital pathology image data using mutual nearest neighbor analysis of Tumor Cells and T Cells for a selected patient, highlighting cell-to-cell.

While this study presents a novel platform for the visualization of ML-driven results for molecular tumor boards, a detailed examination of the underlying ML models and extensive experimental validation remain limited. Further analysis of model performance, interpretability, and generalizability are needed to improve scientific rigor. Additionally, comprehensive clinical validation is required to assess real-world usability and impact of the platform.

Future work will focus on refining the ML algorithms, conducting extensive benchmarking, and evaluating the platform's clinical integration to strengthen its reliability and translational potential. User feedback will also be incorporated to refine interfaces and assess the platform's impact on clinical workflows.

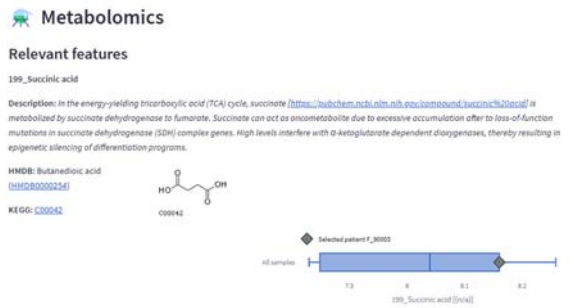


Fig. 5. Example box plot of a selected metabolite putting the value of a chosen patient into perspective of the remaining cohort.

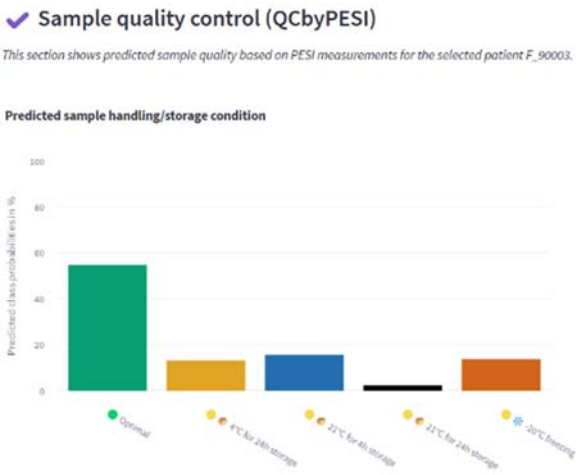


Fig. 6. Example sample quality control showing prediction of the sample quality by the random forest model based on PESI measurements for the selected patient.

4. Conclusions

This study presents an integrated platform for incorporating ML-driven results into molecular tumor boards, addressing critical data interpretation and usability challenges. By offering tools such as FlowSOM, ML-based sample quality prediction, and spatial analysis for fm-IHC images, the platform enhances data-driven decision support for oncology.
Future studies will evaluate the platform's scalability, usability, and clinical impact, ensuring it meets the diverse needs of oncology practitioners and researchers. This platform can potentially advance precision oncology practices by bridging the gap between ML analytics and clinical applications. An online demo of the platform is available upon request.

Acknowledgments

Ethics approval for this study was granted by the Ethics Committees of the Medical University of Graz (33-594 ex 20/21 and 32-194 ex 19/20), Salzburg (1052/2023), and Vorarlberg (EK-2-1/2023). Written informed consent was obtained from all participants prior to enrollment.

References

1]. Rodriguez H., Zenklusen J. C., Staudt L. M., Doroshow J. H., Lowy D. R., The Next Horizon in Precision Oncology – Proteogenomics to Inform Cancer Diagnosis and Treatment, *Cell*, 184, 7, 2021, pp. 1661–70.

[2]. Bhinder B., Gilvary C., Madhukar N. S., Elemento O., Artificial Intelligence in Cancer Research and Precision Medicine, *Cancer Discov*., 11, 4, 2021 April, pp. 900–915.

[3]. Krzyszczyk P., Acevedo A., Davidoff E. J., Timmins L. M., Marrero-Berrios I., Patel M., et al., The growing role of precision and personalized medicine for cancer treatment, *Technology,* 6, 3–4, 2018, pp. 79–100.

[4]. Bartels M., Chibaudel B., Dienstmann R., Lehtiö J., Piccolo A., Michielin O., et al., Evolving and Improving the Sustainability of Molecular Tumor Boards: The Value and Challenges, *Cancers*, 16, 16, 2024, pp. 2888.

[5]. Tan M. J. T., Schrock W. J., Maravilla N. M. A., Ting F. I. L., Choa-Go J. M., Francisco K. K., et al., The Data Scientist as a Mainstay of the Tumor Board: Global Implications and Opportunities for the Global South. *Front Digit Health,* 7, 2025, Available from: https://www.frontiersin.org/journals/digital-health/articles/10.3389/fdgth.2025.1535018/abstract

[6]. Abdulnazar A., Kreuzthaler M., Schulz S., Prietl B., Herbsthofer L., Tumor Board Visualization: Integrating Clinical and Laboratory Insights, in: Mantas J., Hasman A., Demiris G., Saranto K., Marschollek M., Arvanitis TN, et al., (Eds.), Studies in Health Technology and Informatics, *IOS Press*, 2024.

[7]. Van Gassen S., Callebaut B., Van Helden M. J., Lambrecht B. N., Demeester P., Dhaene T., et al., FlowSOM: Using self-organizing maps for visualization and interpretation of cytometry data. *Cytom Part J Int Soc Anal Cytol*., 87, 7, 2015, pp. 636–645.

[8]. Metsalu T., Vilo J., ClustVis: a web tool for visualizing clustering of multivariate data using Principal Component Analysis and heatmap, *Nucleic Acids Res*., 43, 2015, pp. W566–W570.

[9]. Van Herck Y., Antoranz A., Andhari M. D., Milli G., Bechter O., De Smet F., et al., Multiplexed Immunohistochemistry and Digital Pathology as the Foundation for Next-Generation Pathology in Melanoma: Methodological Comparison and Future Clinical Applications, *Front Oncol.,* 11, 2021, 636681.

[10]. Shi L., Habib A., Bi L., Hong H., Begum R., Wen L., Ambient Ionization Mass Spectrometry: Application and Prospective, *Crit Rev Anal Chem*., 54, 6, 2024, pp. 1584–633.

(022)

Towards AI-based Cognitive Training for Adult ADHD Patients

F. Boschello [1], A. Conca [1], I. Donadello [2], G. Giupponi [1], S. Holzer [1] and F. Zini [2]
[1] Psychiatric Service, Bozen-Bolzano Teaching Hospital
[2] Free University of Bozen-Bolzano, Faculty of Engineering
Tel.: + 39 0471 016236
E-mail: floriano.zini@unibz.it

Summary: Our purpose is to enhance the provision of mental health interventions to adults with Attention Deficit Hyperactivity Disorder (ADHD) by using Artificial Intelligence (AI), specifically Reinforcement Learning (RL) and Computational Persuasion (CP), to develop an adaptive and personalized Computerized Cognitive Training (CCT) tool. Our research has three main objectives:

- Assessing the effectiveness of RL-based cognitive training for adults with ADHD.
- Exploring advanced RL algorithms and developing CP techniques to improve personalization and adherence to CCT.
- Conducting a clinical trial to evaluate the effects of AI-based CCT on various cognitive functions.

This is the first project in Italy focusing on CCT for adult ADHD, an area that is largely under-researched globally. The research will be conducted by an interdisciplinary team formed by psychiatrists and psychologists from the Psychiatric Service, Bozen-Bolzano Hospital, and computer scientists of the Free University of Bozen-Bolzano.

Keywords: Computerized cognitive training, ADHD, Personalized medicine, Reinforcement learning, Computational persuasion.

1. Introduction

Attention Deficit Hyperactivity Disorder (ADHD) is a neurodevelopmental disorder often diagnosed in childhood, characterized by inattentiveness, hyperactivity and impulsivity [1]. Worldwide, the prevalence of ADHD is estimated to affect approximately 5% of children and adolescents [2]. Contrary to earlier understanding, several studies showed ADHD manifestations may remain not diagnosed during childhood and become noticeable only in early adulthood, when patients are required to self-manage a wider range of tasks [3]. Among adults the prevalence of ADHD, whether persistent from childhood or newly diagnosed, ranges between 2.5% and 6.7% globally [4]. ADHD imposes a significant public health burden, leading to educational and occupational challenges, unstable personal relationships and even criminal behavior [5]. Moreover, adults with ADHD have roughly nine times higher rates for comorbid psychiatric disorders such as anxiety disorders, major depressive disorders, bipolar disorder and substance use disorder, and twice as high the rate for type 2 diabetes and hypertension as the general population [6]. ADHD in adults is a well-researched condition with established diagnostic and treatment options. The best approach involves a combination of medication (stimulants or non-stimulants, according to the individual characteristics of each patient) and psychosocial interventions [7].

Computerized Cognitive Training (CCT) has already been included among the non-pharmacological interventions that can possibly alleviate the symptoms of ADHD [8]. The cognitive functions the literature suggests as characteristic for the disorder and thus potentially target of CCT include attention (sustained and selective), working memory, inhibition/interference (inhibitory control), set shifting, verbal fluency, speed of processing, and decision-making [9]. The evidence available in the literature about CCT applied to ADHD relates primarily to ADHD in childhood and adolescence [10] and much less to the treatment of ADHD in adults. A 2021 review based on meta-analyses and Randomized Controlled Trials about working memory CCT in ADHD adult patients concluded that CCT interventions provide additional benefits, above and beyond those of pharmacological interventions. The improved working memory performance could be associated, although not consistently, with secondary benefits such as symptom reduction and functional improvements [11]. It is noteworthy that multi domain cognitive training might be superior to single-domain cognitive training, as people with ADHD exhibit deficits in multiple cognitive domains and transfer measures, such as academic achievement, might be the result of a composite influence of different cognitive domains. Working memory remains at the center of the state-of-the-art intervention protocol for CCT, which often encompasses other areas of interventions (such as inhibitory control components) [3,12]. Overall, there is currently insufficient evidence to support the use of CCT as an additional treatment for ADHD and further research is needed to provide clearer recommendations for clinical practice.

In this paper, we present a well-founded research project that aims to exploit Artificial Intelligence (AI) techniques (i.e., Reinforcement Learning (RL) and

Computational Persuasion (CP)) to support an adaptive and personalized CCT intervention in adult ADHD patients. The ultimate goal is to provide an innovative and effective tool to enhance the current treatment approach of their condition.

The rest of the paper is organized as follows. In Section 2 we present our research objectives. In Sections 3 and 4 we elaborate on the activities we intend to carry out to achieve them. Finally, Section 5 highlights some security, confidentiality, and ethical issues of our research and outlines possible solutions.

2. Research Objectives

Adapting the difficulty of the computerized exercises to how individuals perform in their execution is crucial for the success of cognitive training programs [13,14], including those for ADHD. Existing systems for CCT generally embed mechanisms for the execution of sessions consisting in graded exercises, whose difficulty increases gradually according to a set of predefined rules (generally defined by neuropsychologists), considering the results the trainees achieve. For example, several of such systems can automatically increment the difficulty when the trainee performance overcomes a predefined threshold, but the order in which the parameters characterizing the exercise are varied is fixed and predefined, and thus their relative importance for the single trainee is neglected.

In our previous work [15], we contributed to creating truly adaptive mechanisms for cognitive training that personalize the training path for each individual. We embedded RL-based agents into the CCT MS-rehab system [16]. These agents interact with trainees during exercises, learning from their performance to adjust the difficulty level accordingly. The interaction among the RL-agent and the trainee is illustrated in Fig. 1, where the state includes the values of parameters defining an exercise.

We proposed a method for developing difficulty variation policies and tested this method using MS-rehab to conduct experiments with university students. The first experiment showed that a category policy for students effectively improves cognitive training outcomes and performs better than a standard baseline policy to vary the difficulty proposed by neuropsychologists. The second experiment showed that refining the category policy for single students further improved the training process by reducing the number of sessions required to achieve similar cognitive performance.

Based on these very promising results, we intend to extend our work to achieve three research objectives. The first goal is to strengthen the evidence of the effectiveness of RL-based personalization in CCT training. Therefore, we will replicate the experiments done in previous research with adult ADHD patients to check if RL-based CCT is effective and performs better than traditional CCT also for this category of people. A potential limitation of our approach is that the preliminary studies were not conducted on ADHD patients, which may impact the generalizability of the initial findings. However, we are confident to achieve significant results due to the established efficacy of CCT in patients with multiple sclerosis [15] and the similarities between those patients and ADHD patients when it comes down to their neuro-cognitive characteristics.

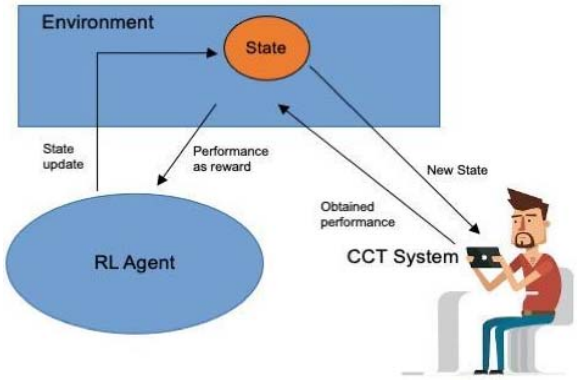


Fig. 1. The user trains with the CCT system and is provided with exercises tailored to his/her performance.

The second goal is to explore more advanced RL algorithms and develop Computational Persuasion (CP) techniques to further improve personalization and ensure adherence to CCT. Our hypothesis is that advanced RL algorithms provide added value for cognitive training of ADHD patients compared to the basic SARSA algorithm [17] that we currently use. Furthermore, as concerns remain about the risks of insufficient patient engagement in digital health and wellness interventions [18], CP may be an effective tool to promote patient adherence to them [19].

Our third and most important goal is to conduct a clinical trial involving adult ADHD patients to evaluate the effects of the personalized CCT system on multiple cognitive functions.

3. Advanced RL algorithms and CP techniques for CCT

This section outlines the advanced RL algorithms and CP techniques we want to explore, which will be integrated into a CCT system developed specifically for adult ADHD patients.

3.1. RL Algorithms

In our previous research [15], the algorithm used by the RL agent to learn the policy for exercise difficult variation was based on tabular SARSA [17]. SARSA is a well-known “traditional” on-policy method, which attempts to improve the policy that is used by the agent to make decisions about the action to take while the

agent is interacting with the environment. In our case, the reward from the environment depended on the performance obtained by the subject that executed the training exercise.

We intend to refine the used algorithm in three directions. The first is investigating other ways to calculate the reward given to the RL agents learning the adaptive difficulty-variation policies. In our previous research, we have adopted a simple model for calculation based on the performance obtained by the subject in the execution of the training exercise, which was a function of the correct, wrong, and missed answers of the subject. We want to investigate the use of other models, e.g.., the Rasch model [20], which has been used in education and healthcare to measure the performance of subjects in psychometric tests.

A drawback of our method is that policy learning needs to be started from scratch for each new category of trainees. Therefore, as a second direction, we want to explore if and how techniques like teacher advice and transfer learning [21] can be used in our context to ease the learning of new difficulty-variation policies from those already elicited using our approach.

The third direction is to use Deep Reinforcement Learning [22] to handle the high-dimensionality space that the Markov Decision Process that models cognitive training can have. In our previous research, we limited the actions to change the difficulty that were allowed in each configuration of the training exercises to limit the number of possible states and thus be able to use a simple algorithm such as tabular SARSA. However, in general, cognitive training exercises can be quite complex (e.g., think about a task-specific video game), and this limitation may preclude the discovery of effective policies to guide training.

3.2. CP Techniques

Maintaining long-term patient adherence to CCT is challenging. Computational Persuasive technologies (CP) combining AI and Cognitive Psychology offer a promising solution [19]. Our goal is to create and evaluate a novel CP algorithm for the CCT of ADHD patients to prevent low adherence. This will be the first application of CP for CCT.

The CP algorithm, implemented by an Automatic Persuasive System (APS), is triggered when a patient shows low engagement. Once the APS is triggered, a dialogue with the patient is started to understand the barriers and problems that prevent them from continuing the CCT. The patient will express their problems as arguments in natural language or with the use of menus through the graphical interface of the ADHD CCT system. The APS will answer with counter arguments that attack the ones expressed by the patient. Such arguments are tailored in a way that should maximize the effect of changing the beliefs of the patient. For example, the arguments of a patient with a low motivation for CCT can be attacked with arguments regarding (1) the possible risks of abandoning the training (e.g., loss of an important opportunity) or (2) the benefits of performing the training at home without having to go to the hospital. If the APS knows that the patient does not love travelling, then the second argument will be proposed as it is the most promising to be accepted. In such a dialogue, the APS and the patient exchange their arguments with the aim of making the patient change their beliefs regarding a particular problem/barrier during the CCT. This would help the patient in overcoming their barriers and continue with the training. Such an APS needs a Knowledge Base of arguments and an argument-selection policy to engage optimized dialogues with the patients. The Knowledge Base of arguments (also known as the argumentation graph) contains a collection of arguments regarding the problems and the barriers that can be encountered during the CCT. Each of these arguments is attacked by a set of counterarguments indicating possible solutions to such problems or different ways of addressing them. These counterarguments will be used by the APS during its turn in the dialogue.

The Knowledge Base of the APS will be built during the project with a strong collaboration between the Free University of Bozen-Bolzano and the Psychiatric Service of the Bozen-Bolzano Hospital. This would result in a questionnaire to be administered to the patients to gather their possible barriers and possible solutions. In addition, literature on Behavior Change [23] will be investigated to formulate further possible counterarguments to the barriers. The Knowledge Base will be implemented as an ontology in the standard Resource Description Framework (RDF) format.

The argument-selection policy is a decision theory algorithm [23-25] that is responsible for selecting the best counter arguments that the APS will propose in the dialogue to maximize their impact in overcoming the patients' barriers. Many policies have been defined in the literature, and they leverage the utility of each argument in a game-theory-based setting. The utility of a single argument is a numeric value that represents how much the argument is useful for one of the involved parties (i.e., the user or the APS) with respect to a persuasive goal. In our example, the argument about reluctance to travel will have a higher utility than the one about missing opportunities. Arguments are selected to maximize the final dialogue utility for each involved party. The effectiveness of the CP algorithms will be assessed by measuring the engagement of the patients with the system. In particular, we will assess whether the CP methods can improve engagement if patients can get reluctant towards the system.

4. Clinical Trial

We plan to conduct a clinical study to measure the impact of a CCT intervention implemented with AI on adult ADHD patients in the following domains: executive functions performance (attention and working memory), daily functioning, and quality of

life. Through CCT sessions administered in combination with a prior established pharmacological therapy, we expect that adult ADHD patients will experience an improved attention and working memory performance (primary outcome) and show improvement in dealing with daily activities and quality of life (secondary outcomes).

We target to recruit 60 adult patients with ADHD who are receiving treatment at the ADHD Outpatient Clinic of the Psychiatric Service, Bozen-Bolzano Hospital. The inclusion criteria will consist of age equal or over 18 years, diagnosis of ADHD (any subtype) according to the DSM -5 (confirmed by means of a semi-structured diagnostic interview (DIVA-5)). The exclusion criteria will include a known diagnosis of mental retardation or dementia and/or an active abuse of alcohol or substances.

After collecting an informed consent, including an extensive explanation of the scientific basis of the project and of data privacy guarantees, the enrolled subjects will be anonymized. A Wechsler Adult Intelligence Scale (WAIS) - IV will be administered to them at baseline to measure the overall cognitive performance and the various subtests scores. Among these subtests scores, the Working Memory Index (WMI) and the Processing Speed Index (PSI) are deemed to be accurate, reliable and valid measures of executive functions [26]. The enrolled subjects will then have access to the AI-based CCT system embedding RL and CP developed in our research project specifically for adult ADHD patients. The subjects will be asked to interactively perform cognitive training tasks with video-game-like layouts related to the attention and memory cognitive areas. An anonymization algorithm will prevent experimenters from tracing the identity of individual subjects.

For methodological reasons, 10 out of 60 recruited subjects will be leveraged for learning the preliminary RL-based policy for the category of adult ADHD patients and once the training phase of the algorithm is completed, will no longer be subjected to the exercises in the CCT package. The remaining 50 ADHD adult patients will be arranged as follows. Thirty of them will be asked to perform two CCT sessions per week for six weeks and during their training sessions they will be challenged with the previously developed RL policy for the category of ADHD patients. In addition, they will interact during sessions with the APS in case they are reluctant to use CCT.

At the same time a comparison group consisting of the twenty remaining patients, who did not perform the AI-based CCT intervention, will be asked to perform two CCT sessions per week for six weeks with a baseline difficulty variation mechanism proposed by experts in cognitive training. They will also not be exposed to the APS.

Within a month after the last administered CCT session, the cognitive performance of all the 50 patients will be measured throughout the WAIS-IV. The primary outcome to be assessed will be the attention and working memory performance (measured by means of the Working Memory Index (WMI) and the Processing Speed Index (PSI)). Moreover, patients will be asked to assess their perceived self-efficacy in the daily routine and their perceived quality of life by means of Visual Analogue Scales (VAS).

The WAIS IV is one of the most widely used neuropsychological assessments in clinical practice, which is of routine use even to characterize the cognitive functioning of neurodevelopmental disorders such as ADHD. Although a cognitive assessment through the WAIS-IV test is not helpful for diagnosing ADHD, it allows obtaining a distinctive cognitive profile of performance (derived specifically from the working memory and the processing speed domains).

We expect to show that a CCT intervention based on RL and CP algorithms used as an add-on to the pharmacological therapy already in place, may improve the short-term executive functions' performance, daily functioning and quality of life of adult ADHD patients. Moreover, our expectation is that the 30 patients exposed to the AI-based adaptive and personalized CCT will outperform those in the comparison group in the dimensions mentioned above.

5. Conclusions

Our research aims to demonstrate that combining AI-driven CCT sessions with the standard pharmacological and psychoeducational approach can significantly enhance the executive functions' performance of adults with ADHD.

Important issues we need to address are the confidentiality and security of patients' data and the potential ethical concerns of AI-based CP. As regards data confidentiality and security, we will pay close attention to the implementation of measures to ensure security and confidentiality, such as that patient data are kept confidential and stored with their consent securely on a protected server, that a secure communication is guaranteed, that access to data is quickly restored when necessary, and that sensitive data are anonymized and encrypted. Regarding CP techniques, we emphasize that CP does not constitute coercion or manipulations, as it does not involve the use of force or imposition on patients and does not employ deception to change patients' beliefs. Therefore, a CTT system that includes CP can be trusted by end-users and ensures ethical compliance.

In conclusion, we would like to emphasize that the interdisciplinary and coordinated approach between the Psychiatric Service, Bozen-Bolzano Hospital and the Free University of Bozen-Bolzano envisaged by our research will enable us to realize an effective CCT tool for personalized medicine that will ensure an overall improvement in patients' burden of symptoms and quality of life. We expect results that demonstrate the potential of combining innovative digital interventions with traditional therapies to achieve comprehensive treatment outcomes.

References

[1]. Song P, Zha M, Yang Q, Zhang Y, Li X, Rudan I. The prevalence of adult attention-deficit hyperactivity disorder: A global systematic review and meta-analysis, *J Glob Health*., 11, 2021, 04009.

[2]. Polanczyk G. V., Willcutt E. G., Salum G. A., Kieling C., Rohde L. A., ADHD prevalence estimates across three decades: an updated systematic review and meta-regression analysis, *Int J Epidemiol*., 43, 2, 2014, pp. 434-442.

[3]. Brown T. E., ADD/ADHD and Impaired Executive Function in Clinical Practice, *Curr Psychiatry Rep*., 10, 5, 2008, pp. 407-411.

[4]. Song P, Zha M, Yang Q, Zhang Y, Li X, Rudan I., The prevalence of adult attention-deficit hyperactivity disorder: A global systematic review and meta-analysis, *J Glob Health*., 11, 2021, 04009.

[5]. Shaw M, Hodgkins P, Caci H, et al., A systematic review and analysis of long-term outcomes in attention deficit hyperactivity disorder: effects of treatment and non-treatment, *BMC Med*., 10, 2012, 99.

[6]. Chen Q, Hartman C. A, Haavik J, et al., Common psychiatric and metabolic comorbidity of adult attention-deficit/hyperactivity disorder: A population-based cross-sectional study, *PLoS One*., 13, 9, 2018, e0204516.

[7]. Kooij J. J. S., Bijlenga D., Salerno L., et al., Updated European Consensus Statement on diagnosis and treatment of adult ADHD, *Eur Psychiatry*, 56, 2019, pp. 14-34.

[8]. Somma F., Rega A., Gigliotta O., Artificial Intelligence-powered cognitive training applications for children with attention deficit hyperactivity disorder: A brief review, in *Proceedings of the Symposium on Psychology-Based Technologies*, 2019.

[9]. Evans S. W., Beauchaine T. P., Chronis-Tuscano A., et al., The Efficacy of Cognitive Videogame Training for ADHD and What FDA Clearance Means for Clinicians, *Evidence-Based Practice in Child and Adolescent Mental Health*, 6, 1, 2021, pp. 116-130.

[10]. Wiest G. M., Rosales K. P., Looney L., Wong E. H., Wiest D. J., Utilizing Cognitive Training to Improve Working Memory, Attention, and Impulsivity in School-Aged Children with ADHD and SLD, *Brain Sci*., 12, 2, 2022, pp. 141.

[11]. Al-Saad M. S. H., Al-Jabri B., Almarzouki A. F., A Review of Working Memory Training in the Management of Attention Deficit Hyperactivity Disorder, *Front Behav Neurosci*., 15, 2021, 686873.

[12]. Dentz A., Guay M. C., Parent V., Romo L., Working Memory Training for Adults With ADHD, *J Atten Disord*., 24, 6, 2020, pp. 918-927.

[13]. Pedullà L., Brichetto G., Tacchino A., et al., Adaptive vs. non-adaptive cognitive training by means of a personalized App: a randomized trial in people with multiple sclerosis, *J Neuroeng Rehabil*., 13, 1, 2016, p. 88.

[14]. Peretz C., Korczyn A. D., Shatil E., Aharonson V., Birnboim S., Giladi N., Computer-based, personalized cognitive training versus classical computer games: a randomized double-blind prospective trial of cognitive stimulation, *Neuroepidemiology*, 36, 2, 2011, pp. 91-99.

[15]. Zini F., Le Piane F., Gaspari M., Adaptive Cognitive Training with Reinforcement Learning, *ACM Transactions on Interactive Intelligent Systems*, 12, 1, 2022, pp. 1-29.

[16]. Gaspari M, Zini F, Stecchi S. Enhancing cognitive rehabilitation in multiple sclerosis with a disease-specific tool, *Disabil Rehabil Assist Technol*., 18, 3, 2023, pp. 313-326.

[17]. Sutton R. S., Barto A. G., Reinforcement Learning: An Introduction, Second Edition, *The MIT Press*, 2018.

[18]. Karekla M., Kasinopoulos O., Neto D. D., et al., Best practices and recommendations for digital interventions to improve engagement and adherence in chronic illness sufferers, *European Psychologist,* 24, 1, 2019, pp. 49-67.

[19]. Hunter A., Towards a framework for computational persuasion with applications in behaviour change, *Argument & Computation*., 9, 1, 2018, pp. 15-40.

[20]. Christensen K. B., Kreiner S., Mesbah M. (Eds.), Rasch Models in Health. *Wiley-ISTE*, 2012.

[21]. Silva F. L. D., Costa A. H. R., A Survey on Transfer Learning for Multiagent Reinforcement Learning Systems, *Journal of Artificial Intelligence Research*, 64, 2019, pp. 645-703.

[22]. François-Lavet V, Henderson P, Islam R, Bellemare M. G., Pineau J., An Introduction to Deep Reinforcement Learning, *Foundations and Trends® in Machine Learning*, 11, 3-4, 2018, pp. 219-354.

[23]. Etminani K., Tao Engström A., Göransson C., Sant'Anna A., Nowaczyk S., How Behavior Change Strategies are used to Design Digital Interventions to Improve Medication Adherence and Blood Pressure Among Patients with Hypertension: Systematic Review, *J Med Internet Res*., 22, 4, 2020, e17201.

[24]. Donadello I., Hunter A., Teso S., Dragoni M., Machine Learning for Utility Prediction in Argument-Based Computational Persuasion, in *Proceedings of the AAAI Conference on Artificial Intelligence*, 36, 5, 2022, pp. 5592-5599.

[25]. Donadello I., Lirio R., Hunter A., Dragoni M., Compromises in Dialogical Argumentation: Aggregated Policies for Biparty Decision Theory, *Frontiers in Artificial Intelligence and Applications*, Vol. 392: ECAI 2024, pp. 3437 – 3444.

[26]. Wilson A. C., Cognitive Profile in Autism and ADHD: A Meta-Analysis of Performance on the WAIS-IV and WISC-V, *Arch Clin Neuropsychol*., 39, 4, 2024, pp. 498-515.

(023)

ProKIP: Rationale and Development of the AI-Nursing-Care-Readiness-Assessment (AINCRA)

K. Seibert [1], D. Domhoff [1], J. Altona [1], S. Jäger [2], F. Bießmann [2], A. Nowak [3], R. Gubser [3], D. Fürstenau [3,8], J. Pohle [4], L. Bergmann [5], D. Walter [6], K. Beier [7] and K. Wolf-Ostermann [1]

[1] Universität Bremen, Institut für Public Health und Pflegeforschung, Grazer Str. 4, 28359 Bremen, Germany
[2] Berliner Hochschule für Technik, Fachbereich VI Informatik und Medien, Luxemburger Str. 10, 13353 Berlin, Germany
[3] Charité Universitätmedizin, Institut für Medizinische Informatik, Invalidenstraße 90, 10115 Berlin, Germany
[4] Alexander von Humboldt Institut für Internet und Gesellschaft, Französische Straße 9, 10117 Berlin, Germany
[5] Verband für Digitalisierung in der Sozialwirtschaft e.V., Leipziger Str. 70/71, 06108 Halle (Saale), Germany
[6] Universität Hildesheim, Institut für Betriebswirtschaft und Wirtschaftsinformatik, Universitätsplatz 1, 31141 Hildesheim, Germany
[7] Universität Bremen, Insitut für Philosophie, Enrique-Schmidt-Str. 7, 28359 Bremen, Germany
Tel.: + 49 421 21868903
[8] Freie Universität Berlin, Fachbereich Wirtschaftwissenschaft, Garystr. 21, 14195 Berlin, Germany
Tel.: + 49 421 21868903
E-mail: kseibert@uni-bremen.de

Summary: AI integration in nursing care often faces challenges due to unaddressed requirements and known pitfalls. Readiness models provide a structured approach to assess project preparedness and enhance capabilities for successful AI adoption. Therefore, we developed the AI-Nursing-Care-Readiness-Assessment (AINCRA), designed for planning, implementing, and evaluating AI projects in nursing. Using a bottom-up approach for the development of maturity models, we utilized a mixed-methods to identify key AI readiness factors through expert workshops (N=21), interviews (N=14), an online survey (N=53), and nominal group consensus (N=5). These factors were validated through literature reviews and a Think Aloud study with experts (N=18) from various disciplines. Our findings resulted in a comprehensive framework encompassing five dimensions: regulatory, processual, technical, social and ethical factors, and community building. Future work will build up on these results creating an assessment tool with different maturity levels to help practitioners evaluate and optimize AI projects throughout their lifecycle, fostering successful AI adoption in nursing care.

Keywords: Artificial intelligence, Nursing, Readiness.

1. Introduction

Scientists and practitioners alike are turning to Artificial Intelligence (AI) both to find a possible solution to the worsening shortage of skilled workers in the care and health professions and to exploit the potential of basic AI functions to ensure and promote high-quality healthcare [1-3].

Following the definition given by the European Commission, AI encompasses systems that exhibit intelligent behavior by assessing their surroundings and making decisions – often independently – to accomplish specific objectives [4]. These AI systems can exist purely as software, operating in virtual environments (such as voice assistants, image recognition tools, search engines, and speech or facial recognition systems) [4]. Alternatively, AI can be integrated into physical devices, including advanced robots, self-driving cars, drones, and Internet of Things (IoT) applications [4].

While AI's potential to support nurses in acute and long-term care has been recognized on a global scale, its adoption at the point-of-care remains limited and use cases outside hospitals have not been explored extensively [1, 2]. Implementation challenges are diverse and go beyond technical and regulatory requirements; social, ethical, and practical factors must also be addressed [5, 6].

Starting in 2022, the German Ministry for Education and Research is funding projects aimed at supporting the integration of AI into nursing practice and promoting the use of nursing and healthcare data for AI systems [7]. These projects are faced with the fact that prior AI-in-Nursing-Care (AINC) projects have been reported to often fail to adequately address requirements and known pitfalls [5], which can lead to delays or even failure of a project.

One way out of this dilemma is making use of AI readiness models. These models provide a systematic approach to assess how a project is set up to achieve its' objectives and offer guidance on how to improve capabilities to foster the integration of AI-based technologies in an organization or at the point-of-care [8, 9]. Although a variety of online resources are available for assessing organization-specific AI readiness or organizational AI adoption capabilities,

only a few tools have been specifically developed for the healthcare sector (e.g. [9]). So far, an evidence-based comprehensive AI-Nursing-Care-Readiness-Assessment (AINCRA) is missing.

Given the promising potential of AI for nursing care, we sought to support those responsible for funding, planning, conducting and evaluating AINC-projects in addressing barriers and facilitators, and in reflecting on their approach to key aspects and success factors of these projects, ensuring a sustainable development and implementation of AI in nursing practice. To do so, we present an overview on the development of a comprehensive AINCRA and highlight dimensions and selected attributes crucial for AINC-project success.

2. Research Background and Related Work

2.1. AI in Nursing Care and Challenges for Research and Development

AI is increasingly being used to support nurses, patients, and their relatives – both in direct care and in the optimization of organizational and workflow processes in nursing [1, 2]. In Germany, products such as hybrid speech assistants for nursing documentation [10], ambient sensors for fall detection or apps for fall-risk assessment [11, 12], intelligent shift and route planning for ambulatory care, AI-supported clinical decision support systems, and (virtual) tools or robots for social participation and engagement of nursing home residents [13, 14] are available on the market. Furthermore, generative AI tools are entering the nursing space, e.g., for nursing education, clinical practice, and research [15].

As the European Federation of Nurses Associations advocates for investing in digital technologies and responsible AI innovation to help frontline nurses free up more time for direct patient care, especially during a period of significant shortages in the healthcare workforce [16], and calls for the participation of nurses in co-design of AI tools [17], the professional discourse has moved beyond the question of whether the nursing profession wants and needs AI systems, focusing instead on how to achieve demand-driven and beneficial AI research and development for nursing practice.

Research programs aim to harness AI's potential to support nurses and care recipients while addressing one of society's major challenges: ensuring high-quality nursing care despite the growing imbalance between the rising demand for care in aging societies and the availability of healthcare and nursing professionals in the labor market [18].

In prior work, we explored needs, application scenarios, requirements, facilitators and barriers for AINC projects and mapped characteristics of successful AINC projects including regulatory, processual, technological, ethical and legal aspects and supportive ecosystems [5]. As with all aspects of digitalization, AI applications often face implementation challenges, ranging from structural barriers such as lack of connectivity to high-capacity infrastructure networks or low availability of machine-readable data, to poor organizational development and staff training as well as lacking financing models [19].

2.2. AI Maturity and Readiness Models

Theoretical and empirical evidence on readiness and maturity models as well as on AI adoption is available [8, 9, 20-23], with few models or frameworks focusing on healthcare settings.

The terms readiness model or maturity model are often used simultaneously to describe a structured mechanism to assess "the current effectiveness of capabilities and ongoing progress in a particular domain" [8]. In the context of organizations, respective models facilitate internal or external benchmarking, highlight potential future improvements and provide guidelines for the evolutionary process of organizational development [8, 21], with stage-growth models being the most common type of model used in Information Systems [21]. Readiness models consist of a series of maturity levels for a class of objects, representing an anticipated, desired, or typical development path for these objects in discrete stages [20]. Objects typically refer to organizations or processes and the lowest level represents the initial state, where only limited capabilities are present in the given domain, while the highest level represents a concept of full maturity [20].

For example, Pumplun et al. developed a maturity model for machine learning (ML) systems in clinics that comprises 3 dimensions and 12 attributes and operationalizes those attributes by 5 corresponding levels [9]. Dimensions include the organization, the adopter system, and patient data, while attributes target, a.o., ML strategy, technical infrastructure, ML methods expertise, ML acceptance or the availability of patient data and corresponding aspects such as unified data formats and quality standards [9].

Factors that influence the adoption and implementation of AI-based technologies in German hospitals have been described by Weinert et al. who identified a lack of resources (staff, knowledge, and financial) as the most important barriers and highlight that few tools have been implemented in routine care, and that many hospitals do not use or plan to use AI in the future [24].

In their discussion of organizational readiness, Akbarighatar et al. [8] point out that organizations tend to prioritize technical and business assessments, neglecting the importance of considering the ethical, societal, and human impacts of AI. These factors, however, have been deemed crucial for AINC projects by experts from nursing science, information and data science, nursing practice, ethics, and nursing education in our prior work [5]. For the development of the AINCRA, we build upon this previous work and expand it with additional methods.

3. Design and Development

We followed a general bottom-up approach for the development of maturity models, starting from a set of specific characteristics or factors categorized into capabilities, from which maturity levels are defined [8]. We further considered general requirements and recommendations for the development of maturity models [20, 21] and took common criticisms of existing maturity models [21] into account when deciding on our methodological approach.

We used a mixed-methods approach, allowing for iterations. Firstly, we identified and categorized AI readiness factors and attributes from the literature sample of a rapid review (N=292) [2], an expert workshop (N=21), qualitative expert interviews (N=14), an online survey (N=53) (see [5] for details), and nominal group consensus in the study team including a ML scientist, two nursing scientists, a business informatics scientist, and an expert from the Association for Digitalization in Social Welfare.

Secondly, these findings were validated with empirical and theoretical evidence from a systematic literature review on AI readiness assessments and AI maturity models in healthcare and nursing (N=8), resulting in five distinct dimensions.

Thirdly, we asked 13 experts working in ongoing German AINC projects for feedback on terms reported in the literature to describe and distinct AI readiness levels in order to establish a general vocabulary for different AINC readiness levels. We then used GPT-4o to create an initial version of the AINCRA from the input of the general vocabulary for level description and the short titles of all attributes identified in the prior steps.

Fourthly, we discussed the initial version in a Think-Aloud study with 18 experts from data and information science, nursing science, nursing management, and nursing practice who have practical experience in conducting AINC projects and will use their feedback for developing the final AINCRA version.

Ethical clearing for our study was granted by the ethics committee of the German Society of Nurisng Science (application 22-030).

4. Results: The AI-Nursing-Care-Readiness-Assessment (AINCRA)

The comprehensive AINCRA consists of five dimensions: regulatory, processual, technical, social and ethical factors, and community building. These dimensions each comprise different numbers of individual attributes – in total, 68 AI-Nursing-Care-Readiness attributes were named. Table 1 presents an overview of the AINCRA dimensions.

Table 1. Dimensions of the AI-Nursing-Care-Readiness-Assessment.

Dimension	Description
Regulatory Requirements (9 Attributes)	This dimension considers the necessary legal frameworks and data protection regulations, which require an analysis of the data pool and consideration of data-sharing models to enable the use of AI systems in nursing practice.
Processual Requirements and Translational Aspects (41 Attributes)	This dimension covers the human, material, temporal, and intangible resources for AINC projects in care facilities or hospitals, as well as their general and AI-specific digitalization level. It also assesses the maturity of overarching strategies for AI, data governance, and IT governance, as well as for long-term external support and evaluation of AI implementation. Reflections on the implementation of demand-oriented research and development of AI systems with practical benefit and added value, along with existing AI-related knowledge and associated competencies in care facilities and hospitals, aim to seamlessly integrate AI systems into existing care processes. Considerations of stakeholders' attitudes towards AI, as well as acceptance of and trust in AI systems in nursing practice are included, to support the use of AI systems and address concerns of those affected by AI use.
Technical Requirements (4 Attributes)	This dimension addresses the availability and functionality of the technical infrastructure in terms of data protection and data security, which are essential for reliable data use and analysis. To enable smooth, secure, and efficient data integration and data exchange between different systems and actors in nursing practice, this dimension also captures the integration of AI systems into data infrastructures or data platforms and the use of technical interoperability standards and nomenclatures.
Social and Ethical Aspects (11 Attributes)	This dimension includes ethical and methodological considerations of voluntariness, privacy, fairness, and transparency, and engages with the ethical-normative value orientations of the nursing profession and individual practice partners in AINC projects. Addressing the impacts of AI use on micro, macro, and meso levels, as well as responsible data management, will support the use of AI systems in nursing practice.
Community Building (3 Attributes)	This dimension aims to promote a network that strengthens the exchange of knowledge and the collective development of AI systems in nursing practice. Researchers, developers, and stakeholders from nursing practice and management are involved in this network.

To guide structured evaluation, an assessment tool with different maturity levels to help practitioners evaluate and optimize AINC projects throughout the project process will be developed.

Participants of the Think-Aloud study emphasized, that the resulting AINCRA should most likely not be conducted by one and the same person of an AINC-project team. Some attributes require a thorough assessment by different disciplines, e.g., the attributes "Practical benefit and added value" and "Reflection on the significance of involving humans as an intermediary between AI systems and actions, and the potentially resulting developmental consequences" which should ideally be assessed by nursing scientists, nurse practitioners and AI developers alike.

A full German and English language version of the AINCRA will be prepared for publication soon. The German language AINCRA will also be made available as an open access online resource.

6. Conclusion

Despite both the need and potential for developing and implementing AI systems in nursing practice, research and development projects as well as organizational implementation efforts still struggle to address relevant barriers and success factors. We have presented an overview of the dimensions of the AINCRA we developed in order to aid those responsible for funding, planning, conducting and evaluating AINC-projects in addressing and overcoming known barriers.

The AINCRA is grounded in theoretical, empirical, and experiential evidence and can be used throughout the AINC-project process to support methodological, organizational and content-related decisions. Decision makers in hospitals and care facilities can utilize the AINCRA to assess their readiness to participate in AINC projects and to develop a roadmap for enhancing their AI-nursing-care specific capabilities.

Future research can build on the AINCRA as a tool with which it is possible to compare between different AINC projects and which can be used in formative or summative evaluations of AINC projects.

Acknowledgements

We would like to thank the German Federal Ministry for Education and Research as the responsible funding body (Grant-Number 16SV8835).

References

[1]. S. O'Connor, Y. Yan, F. J. S. Thilo, H. Felzmann, D. Dowding, and J. J. Lee, Artificial intelligence in nursing and midwifery: A systematic review, *J Clin Nurs,* Vol. 32, No. 13-14, Jul, 2023, pp. 2951-2968.

[2]. K. Seibert, D. Domhoff, D. Bruch, M. Schulte-Althoff, D. Fürstenau, F. Biessmann, and K. Wolf-Ostermann, Application Scenarios for Artificial Intelligence in Nursing Care: Rapid Review, *J Med Internet Res,* Vol. 23, No. 11, Nov 29, 2021, p. e26522.

[3]. F. Busch, L. Hoffmann, C. Rueger, E. H. van Dijk, R. Kader, E. Ortiz-Prado, M. R. Makowski, L. Saba, M. Hadamitzky, J. N. Kather, D. Truhn, R. Cuocolo, L. C. Adams, and K. K. Bressem, Current applications and challenges in large language models for patient care: a systematic review, *Commun Med (Lond),* Vol. 5, No. 1, Jan 21, 2025, pp. 26.

[4]. A Definition of AI: Main Capabilities and Disciplines. Definition developed for the purpose of the AI HLEG's deliverables, Independent High-Level Expert group on Artificial Intelliigence, 2019.

[5]. K. Seibert, D. Domhoff, D. Fürstenau, F. Biessmann, M. Schulte-Althoff, and K. Wolf-Ostermann, Exploring needs and challenges for AI in nursing care – results of an explorative sequential mixed methods study, *BMC Digital Health,* Vol. 1, No. 1, 2023, 13.

[6]. O. M. E. Ramadan, M. M. Alruwaili, A. N. Alruwaili, M. G. Elsehrawy, and S. Alanazi, Facilitators and barriers to AI adoption in nursing practice: a qualitative study of registered nurses' perspectives, *BMC Nurs,* Vol. 23, No. 1, Dec 18, 2024, p. 891.

[7]. Repositorien und KI-Systeme im Pflegealltag nutzbar machen (KIP), *German Federal Ministry for Education and Research*, https://www.interaktive-technologien.de/foerderung/bekanntmachungen/kip

[8]. P. Akbarighatar, I. Pappas, and P. Vassilakopoulou, A sociotechnical perspective for responsible AI maturity models: Findings from a mixed-method literature review, *International Journal of Information Management Data Insights,* Vol. 3, No. 2, 2023, 100193.

[9]. L. Pumplun, M. Fecho, N. Wahl, F. Peters, and P. Buxmann, Adoption of Machine Learning Systems for Medical Diagnostics in Clinics: Qualitative Interview Study, *J Med Internet Res,* Vol. 23, No. 10, Oct 15, 2021, pp. e29301.

[10]. D. Ferizaj, and S. Neumann, Assessing Perceptions and Experiences of an AI-Driven Speech Assistant for Nursing Documentation: A Qualitative Study in German Nursing Homes, *Human-Computer Interaction*, *Lecture Notes in Computer Science*, 2024, pp. 17-34.

[11]. S. A. Alves, S. Temme, S. Motamedi, M. Kura, S. Weber, J. Zeichen, W. Pommer, and A. Baumgart, Evaluating the Prognostic and Clinical Validity of the Fall Risk Score Derived From an AI-Based mHealth App for Fall Prevention: Retrospective Real-World Data Analysis, *JMIR Aging,* Vol. 7, Dec 4, 2024, pp. e55681.

[12]. A. Lukas, I. Maucher, S. Bugler, D. Flemming, and I. Meyer, Security and user acceptance of an intelligent home emergency call system for older people living at home with limited daily living skills and receiving home care, *Z Gerontol Geriatr,* Vol. 54, No. 7, Nov, 2021, pp. 685-694.

[13]. H. Andriesen, Playful Design for Activation. Co-designing serious games for people with moderate to severe dementia to reduce apathy, Human Factors, *TU Delft, Delft University of Technology*, 2017.

[14]. K. Seibert, D. Domhoff, K. Huter, T. Krick, H. Rothgang, and K. Wolf-Ostermann, Application of digital technologies in nursing practice: Results of a mixed methods study on nurses' experiences, needs and perspectives, *Z Evid Fortbild Qual Gesundhwes,* Vol. 158-159, Dec, 2020, pp. 94-106.

[15]. M. Hobensack, H. von Gerich, P. Vyas, J. Withall, L. M. Peltonen, L. J. Block, S. Davies, R. Chan, L. Van

Bulck, H. Cho, R. Paquin, J. Mitchell, M. Topaz, and J. Song, A rapid review on current and potential uses of large language models in nursing, *Int J Nurs Stud,* Vol. 154, Jun, 2024, pp. 104753.

[16]. EFN Policy Statement on improving frontline nurses' time for direct patient care with digitalisation and responsible AI, *European Federation of Nurses Association,* 2024.

[17]. EFN Position Statement on Nurses Co-Designing Artificial Intelligence Tools, *European Federation of Nurses Association,* 2021.

[18]. State of the world's nursing 2020: investing in education, jobs and leadership, *World Health Organization,* Geneva, 2020.

[19]. K. Wolf-Ostermann, and H. Rothgang, Digital technologies in nursing-what can they achieve?, *Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz,* Vol. 67, No. 3, Mar, 2024, pp. 324-331.

[20]. J. Becker, R. Knackstedt, and J. Pöppelbuß, Developing Maturity Models for IT Management, *Business & Information Systems Engineering,* Vol. 1, No. 3, 2009, pp. 213-222.

[21]. L. A. Lasrado, R. Vatrapu, and K. N. Andersen, Maturity Models Development in IS Research: A Literature Review, *Selected Papers of the IRIS*, Issue 6, 2015, 6.

[22]. J. Jöhnk, M. Weißert, and K. Wyrtki, Ready or Not, AI Comes— An Interview Study of Organizational AI Readiness Factors, *Business & Information Systems Engineering,* Vol. 63, No. 1, 2020, pp. 5-20.

[23]. L. Pumplun, C. Tauchert, and M. Heidt, A new organizational chassis for artificial intelligence - Exploring organizational readiness factors, in *Proceedings of the IRIS38 Conference: System design for, with and by users,* Oulu, Finland, August 9-12, 2015.

[24]. L. Weinert, J. Müller, L. Svensson, and O. Heinze, Perspective of Information Technology Decision Makers on Factors Influencing Adoption and Implementation of Artificial Intelligence Technologies in 40 German Hospitals: Descriptive Analysis, *JMIR Medical Informatics,* Vol. 10, No. 6, 2022, 2022.

(024)

Catastrophic Forgetting Mitigation of Melanoma Skin Cancer Detection Using ADANemo

Jesse Orlanda, Nandatama Bagus Adisaka, William Suryadharma Pangestu and Hidayaturrahman
Computer Science Department, School of Computer Science, Bina Nusantara University, Jakarta, 11480, Indonesia
E-mail: jesse.orlanda@binus.ac.id, nandatama.adisaka@binus.ac.id, william.pangestu@binus.ac.id, hidayaturrahman@binus.edu

Summary: Differences in the distribution of characteristics within skin disease datasets often result in domain shifts, which can lead to catastrophic forgetting during the process of melanoma detection. This issue has become a critical concern in AI development as it significantly affects the accuracy and efficiency of diagnostic outcomes. This paper proposes a new model Adaptation with Adapter Network Model (ADANemo), which leverages DANN, ResNet-50, and a network architecture to address this challenge and improve detection accuracy. Several datasets referenced in this study included HAM10000, BCN20000, PROVe AI, ISIC 2019, and ISIC 2020. Experimental results show that ADANemo improves source domain accuracy from 63.71% to 83.71%, significantly reducing knowledge loss.

Keywords: Melanoma, Catastrophic forgetting, Domain adaptation, Adapter, Transfer learning.

1. Introduction

Melanoma is the most dangerous and deadly type of skin cancer. Early detection is crucial to minimize the spread of this disease [1, 2]. However, the challenges in detecting melanoma lie in the diverse distribution of skin cancer characteristics, such as geographical location, skin type, medical imaging devices, and other factors influencing the data. The difference in distribution between the available data and the data to be predicted is called the domain discrepancy [3]. Meanwhile, significant changes in data distribution are termed domain shift [4, 5].

In medical practice, non-invasive diagnostic tools such as dermoscopy are commonly used to detect skin cancer. However, the effectiveness of this tool is highly dependent on the skill and experience of the dermatologist, which can limit its accessibility to a broader population [6, 7]. Previous studies have shown that dermoscopy can only identify approximately 60-90% of melanoma cases based on visual examination [8]. Furthermore, variability in diagnostic outcomes among doctors with differing levels of expertise often leads to inaccuracies, highlighting the need for supporting technologies to enhance diagnostic consistency and reliability.

In addition, Artificial Intelligence (AI)-based skin cancer detection methods have been widely developed to prevent the spread of this disease. Among 14 studies that used shallow techniques, the Support Vector Machine (SVM) emerged as the most popular model for classification [9]. Meanwhile, among 39 studies using deep techniques, the most frequently used models were Convolutional Neural Networks (CNN) and Residual Networks (ResNet). Other models that have been used for skin cancer detection include Naïve Bayes (NB), Logistic Regression (LR), k-Nearest Neighbor (kNN), Random Forest (RF), Inception, AlexNet, MobileNet, Visual Geometry Group (VGG), Xception, DenseNet, along with ensemble and hybrid approaches. [10, 11].

Although many studies have been conducted, the development of these models can only predict data with similar characteristics. When tested on diverse characteristics (referred to as different domains), the models struggle to make accurate predictions [12]. The concept of domain adaptation plays a crucial role in addressing this issue, where the model can minimize domain discrepancy and reduce domain shift between the available data (source domain) and the predicted data (target domain) [4, 5, 13].

However, when data trained on one domain is tested again on the previous domain, the model often forgets the training it has undergone, resulting in suboptimal detection [14, [15]. This phenomenon is known as catastrophic forgetting, where there is a performance decline in previously trained domains [16, 17]. Catastrophic forgetting is a significant concern in AI development as it greatly affects the accuracy and efficiency of testing. Therefore, this study addresses catastrophic forgetting to ensure that models do not forget previously conducted training, thereby maintaining their performance.

Several previous studies on detection have used Domain-Adversarial Neural Networks (DANN) methods to reduce domain shift. DANN works with adversarial learning techniques to generate feature representations independent of the source or target domain, known as domain-invariant. This allows the model to generalize well to the target domain without losing performance [18]. This approach helps the model adapt and maintain diagnostic accuracy across different patient demographics [5, 19]. While helping to reduce domain shift, the model is still perceived to

be suboptimal in minimizing the occurrence of catastrophic forgetting.

Therefore, this study proposes using a new model, Adaptation with Adapter Network Model (ADANemo). ADANemo is based on using an enhanced DANN model with a feature extractor, ResNet-50. The additional implementation of the Network model allows it to process data and analyze patterns to generate predictions. This model is expected to achieve optimal results in reducing catastrophic forgetting and achieving high accuracy in melanoma detection.

2. Methodology

This section describes the process of conducting the research, starting with datasets collection from ISIC Archive, followed by identifying domain discrepancies to ensure good data quality and addressing catastrophic forgetting. Subsequently, the data is trained using the proposed new model and an evaluation is conducted to assess the resulting accuracy.

2.1. Dataset

The dataset utilized in this study is derived from the following sources: HAM10000 [20], BCN20000 [21], and PROVe-AI serve as the source domain [22], while ISIC 2019 [23] and ISIC 2020 [24] act as the target domain. These datasets were sourced from ISIC Archive. Each domain within the dataset is categorized into two groups of labels: melanoma and non-melanoma. All images were resized to 224 × 224 pixels to maintain consistency dimensions. Lastly, the dataset is reorganized into folders, with class_0 containing melanoma images and class_1 containing non melanoma images.

2.2. Determining Catastrophic Forgetting

Differences in data distribution between the source and target domains, such as variations in characteristics, colors, and skin visuals are detected by identifying symptoms of domain discrepancies. These discrepancies are identified through the following analyses:

a. Average color, measuring the mean values of red, green, and blue channels.
b. Average brightness, determining the mean brightness level of the images.
c. Maximum Mean Discrepancy (MMD), measuring the extent of data distribution differences.
d. Wasserstein distance, calculating the distance between two-pixel probability distributions in the images.
e. t-SNE visualizes high-dimensional data in a two or three-dimensional map.

Catastrophic forgetting aims to measure the accuracy difference between a model trained on the source domain and the one trained on the target domain. This process involves training the model on the source domain and subsequently testing it on both domains to obtain accuracy results (A) and (B). The model is then trained on the target domain and retested on both domains to obtain accuracy results (C) and (D). Catastrophic forgetting in the source domain can be identified by calculating the difference between accuracy results (C) and (A). Meanwhile, accuracy values (B) and (D) serve as comparison metrics for the target domain. The flowchart of determining catastrophic forgetting is depicted in Fig. 1.

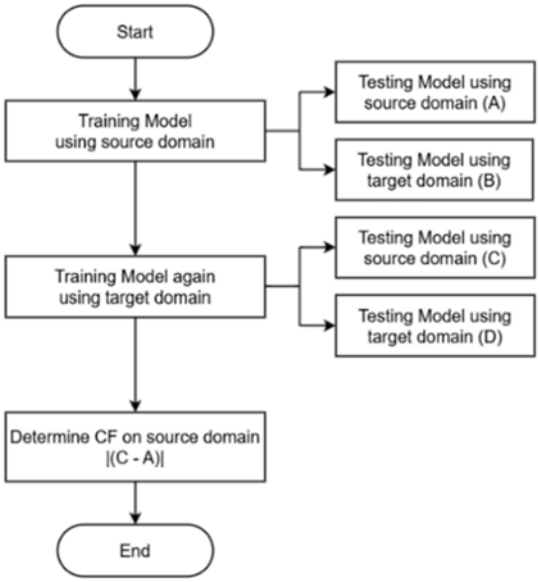


Fig. 1. Flowchart of determining catastrophic forgetting.

2.3. Proposed Architecture

The proposed model, ADANemo follows the DANN approach and consists of four main components:

a. Feature extractor, based on a pre-trained ResNet-50.
b. Adapter, neural network component acts as a transformation layer between feature extractor and the classifiers.
c. Label classifier, for classifying melanoma and non-melanoma labels.
d. Domain classifier, integrated with Gradient Reversal Layer (GRL) to create a confusion model that separates domain-specific characteristics.
e. A pre-trained ResNet-50 was employed as the feature extractor, leveraging its demonstrated efficacy and widespread adoption as a standard model in the field. ADANemo's modular design facilitates straightforward substitution with alternative architectures, enabling adaptation to future advancements.

This approach is enhanced by data augmentation techniques, including resizing, flipping, rotating, affine transformations, and color jittering to minimize

overfitting. Model training will be conducted with specific parameter configurations, such as number of epochs, batch size, learning rate, fully connected layers, and alpha values, as detailed in Table 1.

Table 1. Parameter and value.

Parameter	Value
Feature_extractor	ResNet-50 pretrained
Epoch	30
Alpha	Logistic: from 0 (begin) to 1 (end of epoch)
Batch_size	16
Learning_rate	0.0005
FC_layer	256
Resize	224 × 224 pixel
Random_rotation	30
Random_horizontal_flip	true
Random_affine	degrees=20, translate=(0.2, 0.2), scale=(0.8, 1.2)
Color_jitter	brightness=0.3, contrast=0.3

The model's performance is evaluated on both the source and target domains using accuracy, precision, recall, and F1-score metrics. The results are compared against conventional model evaluation (stage 2) to assess whether the proposed approach mitigates catastrophic forgetting.

3. Result and Discussion

3.1. Domain Discrepancy

The analysis of average color distribution graphs for the Red (R), Green (G), and Blue (B) channels, as shown in Fig. 2, reveals distinct differences in their distribution patterns between the source and target domains.

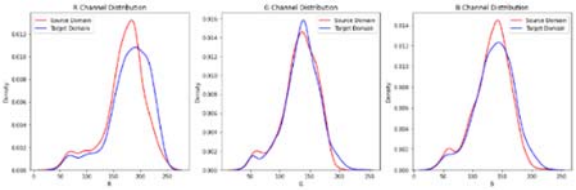


Fig. 2. Distribution of average color values.

In the Red channel, the source domain (represented by the orange line) tends to have higher values, whereas the target domain (depicted by the blue line) displays a subtle rightward shift. Similarly, the Green channel shows comparable trends, however, the target domain exhibits a more pronounced peak. Notably, the Blue channel presents unique variations, characterized by a higher peak in the source domain.

Despite the relatively minor discrepancies observed when comparing RGB distributions between domains, these subtle distinctions still indicate potential domain discrepancies, particularly within both the Red and Blue channels.

As presented in Fig. 3, the brightness distribution graph highlights slight variations between source and target domains. Specifically, the source domain (illustrated using a blue line) features a more concentrated luminance distribution with a more prominent peak compared to the target domain (indicated by a green line). This latter domain exhibits a broader distribution with tendencies toward lower luminance values. These luminosity differences further reinforce the indication of domain discrepancies.

The quantitative analysis, as illustrated in Table 2, provides additional insight through Maximum Mean Discrepancy (MMD) and wasserstein distance calculations. The MMD value of 0.0042 indicates a marginal distributional difference, while the wasserstein distance of 1.1833 reveals a significant distributional separation between source and target domains.

As shown in Fig. 4, the t-SNE visualization illustrates the feature distribution of the source and target domain datasets, revealing both overlapping and distinct patterns. While the two domains largely overlap, a particular cluster of target domain samples (orange points) is observed at the bottom of the plot. This separation suggests a unique feature distribution in the target domain and indicates signs of domain discrepancies.

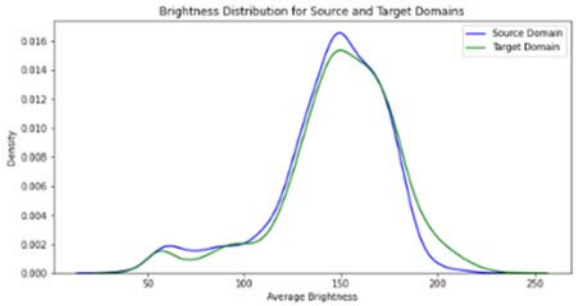


Fig. 3. Distribution of average rightness values.

Table 2. Maximum Mean Discrepancy (MMD) and wasserstein distance metrics.

Method	Value
MMD	0.0042
Wasserstein Distance	1.1833

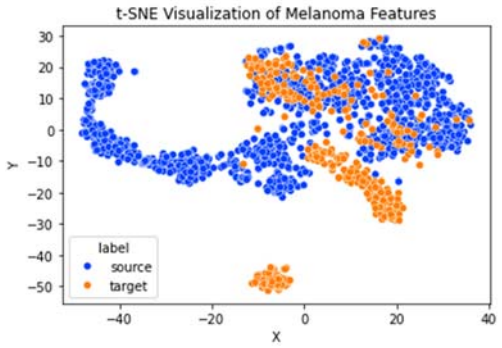


Fig. 4. Visualization of Melanoma Features.

3.2. Catastrophic Forgetting

The catastrophic forgetting analysis investigates the performance of a pre-trained ResNet-50 model when trained on both source and target domains. This investigation yields critical insights into the model's knowledge retention and adaptation capabilities. In evaluating the source domain performance, the initial accuracy was recorded at 83.75% (A), which significantly decreased to 63.71% (C) following training on the target domain. This 20.04% reduction in accuracy illustrates substantial knowledge erosion, indicating a diminished capacity for the model to recognize patterns in the original source domain. The model's accuracy without knowledge of the target domain was 68.38% (B). However, after retraining with the target domain, accuracy improved to 85.50% (D).

The proposed methodology aims to address these performance variations by developing strategies to mitigate the decline in source domain accuracy while preserving learning improvements in the target domain. The primary objectives include recovering performance closer to the original 83.75% accuracy in the source domain and maintaining the unsupervised learning gains observed in the target domain. The result shown in Table 3.

Table 3. Performance evaluation metrics for source and target domains.

Parameter	(A)	(B)	(C)	(D)
Accuracy	83.75%	**68.38%**	**63.71%**	85.50%
Precision	80.04%	67.05%	67.67%	83.97%
Recall	89.92%	72.25%	52.50%	87.75%
F1-score	84.69%	69.55%	59.13%	85.82%

3.3. ADANemo Approach

The ADANemo architecture was implemented through training over 30 epochs, utilizing a learning rate of 0.0005 and a batch size of 16. The implementation was done in PyTorch on computational hardware featuring an AMD Ryzen 9 5900HX CPU (3.30 6Hz), 16 GB RAM, and an NVIDIA GeForce RTX 3060 GPU, executed via Jupyter Notebook server.

Performance evaluation of the proposed model involved comprehensive metrics including accuracy, precision, recall, and F1-score, as detailed in Table 4.

The evaluation results for the source domain are illustrated through the confusion matrix in Fig. 5, which shows that the ADANemo architecture tested against this domain yielded performance results as described in Table 5.

Meanwhile, evaluation results for the ADANemo architecture in the target domain can be observed in Table 6 and Fig. 6.

Table 4. Accuracy, precision, recall, and F1-score metrics.

$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$	(1)
$Precision = \frac{TP}{TP + FP}$	(2)
$Recall = \frac{TP}{TP + FN}$	(3)
$F1\ Score = \frac{2 \times Precision \times Recall}{Precision + Recall}$	(4)

Table 5. Accuracy, precision, recall, and F1-score metrics for source domain testing dataset.

Parameter	Metrics Performance
Accuracy	83.71%
Precision	90.65%
Recall	75.17%
F1-score	82.20%

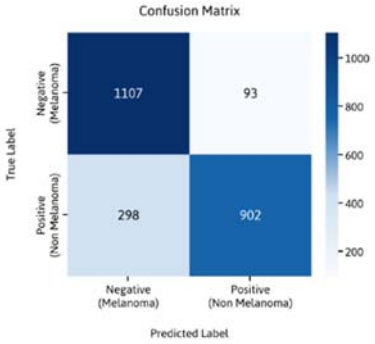


Fig. 5. Confusion matrix on source domain testing dataset.

Table 6. Accuracy, precision, recall, and F1-score metrics for target domain testing dataset.

Parameter	Metrics Performance
Accuracy	69.58%
Precision	72.86%
Recall	62.42%
F1-score	67.22%

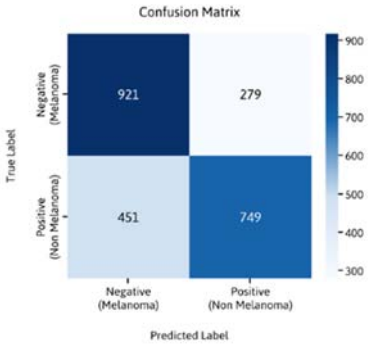


Fig. 6. Confusion matrix on target domain testing dataset.

A comparative analysis between the ADANemo architecture and baseline models revealed significant insights. Both models utilized a pre-trained ResNet-50 model, with performance comparisons documented in Table 7.

Table 7. Performance comparison of baseline model and ADANemo architecture.

Model	Domains	Accuracy
Baseline	Source	63.71% (i)
	Target	68.38% (ii)
ADANemo	Source	83.71% (iii)
	Target	69.58% (iv)

Description:
(i). Baseline model accuracy result at point (C)
(ii). Baseline model accuracy result at point (B)
(iii). ADANemo accuracy result at source domain
(iv). ADANemo accuracy result at target domain

Fig. 7 graphically demonstrates ADANemo's performance. In the source domain, while the baseline model suffered from catastrophic forgetting, ADANemo achieved an impressive accuracy improvement of 20% (from 63.71% to 83.71%). This indicates ADANemo's enhanced capability for learning across both source and target domains.

However, the architecture's performance on the target domain improved by 1.2% (from 68.38% to 69.58%). While modest, this improvement is valuable given our primary focus on source domain catastrophic forgetting mitigation, which achieved a substantial 20% accuracy gain. This demonstrates the potential optimization opportunities within the unsupervised learning methodology.

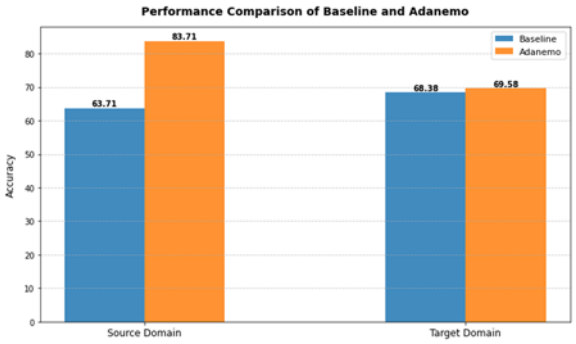


Fig. 7. Accuracy comparison of baseline model and ADANemo architecture.

The loss function visualization in Fig. 8 provides further insight into the model's learning challenges. While there is an overall downward trend in label loss, certain epoch-level instabilities are evident. The domain loss remained relatively constant at approximately 0.6, contributing to a marginally significant average loss reduction.

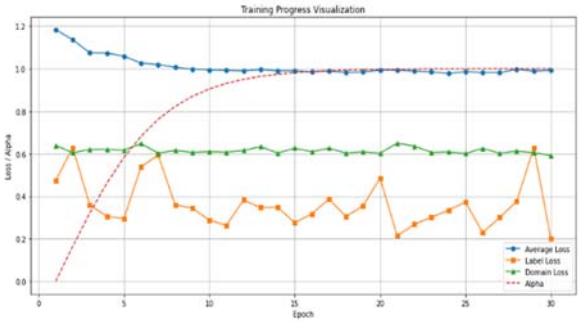


Fig. 8. Loss function trends on model training.

4. Conclusions

The ADANemo has effectively reduced catastrophic forgetting in the source domain. The model shows a significant accuracy improvement, increasing from 63.71% with the baseline model to 83.71% with ADANemo. However, its ability to identify patterns in the target domain using an unsupervised learning approach is still lacking. Future studies should focus on improving adaptation techniques and exploring advanced optimization strategies to enhance the model's pattern recognition in the target domain and boost overall performance in future research.

References

[1]. R. Groves, Quantifying the mechanical properties of skin in vivo and ex vivo to optimize microneedle device design, Ph.D. dissertation, *Cardiff Univ.,* Cardiff, U.K., 2011.
[2]. Z. R. Garrison, C. M. Hall, R. M. Fey, et al., Advances in early detection of melanoma and the future of at home testing, *Life*, Vol. 13, No. 4, 2023, p. 974.
[3]. Z. Han, H. Sun, and Y. Yin, Learning transferable parameters for unsupervised domain adaptation, *IEEE Trans. Image Process.,* Vol. 31, 2022, pp. 6424–6439.
[4]. S. Chamarthi, K. Fogelberg, T. J. Brinker, and J. Niebling, Mitigating the influence of domain shift in skin lesion classification: A benchmark study of unsupervised domain adaptation methods, *Inform. Med. Unlocked,* Vol. 44, 2024, p. 101430.
[5]. S. Q. Gilani, M. Umair, M. Naqvi, O. Marques, and H. C. Kim, Adversarial training-based domain adaptation of skin cancer images, *Life*, Vol. 14, No. 8, 2024, p. 1009.
[6]. C. Xin, Z. Liu, K. Zhao, et al., An improved transformer network for skin cancer classification, *Comput. Biol. Med.,* Vol. 149, 2022, p. 105939.
[7]. A. A. Nugroho, I. Slamet, and S. Sugiyanto, Skin cancer identification system of HAM10000 skin cancer dataset using convolutional neural network, in *AIP Conf. Proc*., Vol. 2202, 2019, 020039.
[8]. A. S. Qureshi and T. Roos, Transfer learning with ensembles of deep neural networks for skin cancer detection in imbalanced data sets, *Neural Process. Lett.,* Vol. 55, No. 4, 2023, pp. 4461–4479.
[9]. G. B. T. R., An efficient skin cancer diagnostic system using bendlet transform and support vector machine,

An. Acad. Bras. Ciênc., Vol. 92, No. 1, 2020, e20190554.

[10]. T. M. de Carvalho, E. Noels, M. Wakkee, A. Udrea, and T. Nijsten, Development of smartphone apps for skin cancer risk assessment: Progress and promise, *JMIR Dermatol.,* Vol. 2, No. 1, 2019, e13376.

[11]. A. Takiddin, J. Schneider, Y. Yang, A. Abd-Alrazaq, and M. Househ, Artificial intelligence for skin cancer detection: Scoping review, *J. Med. Internet Res.,* Vol. 23, No. 11, 2021, e22934.

[12]. A. C. Morgado, C. Andrade, L. F. Teixeira, and M. J. M. Vasconcelos, Incremental learning for dermatological imaging modality classification, *J. Imaging,* Vol. 7, No. 9, 2021, p. 180.

[13]. N. P. García-de-la-Puente, M. López-Pérez, L. Launet, and V. Naranjo, Domain adaptation for unsupervised cancer detection: An application for skin whole slides images from an interhospital dataset, in *Proceedings of the Medical Image Computing and Computer Assisted Intervention Conference (MICCAI' 2024)*, 2024, pp. 58–68.

[14]. D. Saunders and S. DeNeefe, Domain adapted machine translation: What does catastrophic forgetting forget and why?, *arXiv preprint arXiv:2412.17537*, 2024.

[15]. P. Kumari, J. Chauhan, A. Bozorgpour, R. Azad, and D. Merhof, Continual learning in medical imaging analysis: A comprehensive review of recent advancements and future prospects, *arXiv preprint arXiv:2312.17004,* 2023.

[16]. P. Bándi, M. Balkenhol, M. Van Dijk, B. van Ginneken, J. van der Laak, and G. Litjens, Domain adaptation strategies for cancer-independent detection of lymph node metastases, *arXiv preprint arXiv:2207.06193,* 2022.

[17]. B. Thompson, J. Gwinnup, H. Khayrallah, K. Duh, and P. Koehn, Overcoming catastrophic forgetting during domain adaptation of neural machine translation, in *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,* 2019, pp. 2062–2068.

[18]. Q. Gou and G. Cui, Adversarial selective domain adaptation with feature cluster for skin cancer diagnosis, 2024, in print.

[19]. Y. Yu, P. Y. Chen, Y. Zhou, and J. Mei, Adversarial sample enhanced domain adaptation: A case study on predictive modeling with electronic health records, *arXiv preprint arXiv:2101.04853*, 2021.

[20]. ISIC Archive, HAM10000. [Online]. Available: https://api.isic-archive.com/collections/212 [Accessed: Dec. 8, 2024].

[21]. ISIC Archive, BCN20000. [Online]. Available: https://api.isic-archive.com/collections/249 [Accessed: Dec. 8, 2024].

[22]. ISIC Archive, PROVe-AI. [Online]. Available: https://api.isic-archive.com/collections/218 [Accessed: Dec. 8, 2024].

[23]. ISIC Archive, Challenge 2019: Training, 2019. [Online]. Available: https://api.isic-archive.com/collections/65 [Accessed: Dec. 8, 2024].

[24]. ISIC Archive, Challenge 2020: Test, 2020. [Online]. Available: https://api.isic-archive.com/collections/68 [Accessed: Dec. 8, 2024].

(025)

AI for Education: A Generative AI-Powered Cognitive Tool for Medical Students' Self-Assessment

A. Pitrone [1] and I. Haddiya [2]
[1] Loop AI Global LLC, 1732 1St Ave, Suite 21427, New York, NY, 10128, United States
[2] Department of Nephrology, Faculty of Medicine and Pharmacy, University Mohammed First, Oujda, Morocco
Tel.: + 0039 3423219559
E-mail: andrea@loop.ai

Summary: This article presents a Generative AI-powered Cognitive Application designed to help medical students automatically self-assess their knowledge while preparing for exams on key Nephrology topics.
The system functions as a Cognitive Agent, mimicking the evaluation process of a human investigator—such as a professor—by assessing not only the accuracy and completeness of students' answers but also their confidence in each topic. This dual-layered evaluation ensures a more comprehensive and insightful self-assessment, helping students identify areas where they need improvement.

By leveraging advanced AI-driven analysis, this application aims to provide personalized feedback and enhance students' learning experiences. With the potential to be expanded across various medical fields, it represents a significant step forward in integrating AI into medical education, fostering more effective and data-driven learning strategies for future healthcare professionals.

Keywords: Medical students, Self-assessment, Generative AI, Cognitive application, Large language model (LLM), Convolutional neural network (CNN), Micro-facial expressions.

1. Introduction

Accurate self-assessment is vital for doctors to identify strengths and weaknesses, driving lifelong learning and skill refinement [1].

Research underscores its key role in medical education, enhancing both professional growth and patient care. [2].

Medical diagnoses and treatments follow standardized morbidity protocols — a globally recognized set of steps designed to ensure consistency and accuracy in patient care. Assessing medical students' understanding of these protocols requires verifying their knowledge using precise medical terminology, abbreviations, synonyms, and key diagnostic and treatment concepts.

This article introduces a Generative AI-powered Cognitive Application that enables automated self-assessment for students preparing for medical exams (e.g., Nephrology). The system, functioning as a Cognitive Agent, is designed to match the evaluation accuracy of a human investigator (e.g., a professor), assessing not only the accuracy and completeness of responses but also the student's confidence in each topic.

The study initially led to the development of a Cognitive Agent designed to assess students' competence in Hypertension [3]. Building on its promising results, the scope was expanded to include nearly all Nephrology topics as documented in this article, thus introducing significant novelty. Future efforts will focus on extending its application to Cardiology and other medical fields, further enhancing medical education.

2. Implementation Methods

We conducted a prospective, descriptive, and analytical study at the Mohammed VI University Hospital of Oujda, Morocco, involving fifty fifth-year volunteer medical students. Participants were provided with comprehensive documentation on the diagnosis, management, and pathophysiology of main Nephrology-related topics:

- Hypertension [4];
- Chronic Kidney Disease [5];
- Diagnosis, Evaluation, Prevention, and Treatment of Chronic Kidney Disease–Mineral and Bone Disorder [6];
- Diabetes Management in Chronic Kidney Disease [7];
- Management of Glomerular Diseases [8];
- Lipid Management in Chronic Kidney Disease [9];
- Management of Lupus Nephritis [10];
- Acute Kidney Injury [11].

To measure the accuracy of the Cognitive Agent, we compared students' knowledge evaluation results using two assessment methods:

- AI-Based Exam Simulation: Students answered ten open-ended questions automatically generated by the Cognitive Agent.
- Human Investigator-Based Assessment: Students participated in a 15-minute oral examination, during which a Professor asked the same set of questions generated by the AI system.

This comparative approach, combined with the key findings from the "user satisfaction survey (please refer to par. 5), validated the effectiveness and

reliability of the AI-driven self-assessment tool in evaluating medical students' knowledge.

3. Cognitive Agent Operational Principles

The Cognitive Agent is initially provided with reference documentation for each medical topic [4-12].

During the exam simulation, the Cognitive Agent autonomously generates ten direct open-ended questions for the candidate and records a video of the session. It then evaluates the correctness and completeness of the answers based on the following criteria:

- Semantic Similarity: The wording of the response is compared to the reference documentation to ensure conceptual alignment.
- Keyword and Concept Inclusion: The answer must contain all relevant concepts, keywords, synonyms, or abbreviations. If only a partial set is detected, a completeness percentage is calculated.

Additionally, the Cognitive Agent analyzes the student's micro-facial expressions—based on Paul Ekman's theory [13]—to assess their reliability and self-confidence in each topic while responding.

4. Cognitive Agent Components

The Cognitive Agent is composed of two primary AI components:

- Natural Language Engine (NLE): This component handles natural language understanding, processing, and generation. It is powered by an open-source Large Language Model (LLM) that has been fine-tuned with medical-specific content. This content is derived from the documentation provided to the students, which the Cognitive Agent uses to automatically generate exam simulation questions. For this component, we have selected the LLAMA 3.1 Large Language Model (LLM), which contains 70 billion parameters and is based on an optimized Transformer architecture.

Fig. 1 illustrates the cognitive components within the Natural Language Engine (NLE) module.

- Multimodal Sentiment Analyzer (MSA): This component is a CNN-based component developed to identify the seven basic emotions defined by Paul Ekman: Happiness, Sadness, Fear, Disgust, Anger, Contempt, and Surprise - it is exploited to assess emotions from candidates' exam videos by analyzing:
  - Micro-facial expressions from both individual frames and the full video.
  - Audio tone and nuances in the candidate's voice.
  - Text transcribed from speech using speech-to-text conversion.

Fig. 2 showcases an example of Multimodal Sentiment Analysis, demonstrating how these categories are applied in practice.

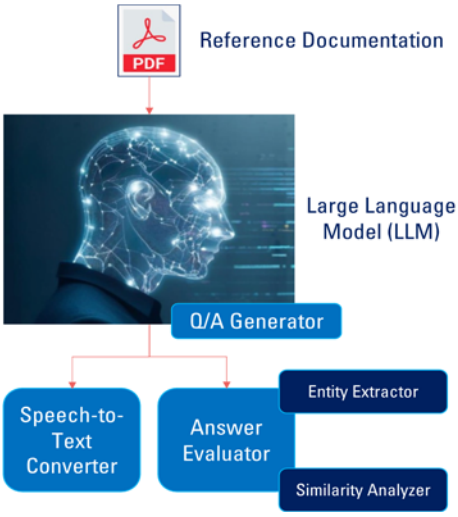


Fig. 1. Components of the AI Agent's Natural Language Engine.

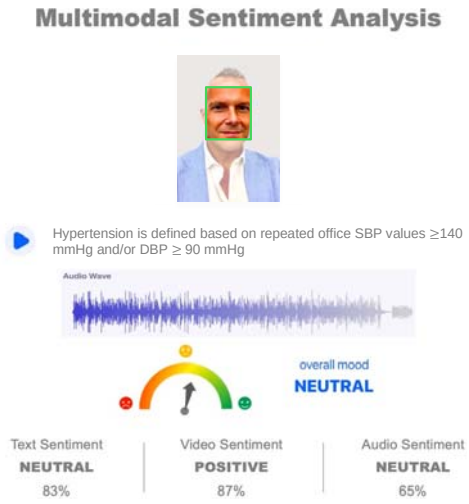


Fig. 2. Example of Facial Expression Analysis.

Fig. 3 illustrates the cognitive components of the Multimodal Sentiment Analysis (MSA) module.

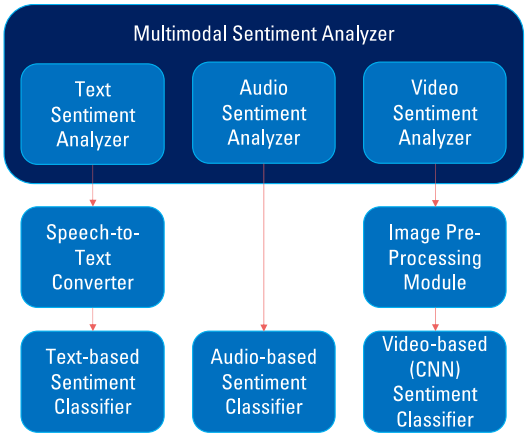


Fig. 3. Components of the AI Agent's Multimodal Sentiment Analysis module.

5. Results Obtained

The "Exam Readiness %" metric represents a student's preparedness for the exam. It is calculated by evaluating the accuracy and completeness of each response, averaging the percentage scores from all ten open-ended questions, and factoring in the student's confidence, as assessed by the MSA module.

To measure the Cognitive Agent's accuracy, we have adopted the Mean Squared Error (MSE), as calculated in equation (1).

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - p_i)^2 \qquad (1)$$

where: n is the number of predictions (n =50 for our study), Y is the vector of observed values (i.e., Human Investigator's evaluation) of the variable being predicted, P is the predicted values.

The assessments provided by the Cognitive Agent closely match those of the Human Investigator, as shown in the following figure. This comparison highlights the alignment between the evaluations from both the Cognitive Agent and the Human Investigator (please refer to Fig. 4).

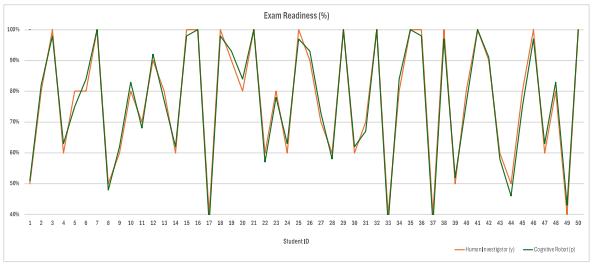


Fig. 4. Results Obtained – Exam Readiness Percentage.

The calculated Mean Squared Error (MSE) value is 0.0007, confirming the high accuracy of the AI solution (as represented in Fig. 5).

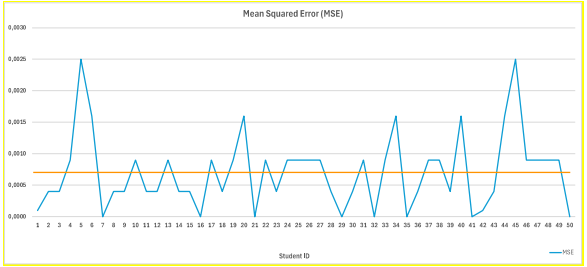


Fig. 5. Results Obtained – Mean Squared Error.

Lastly, a user satisfaction survey was conducted to evaluate the effectiveness and overall user experience of an AI-driven solution designed for medical education.

The results demonstrated a highly positive reception, with 94% of respondents affirming the solution's effectiveness in enhancing their learning process.

Users particularly appreciated the AI's ability to provide accurate self-assessments, personalized feedback, and targeted recommendations, which significantly improved their confidence in exam preparation.

Additionally, the survey revealed a strong willingness to recommend the solution, indicating high trust and perceived value among users.

The overwhelmingly positive feedback suggests that the AI solution successfully addresses critical learning needs, offering an intuitive and efficient tool for medical students.

These findings reinforce the potential for AI-driven educational technologies to revolutionize self-assessment methods, paving the way for broader applications in medical and other professional training domains.

Future improvements will further refine the system based on user insights and evolving educational requirements.

Fig. 6 presents the results of the user satisfaction survey conducted with fifty fifth-year volunteer medical students, validating the solution's effectiveness.

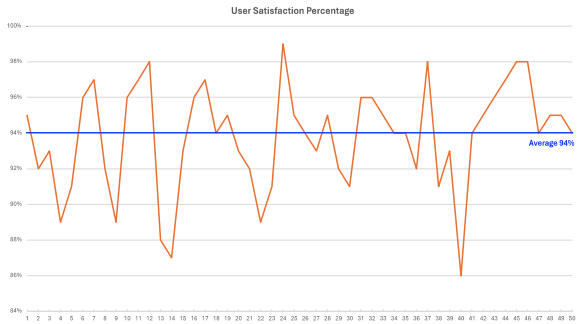


Fig. 6. User Satisfaction Survey results.

6. Prospects for Future Research

Our future research will focus on the following key areas:

- Expanding Student Participation: Increasing the number of participants to strengthen the validation process and further enhance the accuracy of the Cognitive Agent.
- Expanding Medical Applications: Extending the Cognitive Agent to additional domains, such as Cardiology and Endocrinology, to enhance its generalizability.
- Developing Real-Time AI Adaptation: Implementing an adaptive AI system that generates follow-up questions in real time, based on detected uncertainties in student responses, to ensure a deeper understanding of each topic.

7. Conclusions

The Cognitive Agent demonstrated a high level of accuracy (90%), as validated by comparing its evaluation results with those of the Human Investigator.

Additionally, a Customer Satisfaction survey completed by the participating students confirmed the usefulness and reliability of the Cognitive Agent. The positive feedback indicates that students would be willing to adopt this automated AI-driven assessment tool for other topics and fields of study.

This study was carried out in compliance with the Helsinki declaration on the protection of individuals. Students' participation was voluntary. All participants were informed of the objectives and procedures of the study. There was no request for personal information and the confidentiality of the data was guaranteed.

References

[1]. S. Prediger, K. Schick, F. Fincke, S. Fürstenberg, V. Oubaid, M. Kadmon, PO. Berberat, S. Harendza, Validation of a competence-based assessment of medical students' performance in the physician's role, *BMC Med. Educ.,* 20, 2020, art. 6.

[2]. M. Wijnen-Meijer, M. van der Schaaf, K. Nillesen, S. Harendza, O. Ten Cate, Essential facets of competence that enable trust in graduates: a Delphi study among physician educators in the Netherlands, *J Grad. Med. Educ.,* 5, 1, 2013, pp. 46–53.

[3]. I. Haddiya, A. Pitrone, Generative AI-based Cognitive Robot for exam candidates' knowledge self-assessment, in *Proceedings of the IEEE Conference on Artificial Intelligence (CAI)*, 2024, pp. 643-648.

[4]. G. Mancia et al., 2023 ESH Guidelines for the management of arterial hypertension. The Task Force for the management of arterial hypertension of the European Society of Hypertension Endorsed by the European Renal Association (ERA) and the International Society of Hypertension (ISH)*: Journal of Hypertension,* 41, 12, 2023, pp. 1874-2071.

[5]. KDIGO 2024 Clinical Practice Guideline for the Evaluation and Management of Chronic Kidney Disease, Kidney Disease Improving Global Outcomes (KDIGO), *Official Journal of the International Society of Nephrology,* Volume 105, Issue 4S, April 2024, pp. S117-S314.

[6]. KDIGO 2017 Clinical Practice Guideline Update for the Diagnosis, Evaluation, Prevention, and Treatment of Chronic Kidney Disease–Mineral and Bone Disorder (CKD-MBD), *Kidney Int Suppl,* 7, 1, 2017, pp. 1-59.

[7]. KDIGO 2022 Clinical Practice Guideline for Diabetes Management in Chronic Kidney Disease, *Kidney Int.,* 102, 5S, 2022, pp. S1-S127.

[8]. KDIGO 2021 Clinical Practice Guideline for the Management of Glomerular Diseases, *Official Journal of the International Society of Nephrology, Volume 100, Issue 4S*, Oct2021.

[9]. KDIGO 2013 Clinical Practice Guideline for Lipid Management in Chronic Kidney Disease, *Official Journal of International Society of Nephrology, Volume 3, Issue 3*, Nov2013.

[10]. KDIGO 2024 Clinical Practice Guideline for the Management of Lupus Nephritis, *Official Journal of the International Society of Nephrology, Volume 105, Issue 1S*, Jan2024.

[11]. KDIGO 2012 Clinical Practice Guideline for Acute Kidney Injury, *Official Journal of the International Society of Nephrology, Volume 2, Issue 1*, March 2012.

[12]. E.V. Lerma, M.A. Sparks, J.M. Topf, Nephrology Secrets, Fourth Edition, *Elsevier*, 2015.

[13]. K. Ramsland, The man of 1,000 faces: Paul Ekman and the science of facial analysis, *Springfield,* Vol. 22, Issue 1, 2012, pp. 65-70.

(026)

Detecting and Pre-coding Multiple Tumours in Pathology Reports

T. Niemi, G. Defossez and J.-L. Bulliard
Centre for Primary Care and Public Health (Unisanté), University of Lausanne, Lausanne, Switzerland
Tel.: + 41 213148012
E-mail: tapio.niemi@unisante.ch

Summary: Population-based cancer registries record cancer cases using international standards. This requires processing numerous free text pathology reports describing one or multiple diagnoses. We developed three convolutional neural network (CNN) methods to detect and pre-code multi-tumour reports: the cancer type detection (M1), multi-task - multi-value CNN for pre-coding (M2), and CNN enhanced by an attention layer for key term detection (M3). We evaluated each method's performance against invasive multi-tumour skin cancer reports and achieved the following results: M1: sensitivity 0.89, specificity 0.99, and positive predictive value (PPV) 0.53; M2: 0.94, 0.66, and 0.04, M3: 0.78, 0.93, and 0.18 respectively. M1 was only able to detect multiple tumours if they represented different morphology groups. When comparing M2 and M3 against positive reports, the accuracies were following: for the sub site 48 and 62%, for morphology group 71 and 81%, and for laterality 82 and 71% respectively. All three values were correct in 29% (M2), and 38-52% of cases (M3). The method will be piloted in the Vaud Cancer Registry and, if successful, integrated to the automated document management system.

Keywords: Automated medical coding, Multiple tumours, CNN, Noisy and heterogenous data, Skin cancer, Cancer registry.

1. Introduction

In many countries population-based cancer registries record invasive tumours based on a large number of free text pathology reports. This task is resource-intensive and difficult to automatise because of lack of standardised report types and templates. In this work, we focus on skin cancers that are the most frequent cancers world-wide in fair-skinned populations [1], and especially on squamous cell carcinoma (SCC) and cutaneous malignant melanoma (CMM). A special challenge in skin cancer pathological reports is that they usually contain multiple diagnoses (one for each biopsy). These diagnoses can describe different tumours or negative results. Depending on types of diagnosis, none, one, or multiple tumours need to be recorded based on the international cancer registration standards [2]. For each case at least the sub site, morphology, behaviour, and laterality are coded.

We developed a method based on neural networks using previously coded cases as training data. Neural networks are commonly applied to medical coding [3-5] but usually a report contains only one positive diagnosis. However, in our case, one or multiple codes can be assigned to one report and coding may also depend on previous tumours of the patient, not just the report at hand. This all means that our training data contains a significant amount of noise. Attention methos have been applied in medical information extraction and coding [3, 6, 7] but not for detecting multiple diagnoses in a single report.

Our aim is to introduce a method that can 1) identify whether a report contains diagnosis of multiple invasive tumours, and 2) extract the key terms required for pre-coding sub site, morphology, behaviour, grade, laterality, and first line of treatment of skin cancer cases identified in diagnoses. In this paper, we focus only for sub site, morphology, and laterality.

2. Material and Methods

We used previously coded and human-verified cases as training, validation, and test data. The dataset included skin cancer cases diagnosed between 2014 and 2022, consisting primarily of pathology reports in PDF format within 90 days from the diagnosis date, i.e. the first sample collected. We divided the data into three sets: training (2014-2020, 80%), validation (2014-2020, 20%), and test (2021-2022) sets. Our training data consisted of free-text pathological reports and their diagnosis codes retrieved from the cancer registry database. In case of multiple diagnoses in a report, the report was duplicated by medical coders. In some cases, coding may also have depended on previous tumours of the patient; thus, our training data contained significant among of noise. Therefore, noise robust loss functions [8] were required.

Text preprocessing included basic NLP methods such as lemmatization, removing punctuation, numbers, and accents. This was done by using spaCY library [9]. Known medical abbreviations were expanded and person and organisation names removed by using named entity recognition and document metadata. Words in pre-processed text were transformed to word embeddings by using Word2vec method [10].

Our method was based on convolutional neural networks (CNN). The main requirements were that it could work with multiple diagnoses in a single PDF report, and previously coded cases in the cancer registry database could be used as training data.
A CNN model can return multiple prediction or codes, but the difficulty is to know whether the codes indicate

multiple tumours or alternative predictions: since we did not know how many diagnoses there were, we had to assume that a high enough probability (> 0.5 in a well calibrated model) indicated a correct code for each tumour. Unfortunately, deep neural network models such as CNN are often poorly calibrated [11]. An additional challenge was that the same code could have assigned multiple times to the same multi-tumour report, but a normal CNN model does not support this. Therefore, we developed and tested the following methods:

- M1: Detecting cancer types using a different binary model for each main cancer type.
- M2: Predicting cancer sub sites, morphology types, and lateralities by a multi-task/label CNN trained for pre-coding.
- M3: Using an additional attention layer to detect key terms in which the prediction is based and input these terms to a pre-coding CNN model.

Since each binary model (M1) detected one cancer type, both types (SCC and CMM) required their own models. This gave us an accurate method (usually over 95% accuracy) to detect different cancer types, but not multiple occurrences of the same type in one report.

Our attention layer used in M3 (Fig. 1) was able to learn to detect key diagnostic terms in texts without a need to have these terms labelled in the training data. In our CNN architecture, an attention layer followed each convolutional layer (one attention layer for each code or code group). Each attention layer contained a dense layer having a sigmoid activation function. It computed a single weight for each filter window defined by the convolutional layer. In our case, this corresponded 3-6 words in the input data. A dot product layer was used to multiply the output of the convolutional layer with the attention layer weights. Batch normalisation layers were applied after each convolutional layer and after the dot product layer. During the training the attention layer learned weights for each convolutional filter window containing 3 (sub-site) to 5 (morphology) input terms. The filters were sliding by one-word steps over the input text. In this way, each filter size long piece of text was evaluated separately and there were as many filter windows as words in the input text. This approach frequently produced overlapping key terms. They could be merged but this could have caused losing some diagnoses if there were not enough other words between them in the text. Therefore, we used key term merging only for the multi-tumour report detection and not for report pre-coding.

Our training samples were labelled using international ICD-O codes. The key terms were not included in labels. In the prediction step, we gave a pre-processed pathology reports as input data to the model. The model predicted ICD-O codes and laterality. The key terms were extracted by retrieving the activation weights from the reshape layer located after the dense layer of the attention layer. In our Keras implementation, we created a sub model that outputs these weights, so in this sense the key term model (M3) was a separate model but trained as a part of the full model (M2). An example of extracted terms, weights, and predicted ICD-O codes with their confidence probabilities is given in Fig. 2.

Key terms were detected separately for tumour sub sites and morphologies. Relevant key terms were selected based on their weights estimated by the attention layer. These terms were given as input to the pre-coding CNN to predict sub site, laterality and morphology codes for each identified term (3-6 words). Prediction confidence probabilities returned by the CNN model were used to select the most likely codes. Finally, we could apply simple heuristics based on coding rules to decide the final coding (e.g. only one SCC case but all melanoma cases to be coded and recorded).

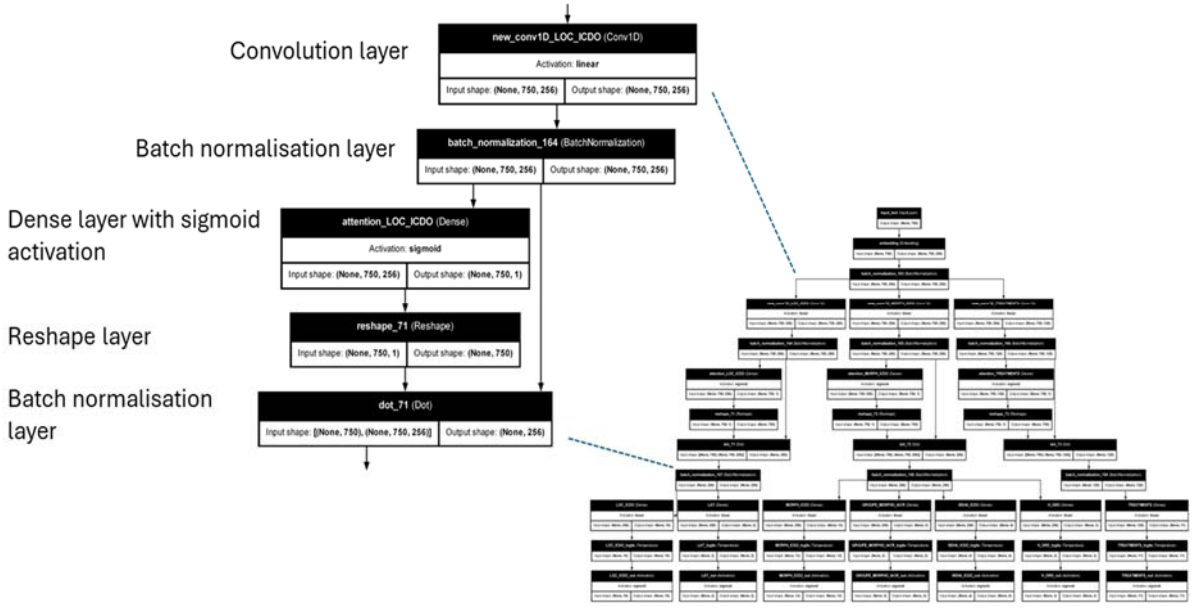


Fig. 1. CNN model and its attention layer.

Word index	Weight	Detected sub site description	Detected ICDO Site	Probability	Detected morphology description	Predicted ICDO-Morpho	Probability
37	0.998	shave localisation **omoplate gauche**	**C44.5**	[0.91]		8090	[0.34]
143	0.892	prelevement droite avec mon	C44.3	[0.48]		8090	[0.34]
31	0.517		C44.9	[0.48]	votre ref mode prelevement exc	8070	[0.35]
39	0.886		C44.9	[0.48]	omoplate gauche droite clinique	8090	[0.32]
84	0.932		C44.9	[0.48]	retrouver un melanocyte sans co	8090	[0.37]
102	0.998		C44.9	[0.48]	**melanome** extension superficiel indic breslow	**8743**	[0.81]
131	0.995		C44.9	[0.48]	**neurofibrome** stricto sensu il agir shav millimetre siege	**9540**	[0.83]

Fig. 2. Example of detection results. The highest probability values match with values in test data set.

3. Results

We first compared the results against our gold standard containing 1352 tumours, 2170 documents but only 27 (M1) or 49 (M2-3) relevant multi-tumour cases. The number was low since only the first SCC case of each patient has been recorded and many reports contained multiple SCC diagnoses. Therefore, not all multi-tumour cases were labelled in the gold standard. For all models, detection performance for multi-tumour reports strongly depended on used parameters (Fig. 3). We selected the threshold by maximising Youden's index (sensitivity + specificity - 1). M1's performance (only tumours representing different morphology groups) was: sensitivity 0.89, specificity 0.99, and PPV 0.53; M2's: 0.94, 0.66, and 0.04; and M3's 0.78, 0.93, and 0.18 respectively (Table 1). PPVs reminding low probably indicated that our gold standard did not contain all multi-tumour cases as explained above.

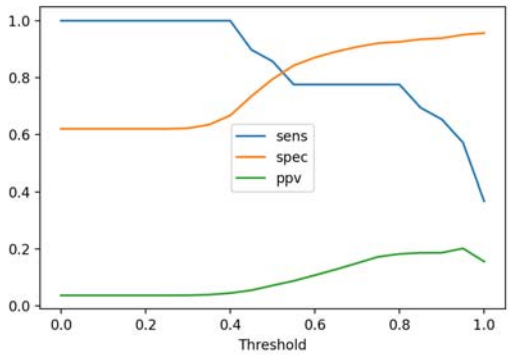


Fig. 3. Sensitivity, specificity, and PPV in different thresholds for M3.

Table 1. Multi-tumour report detection performance.

	M1	M2	M3
n	2170	2170	2170
Pos	27	49	49
Sens	0.889	0.939	0.776
Spec	0.990	0.663	0.926
PPV	0.533	0.041	0.182

When comparing against only positive cases in our gold standard, M1 detected 92% of reports correctly as multi-tumour cases (only cases representing two different morphology groups), M2 100% (predicted 2 values at least one of the three fields), and M3 95%.

Since M2 was not able to detect the order of codes, the corresponding performances are computed without considering the order. A threshold 0.5 was used. The accuracies were: 48% (sub site), 71% (morphology group), and 82% (laterality). M2 coded all reports correctly in 29% of cases and 43% if no laterality was considered.

M3 correctly coded sub sites in 62-76% (total match – subset match), morphology groups in 81-100%, and laterality in 71-76% of reports. Both sub site and morphology group were correct in 43-67% and all three (sub site, morphology group and laterality) in 38-52% of reports. By the subset match we mean that possible additional tumours existing in the reports and detected by the model were ignored in comparison. The threshold values were selected by using Youden's index. Results are illustrated in Fig. 4.

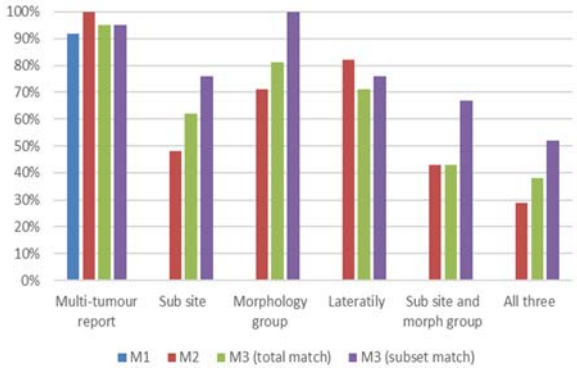


Fig. 4. Results of 3 methods (M1 is only for multi-tumour detection).

4. Discussion and Conclusions

We implemented a three-step method for detecting multiple skin cancers in free-text pathology reports. The model aims at detecting all positive diagnoses independently of the type or structure of the report. The model does not require multi-labelled or cleaned

training data. Due to the 'intelligent' attention layer, the training is purely based on labelled reports and no key terms are required in the training data.

The method indicated good performance when comparing against previously coded cases in the cancer registry database. However multiple tumours of different types in the same sub site and the same type of tumours in different sub sites are still challenging. Confusions in detecting the correct order of diagnoses reduced performance, especially in the case of the laterality. On the other hand, some documents in our gold standard contained references to previous tumours and thus not necessarily clear multi-tumour documents themselves.

Evaluating the performance of the methods was challenging because of a small number of coded multiple tumours in the gold standard. This is mostly because of coding rules that allow coding only the first incidence of SCC. Therefore, the report can have multiple invasive tumours but only one of them is linked to the gold standard. Errors in key term detection and predicted codes were usually due to incorrect text extraction and pre-procession. The used NLP method is not deterministic and can, for example, sometimes produce different lemmatised versions for the same input word. Additionally, final letters such as 's' or 'e' in the text were sometimes incorrect removed during the NLP process. The structure of the document can also be too complex, especially if containing internal references inside the document.

There are always multiple alternative methods suitable in the same data science problem. We chose the CNN based approach since our use case can be seen as a pattern matching problem. CNN is also computationally efficient and thus possible to run on an office PC. As a pattern matching method, CNN works coherently: the same expression in the input text gives the same output independently of the large context. This is a desired behaviour since different diagnoses in a report are usually independent. In this sense, a CNN-based model could outperform new transformer-based models.

The Hierarchical Attention Networks (HAN) approach could offer additional benefits [5]. HAN utilises a hierarchical structure of documents such as lines and words. A challenge would be to identify the hierarchy of pathology reports correctly. There are also recent developments in other alternative methods such as Large Language Models (LLM). However, these are often computationally heavier and data protection rules limited us from using external resources. Additionally, in existing studies LLM models have shown insufficient performance in medical coding [12, 13].

Although the method itself is general, we have tested it only with a small set of skin cancer pathology reports from French-speaking Switzerland. We also limited the method to cases in which the multiple tumours in the same report had different morphology codes (but the morphology groups could have been the same). Although this was not frequent, it could have caused losing a small number of multi-tumour cases.

This study is still work-in-progress. If the method will be found useful in pilot tests, it will be later integrated in the automated document management system of the Vaud Cancer Registry.

References

[1]. W. Hu, L. Fang, R. Ni, H. Zhang, G. Pan, Changing trends in the disease burden of non-melanoma skin cancer globally from 1990 to 2019 and its predicted level in 25 years, *BMC Cancer,* 22, 1, 2022, 836.

[2]. International classification of diseases for oncology (ICD-O), Third Edition ed., *World Health Organization,* 2013.

[3]. S. Gao, M. T. Young, J. X. Qiu, H.-J. Yoon, J. B. Christian, P. A. Fearn, G. D. Tourassi, A. Ramanthan, Hierarchical attention networks for information extraction from cancer pathology reports, *Journal of the American Medical Informatics Association*, 25, 3, 2018, pp. 321-330.

[4]. M. Alawad, S. Gao, J. X. Qiu, H. J. Yoon, J. Blair Christian, L. Penberthy, B. Mumphrey, X.-C. Wu, L. Coyle, G. Tourassi, Automatic extraction of cancer registry reportable information from free-text pathology reports using multitask convolutional neural networks, *Journal of the American Medical Informatics Association,* 27, 1, 2019, pp. 89-98.

[5]. A. Rios, E. B. Durbin, I. Hands, R. Kavuluru, Assigning ICD-O-3 Codes to Pathology Reports using Neural Multi-Task Training with Hierarchical Regularization, *ACM BCB,* 2021, 2021, 32.

[6]. B. Shin, F.H. Chokshi, T. Lee, J.D. Choi, Classification of radiology reports using neural attention models, in *Proceedings of the International Joint Conference on Neural Networks (IJCNN' 2017),* 2017, pp. 4363 - 4370.

[7]. T. Niemi, G. Defossez, S. Germann, J.-L. Bulliard, Pre-coding skin cancer from free-text pathology reports using noise-robust neural networks, Submitted for peer-review, 2025.

[8]. Y. Wang, X. Ma, Z. Chen, Y. Luo, J. Yi, J. Bailey, Symmetric cross entropy for robust learning with noisy labels, in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 322-330.

[9]. M. Honnibal, I. Montani, S. Van Landeghem, A. Boyd, SpaCy: Industrial Strength Natural Language Processing in Python, 2020.

[10]. T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, arXiv preprint *arXiv:1301.3781*, 2013.

[11]. S. A. Balanya, J. Maroñas, D. Ramos, Adaptive temperature scaling for Robust calibration of deep neural networks, *Neural Computing and Applications,* 36, 14, 2024, pp. 8073-8095.

[12]. A. Soroush, B.S. Glicksberg, E. Zimlichman, Y. Barash, R. Freeman, A.W. Charney, G.N. Nadkarni, E. Klang, Large Language Models Are Poor Medical Coders — Benchmarking of Medical Code Querying, *NEJM AI,* 1, 5, 2024.

[13]. A. Simmons, K. Takkavatakarn, M. McDougal, B. Dilcher, J. Pincavitch, L. Meadows, J. Kauffman, E. Klang, R. Wig, G. Smith, A. Soroush, R. Freeman, D.J. Apakama, A.W. Charney, R. Kohli-Seth, G.N. Nadkarni, A. Sakhuja, Benchmarking Large Language Models for Extraction of International Classification of Diseases Codes from Clinical Documentation, *Cold Spring Harbor Laboratory*, 2024.

(027)

Deep Learning Explainability on Non-Hodgkin Lymphoma: Relapse & Treatment

María L. Reyna Cruz [1], Martine Ceberio [1], Christoph Lauter [1] and Jesus Manuel López Valles [2]
[1] University of Texas at El Paso, CR2G Laboratory, 500 W. University Avenue.,
79968 El Paso Tx., United States of America
[2] Instituto Mexicano del Seguro Social, 8560 Av. Vicente Guerrero, Las Quintas,
31240 Ciudad Juarez, Chihuahua, México
Tel.: + 001 152669878
E-mail: mlreynacruz@miners.utep.com

Summary: Non-Hodgkin Lymphoma, a type of blood cancer, poses significant diagnosis and treatment challenges that could benefit from machine learning (ML) approaches. Accurate and timely diagnosis is crucial for effective treatment of Non-Hodgkin Lymphoma. In cases of relapse, personalized treatment becomes even more essential. Wellestablished guidelines exist for first-time diagnoses of Non-Hodgkin Lymphoma. However, when a patient experiences a relapse, selecting the best treatment can be time-consuming due to the numerous available options. This study focuses on uncertainty quantification and enhancing the explainability of ML models and ensuring their clinical integration to improve Non-Hodgkin lymphoma treatment when a relapse occurs. We emphasize the use of explainable AI techniques, such as SHAP and LIME, to improve the interpretability of the ML model's decisions. This ensures that clinicians can better understand and trust the model's predictions. Additionally, we analyze different types of uncertainty (epistemic, aleatoric, and distributional) and discuss how they affect decision-making. Our goal is not only to develop an ML-based treatment recommendation tool but to ensure that its predictions are reliable, interpretable, and ethically sound before clinical deployment.

Keywords: Non-hodgkin limphoma, Relapse, Uncertainty quantification, Explainability, Deep learning.

1. Introduction

Non-Hodkin Lymphoma is a type of blood cancer that poses significant challenge in diagnosis and treatment. It is the most common type of lymphoma, with 544,352 new cases reported globally in 2020, making it the third most common cancer among individuals aged 0-19 years. [1] This type of lymphoma has a variability in its clinical behaviour. being the most common fever, heavy sweating (diaphoresis), weight loss, adenopathy (progresive growth, non-painful), splenomegaly, and hepathomegaly. [2]

The survival rates range from 79% for one year, 63% for five years, to 51% for ten years [2]. This highlights the critical importance of early and accurate treatment for Non-Hodgkin lymphoma, as timely intervention can significantly improve survival rates. However, treatment becomes more complex in cases of relapse.

The disease, after being treated successfully once, may return, requiring a reassessment of treatment strategies. With numerous treatment options available, each patient's case presents a unique challenge. Currently, in the guidelines there are a lot of medication options to choose as treatment for when a patient has a relapse of Non-Hodgkin lymphoma; from those the physician must choose one for each patient. A personalized treatment plan addresses the specific needs of each patient. Identifying the most effective treatment plan for a patient who has experienced relapse is a time-consuming and difficult process. This process is subjective and may vary depending on the physician's experience and interpretation of clinical data.

In the past years machine learning (ML) has been proven a useful aid for several diseases, talking about diagnose and treatment; models can process a vast amount of data and identify patterns not immediately detected by the human clinitian. Being it so helpful, a question arises: Why has it not been more widely adopted? The answer being clear, Uncertainty and explainability. The way ML models work are considered as a" black box" making predictions without a clear insight as to how the conclusion were reached. In the context of health and life-death decisions, such as cancer, this lack of transparency is a barrier for the daily clinical practice.

The problem of having a black-box is serious when we talk about relapsed Non-Hodgkin Lymphoma, where the treatment is not an established algorithm and requires thoughtful consideration of multiple factors. Making a model that could provide clear, interpretable, and explainable recommendations, while incorporating datasets of prior cases would not only help for the treatment selection process but also offer insights that may not appear as obvious to an experienced physician.

By integrating machine learning with data from patients treated for Non-Hodgkin Lymphoma relapse, this study aims to improve the accuracy and reliability of treatment recommendations by integrating explainable AI techniques such as SHAP (SHapley Additive exPlanations), and LIME (Local

Interpretable Modelagnostic Explanations). This work explores explainability techniques and Uncertainty Quantification methods to improve the transparency and robustness of ML-driven lymphoma treatment recommendations.

The implementation of an explainable ML-based treatment tool will significantly improve the accuracy, reliability, and clinical acceptance of Non-Hodgkin lymphoma treatment aiding tool compared to traditional methods that involve looking though all guidelines and medications [3]. Offering quicker, more reliable, and personalized treatment recommendations focusing on interpretability, we can ensure that the ML models developed in this study are practical and trustworthy for clinical use.

2. Background

Lymphoma comprises a group of malignant neoplasms of lymphocytes, that concerns lymphatic tissue, bone marrow, or extranodal sites [4]. It is categorized mainly into two types: Hodgkin lymphoma (HL) and Non-Hodgkin lymphoma (NHL) [5]. Non-hodgkin Lymphoma is the most common making the 90 % of all lymphomas [3].

The treatment of Non-Hodgkin Lymphoma is so far well established by guidelines, the problem arises when the patient gets diagnosed again after a period of being free (remission) from the disease. Relapse of Non-Hodgkin Lymphoma, presents additional complexities for treatment. The relapse rates for NonHodgkin Lymphoma varies depending the subtype; the disease's behaviour after relapse is often more aggressive and resistant to initial treatments. Patients who relapse require reassessment and modification of the treatment strategy, needing personalized therapies to target the progression.

Treatment guides provide several options. The treatment strategies may vary based on the type of Non-Hodgkin Lymphoma (e.g., diffuse large B-cell lymphoma, follicular lymphoma) and specific characteristics of the relapse. Guidelines recommend high-dose chemotherapy, followed by autologous stem cell transplant (ASCT) used often as standard care for relapsed Diffuse Large B Cell Lymphoma and other aggressive Non-Hodgkin Lymphomas. The success of ASCT depends on factors such as the patient's response to prior therapies and the length of remission period before relapse. [6]

However not all patients are candidates for ASCT, where alternative therapies are implemented. Treatments such as: Monoclonal antibodies such as rituximab can be reintroduced for patients who previously responded well. Targeted therapies including Bruton's Tyrosine Kinase (BTK) inhibitors, and Phosphoinositide (PI3K) inhibitors are considered effective options for patients with certain genetic profiles not candidates for ASCT. Chimeric Antigen Receptor (CAR) T-cell therapy is mentioned as a potential option for patients who has exhausted other therapies and had failed 2 or more lines of therapy. The group of immunomodulatory drugs (e.g., lenalidomide) are also implemented when traditional therapies are ineffective. [6]

The choice of therapy depends on factors like response of patient's initial treatment, overall health, comorbidities, and specific molecular and genetic features of the lymphoma. These numerous treatment options, while offering a broad spectrum of choices, create a challenge for the physician, as selecting the optimal regimen requires comprehensive analysis of clinical data, patient history, and the characteristics of the disease.

The role of machine learning in this context becomes useful in assisting clinicians by providing data-driven recommendations. Machine Learning models can analyze vast datasets from previous cases, evaluating the success rates of various treatments in similar relapse scenarios. Adding explainability techniques and being able to quantify the uncertainty can enhance this process making the decision process transparent, helping clinicians to understand why specific treatments are being suggested and thus inserting confidence in the recommendations; reducing the time-consuming process and subjectivity in clinical decision making.

3. Methodology

Recent advancements in machine learning, particularly deep learning, have shown promise in automating and improving the accuracy of medical diagnoses and treatments. Convolutional neural networks (CNNs) have been particularly effective in analysis tasks [7], making them suitable for treatment of Non-Hodgkin Lymphoma. Despite the advancements, a significant challenge remains in making these ML models interpretable and trustworthy for clinical use.

3.1. Data Collection

To develop an accurate predictive model for Non-Hodgkin Lymphoma (NHL), we will use clinical data from patients, including survival data at 5 and 10 years. Key features include age, gender, lymphoma type, tumor size and location, laboratory results, treatment responses, and genetic mutations. These features will serve as the foundation for model development and validation. We are currently in the process of obtaining approval from the Institutional Review Board (IRB) of the Mexican institution providing the data. This step ensures compliance with ethical standards for patient data usage and guarantees that all necessary de-identification and privacy measures are in place.

To ensure model generalizability, we will employ stratified sampling to maintain a representative distribution of key variables such as lymphoma subtype, disease stage, and treatment response. Given the potential for class imbalance—particularly in outcomes such as relapse or specific genetic

mutations—we will address this through oversampling of minority classes, weighting adjustments, or synthetic data generation techniques where necessary.

3.2. Explainability Techniques: [8], [9] and [10]

By integrating Explainable AI techniques into the ML models for lymphoma treatment, clinicians can better understand and trust the automated therapeutic outcomes.

SHAP (SHapley Additive exPlanations): Compute the contribution of each feature to the prediction to understand feature importance.

LIME (Local Interpretable Model-agnostic Explanations): Approximate the model locally with an interpretable model to provide insights into individual predictions.

Attention Mechanisms: Implement attention mechanisms to highlight which parts of the input data the model focuses on during prediction.

Quantify Error: Measure and quantify the errors in predictions to understand where the model might be making incorrect decisions and how significant these errors are.

Statistical Properties: Analyze the statistical properties of the features and predictions to ensure the model's outputs are consistent and reliable.

3.3. Uncertainty Quantification

Machine Learning models must not only be accurate but also provide a measure of confidence in their predictions. We categorize uncertainty into three types: [11]

Epistemic (Model) Uncertainty: Arises from limitations in the model or lack of data; can be reduced by improving training data.

Bayesian Neural Networks and Monte Carlo Dropout can help quantify epistemic uncertainties, ensuring that ML-based recommendations reflect a level of confidence that clinicians can consider when making treatment decisions. [12]

In Bayesian Neural Networks, model weights are treated as probability distributions rather than fixed values [13], which allows us to quantify the uncertainty in the model's predictions. This is useful for predicting the likelihood of Non-Hodgkin Lymphoma treatment success, where understanding the level of certainty can inform treatment decisions. For example, if the model predicts that a certain treatment has a 70% chance of success, a Bayesian approach might reveal that the true confidence interval for this prediction could be anywhere between 60% and 80%, translated from a mean of 70% and variance of 10%. This allows the model to express the uncertainty that arises from the fact that there is not a single" correct" set of weights but instead a range of possible sets of weights that could explain the data. By sampling different weight configurations, we can calculate the spread of predictions and understand how uncertain the model is about a given prediction.

Monte Carlo Dropout involves applying dropout not only during training/inference but also during training. This allows us to run the network multiple times with different random dropout configurations, producing a distribution of predictions. The variance in these predictions helps us quantify the uncertainty associated with each output. The goal is to capture how confident or uncertain the model is about its predictions, especially when the input is uncertain or ambiguous and the model's behavior is not deterministic.

Aleatoric (Data) Uncertainty: Stems from inherent noise in medical data and cannot be eliminated, only accounted for.

Distributional Uncertainty: Occurs when new patient data deviates from the training distribution. In other words, if the model was trained using data from a specific group of patients, but the new patient data comes from a different group (e.g., with different age, gender, or health conditions), the model might not perform as well because it hasn't seen this kind of data before. This can lead to risks in real-world applications, where the model might not be able to make accurate predictions for patients that fall outside the patterns it learned during training.

Distributional uncertainty can be reduced by having a more diverse and representative dataset. If the model is trained on a broad range of data that covers various patient characteristics (such as age, gender, medical history, etc.), it will be better equipped to handle new, unseen cases that are closer to the real world variation, which it was addressed at the Data collection section.

3.4. Model Training

- Data Splitting: The model will be trained using a real world clinical dataset, which will be divided into training, validation, and test sets, to ensure unbiased evaluation and prevent overfitting.
- Training: The model needs to be trained to predict outcomes based on patient characteristics, with survival rates (e.g., 5- and 10-year survival) serving as the target variables. Implement cross-validation to fine-tune hyperparameters and enhance model robustness.
- Validation: Model performance will be assessed using the validation set, where ground truth survival outcomes are available. Metrics such as accuracy, AUC-ROC, precision, and recall will be used to monitor performance and refine the model.
- Inference: Once trained, the model will be used to generate predictions for new patients who have not yet undergone treatment. Since ground truth outcomes are unknown for these patients, model predictions will be clinically validated by comparing them to expert judgment and established treatment guidelines. Clinicians can flag cases where the model's predicted treatment does not align with their expertise or known best practices. These flagged cases can be analyzed to identify potential biases or gaps in the training data.

• Components of the output:

Predicted Treatment Pathway: The model will suggest potential treatment strategies based on patient history and clinical data.

Explainability Insights: Feature importance will be visualized using SHAP values, with LIME explanations providing interpretability at the individual prediction level.

3.5. Ethical Considerations

Ethical concerns include patient privacy, the potential risks of over-reliance on automated decision-making, and the need for transparency in the model's decision-making process.

Ensuring patient privacy and confidentiality will be a priority, as the model will be built using sensitive clinical data. Data anonymization and compliance with data protection regulations, such as HIPAA, will be implemented.

The risk of over-reliance on automated decision-making poses significant risks, as incorrect or suboptimal treatment recommendations may arise if the model fails to capture complex clinical variations.

To mitigate this, the integration of uncertainty-aware machine learning systems, which quantify uncertainty in model outputs, will allow clinicians to assess the reliability of predictions before making decisions.

We will explore clinician-in-the-loop approaches to ensure human oversight, where clinicians review the model's predictions and integrate their expertise, thus ensuring that automated recommendations align with clinical judgment.

This approach will mitigate ethical and medical risks, as it provides transparency and insight into the model's limitations. Ultimately, ensuring that predictions are ethically sound and clinically safe is a crucial step before any deployment.

3. Results

As the goal of this research is not to immediately deploy a generic machine learning (ML) model but to first prioritize ensuring the quality of the dataset and reducing uncertainty before clinical application, we do not present model results at this stage. Instead, we reference studies that have successfully implemented explainability techniques and uncertainty quantification in various domains. These studies demonstrate the effectiveness of these techniques in improving model reliability, interpretability, and clinical applicability, providing evidence that such techniques can enhance decision-making in complex, high-stakes environments, such as lymphoma treatment, even before model deployment.

• SHAP-Based Feature Attribution in Predictive Process Monitoring [14]

A study published in Annals of Operations Research applied SHAP values to analyze prediction uncertainty in industrial process monitoring. By decomposing model decisions into interpretable feature attributions, they identified high-risk scenarios where the model was overconfident. Their findings indicated that integrating SHAP-based explainability reduced error rates and increased human trust in automated systems. We expect that SHAP values will enhance lymphoma treatment model interpretability by highlighting the contribution of clinical variables (e.g., age, tumor markers, genetic profiles) to treatment recommendations.

• Explainability and Uncertainty in Educational Data Mining. [15]

Thuy and Benoit (2024) integrated uncertainty estimation into neural networks for predicting student performance in educational data mining. Their approach involved a classification-with-rejection mechanism, which flagged uncertain cases for expert review. They found that incorporating uncertainty awareness improved model robustness, particularly under distribution shifts where unseen test samples deviated from training data. The study demonstrated that models with uncertainty quantification significantly reduced incorrect high-confidence predictions. This reinforces the importance of uncertainty-aware decision-making in critical applications. This result supports our approach by showing that uncertainty estimation can prevent misleadingly confident ML-based recommendations in lymphoma treatment, ensuring that ambiguous cases are referred for expert validation rather than blindly trusted.

• Uncertainty-Aware Image Classification Using Bayesian Neural Networks [16]

A study in Knowledge-Based Systems implemented Bayesian neural networks and entropy-based rejection mechanisms for image classification tasks. They found that using Monte Carlo Dropout significantly improved the model's ability to express uncertainty compared to deterministic deep learning models. This is particularly relevant for our research, as lymphoma diagnosis and treatment planning involve similar risks of misclassification. By incorporating Bayesian techniques, we can ensure that our model does not provide overconfident yet incorrect treatment predictions, thereby increasing the reliability of ML-assisted decision-making in oncology.

5. Conclusion

Early and accurate diagnosis of lymphoma is critical for initiating timely and effective treatment. Traditional diagnostic approaches at relapse often rely on subjective assessments, which can lead to delays and suboptimal outcomes. Machine learning based tools offer the potential to provide consistent, non-invasive, and efficient diagnostic support, enabling more precise and timely decision-making, ultimately improving patient outcomes while reducing healthcare costs.

However, to realize the full potential of these tools, it is essential to address the inherent complexities and

limitations of ML models. Explainability techniques and uncertainty quantification play a vital role in ensuring that clinicians can confidently rely on these tools. By providing clear insights into the factors influencing model predictions and quantifying the uncertainty associated with those predictions, these techniques foster trust and promote the integration of ML-based systems into clinical practice.

This work emphasizes the significance of not only achieving high accuracy in ML models but also prioritizing explainability and uncertainty quantification to enhance clinician trust and facilitate the responsible adoption of AI-driven recommendations in healthcare. It is through this combination of model transparency and uncertainty awareness that we can improve both the clinical applicability and reliability of ML tools in lymphoma diagnosis and treatment, paving the way for more effective, patient-centered care.

References

[1]. Cuenta de Alto Costo, Día mundial del linfoma 2023, 2023. [Online]. https://cuentadealtocosto.org/cancer/dia-mundial-del-linfoma-2023/

[2]. Guía de Practica Clíinica, Linfomas No Hodgkin en el Adulto, *Instituto Mexicano del Seguro Social,* Mexico, 2009.

[3]. K. R. Shankland, J. O. Armitage, and B. W. Hancock, Non-hodgkin lymphoma, *The Lancet*, Vol. 380, No. 9844, 2012, pp. 848–857.

[4]. W. D. Lewis, S. Lilly, and K. L. Jones, Lymphoma: diagnosis and treatment, *American Family Physician*, Vol. 101, No. 1, 2020, pp. 34–41.

[5]. L. de Leval and E. S. Jaffe, Lymphoma classification, *The Cancer Journal*, Vol. 26, No. 3, 2020, pp. 176–185.

[6]. N. C. C. N. (NCCN), B-Cell Lymphomas, version 3.2024 ed., *National Comprehensive Cancer Network, Inc.,* 2024, [Online]. https://www.nccn.org

[7]. L. V. Haar, T. Elvira, and O. Ochoa, An analysis of explainability methods for convolutional neural networks, *Engineering Applications of Artificial Intelligence*, Vol. 117, 2023, p. 105606.

[8]. A. Bhattacharya, Applied Machine Learning Explainability Techniques: Make ML models explainable and trustworthy for practical applications using LIME, SHAP, and more, *Packt Publishing Ltd*, 2022.

[9]. F. Dallanoce, Explainable ai: A comprehensive review of the main methods, *MLearning. ai*, 2022.

[10]. D. Gopinath, Picking an explainability technique, Oct 2021. [Online]. https://towardsdatascience.com/picking-an-explainability-technique-48e807d687b9

[11]. J. Gawlikowski, C. R. N. Tassi, M. Ali, J. Lee, M. Humt, J. Feng, A. Kruspe, R. Triebel, P. Jung, R. Roscher et al., A survey of uncertainty in deep neural networks, *Artificial Intelligence Review*, Vol. 56, No. Suppl 1, 2023, pp. 1513–1589.

[12]. P. Wang, N. C. Bouaynaya, L. Mihaylova, J. Wang, Q. Zhang, and R. He, Bayesian neural networks uncertainty quantification with cubature rules, in *Proceedings of the IEEE International Joint Conference on Neural Networks (IJCNN' 2020)*, 2020, pp. 1–7.

[13]. Y. Gal and Z. Ghahramani, Dropout as a Bayesian approximation: Representing model uncertainty in deep learning, in *Proceedings of the International Conference on Machine Learning* (*PMLR)*, 2016, pp. 1050–1059.

[14]. N. Mehdiyev, M. Majlatow, and P. Fettke, Quantifying and explaining machine learning uncertainty in predictive process monitoring: an operations research perspective, *Annals of Operations Research*, 2024, pp. 1–40.

[15]. A. Thuy and D. F. Benoit, Explainability through uncertainty: Trustworthy decision-making with neural networks, *European Journal of Operational Research*, Vol. 317, No. 2, 2024, pp. 330–340.

[16]. X. Zhang, F. T. Chan, and S. Mahadevan, Explainable machine learning in image classification models: An uncertainty quantification perspective, *Know.-Based Syst*., Vol. 243, 2022, 108418.

(029)

Risk Stratification and Early Diagnosis of Heart Failure

Borut Flis [1], Petar Vračar [1], Matej Pičulin [1], Djordje Jakovljević [2], Nenad Filipović [3] and Zoran Bosnić [1]
[1] University of Ljubljana, Faculty of Computer and Inf. Science, Večna pot 113, 1000 Ljubljana, Slovenia
[2] Clinical Sciences and Translational Medicine, RC for Health and Life Sciences, Coventry University, UK
[3] Bioengineering Research and Development Center, BioIRC, Kragujevac, Serbia
E-mail: zoran.bosnic@fri.uni-lj.si

Summary: Heart failure (HF) affects over 64.3 million people worldwide. As a part of the StratifyHF project, we developed a decision support system (DSS) to enhance HF prediction and diagnosis through machine learning (ML) approaches. The DSS comprises two modules: Early diagnosis and Risk stratification module; both are critical to improving outcomes yet remain underutilized in clinical practice. The Early Diagnosis Module attempts to identify HF before diagnostic completion, prioritizing physical examination, symptoms, blood biomarkers, and patient history while excluding post-diagnosis attributes. Multiple ML models were trained using 10-fold cross-validation, achieving promising results despite challenges posed by incomplete data. The Risk Stratification Module focuses on predicting HF risk without prior diagnoses. Using XGBoost and Random Forest models, we achieved accuracy of 0.895, sensitivity of 0.984 and an F1 score of 0.937. These results demonstrate the feasibility of integrating ML-based predictive models into clinical workflows, offering significant potential to improve early HF diagnosis and risk management.

Keywords: Heart failure, Decision support system, Risk stratification, AI in medicine, Early diagnosis.

1. Introduction

Heart failure (HF) is a complex syndrome of impaired heart function, affecting at least 64.3 million people globally and 15 million in Europe [1]. Its prevalence is ~2% in the general population and >10% in those over 70, with higher rates in women after 79 years. HF incidence has risen due to prolonged survival and aging populations, increasing diagnosed cases by 12% over the past decade [2]. Accurate risk prediction, diagnosis, and disease progression assessment are vital for implementing effective prevention and treatment strategies to reduce HF morbidity, mortality, and healthcare burden. Evidence shows that HF is often underdiagnosed in primary care [3]. General practitioners play a key role in early diagnosis and referrals, but clinical risk stratification models, recommended in guidelines, remain underutilized [4]. We developed a decision support system (DSS) as a part of our project StratifyHF. The DSS comprises two modules: **Early Diagnosis** and **Risk Stratification**.

2. DSS Modules

2.1. Risk Stratification

The **Risk Stratification** module categorizes patients into distinct risk classes to predict the likelihood of heart failure at the next recorded time point in a patient's medical history. The dataset consists of 2,331 cases, with irregular such time intervals between measurements, ranging from months to several years or even a decade. Various imputation methods, including KNN, median, and missing value flags, were tested to address missing data, with binary attributes indicating the presence of data for specific patients.

The module's design incorporates feedback from clinical partners and emphasizes iterative refinement and the integration of additional data sources, such as virtual patients, to improve model accuracy and reliability.

2.2. Early Diagnosis

The **Early Diagnosis** module aims to identify potential **HF** diagnoses based on a patient's current clinical profile, prioritizing accurate identification before completing all diagnostic tests. Early diagnosis is often challenging for general practitioners, as the initial symptoms can be difficult to distinguish from conditions like obesity or lung disease [5]. Using a dataset of 10,314 examples, the module evaluates combinations of medical modalities – such as physical examination data, symptoms, blood biomarkers, imaging results (ECG, echocardiography, MRI, X-ray), patient history, and medications—to develop cost-effective diagnostic models. Attributes available only after diagnosis were excluded to ensure applicability to early-stage diagnosis.

This framework provides a foundation for integrating predictive models into clinical decision-support systems and evaluating their cost-effectiveness in diagnostic pathways.

3. Preliminary Experiments

3.1. Risk Stratification

The **Risk Stratification** module was evaluated using two machine learning models, XGBoost and

Random Forest, assessed with multiple metrics: sensitivity, specificity, AUC, and F1 score. Model evaluation was performed using the 10-fold cross-validation, ensuring that the same patient (if they have multiple measurements at different times) does not appear in both the training and test folds. In each iteration of cross-validation, internal cross-validation was performed on the training data for hyperparameter optimization. The most effective imputation method was KNN, enhanced with an additional indicator column to distinguish between real and imputed values.

During cross-validation, the model achieved an accuracy of 0.887, with a majority-class (positive heart failure diagnosis) accuracy of 0.817. Key performance metrics included a sensitivity of 0.974, specificity of 0.536, and F1-score of 0.932. Fig. 1 illustrates the ROC curve and AUC score.

3.2. Early Diagnosis

The **Early Diagnosis** evaluation contained experiments with decision trees, naïve Bayes, random forests, SVM, and XGBoost, using feature selection from patient modalities and 10-fold cross-validation. The default accuracy was 0.567. To handle the dataset's extensive missing values, models were selected based on their ability to manage incomplete data effectively, with feature importance analyzed using random forests.

Models using physical exams, blood markers, symptoms, and patient history were deemed promising by medical partners, while those incorporating all modalities fell short due to significant missing data in critical tests like echo and MRI. Table 1 shows the results achieved by different attribute groups.

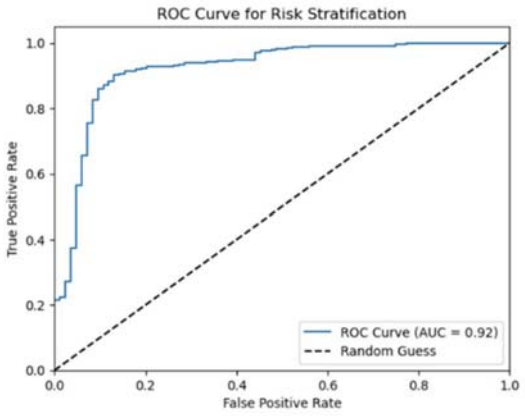


Fig. 1. ROC Curve for Risk Stratification.

Table 1. Classification report by attribute groups for early diagnosis.

Features	Accuracy	Sensitivity	Specificity	PPV
Age, Weight, Height	0.681	0.631	0.748	0.766
PHY + PAT + SYM	0.860	0.838	0.889	0.908
PHY + PAT + SYM + BLO	0.874	0.857	0.897	0.916
All (incl. Echo, ECG, X-ray, MRI)	0.900	0.894	0.908	0.927

Legend: PHY – physical examination, PAT – patient history, SYM – symptoms, BLO – blood markers; PPV – positive predicted value

4. Conclusions

The presented decision support system design demonstrates the potential of machine learning to improve heart failure (HF) diagnosis and risk stratification. By leveraging diverse clinical data and employing robust modeling techniques, the Early Diagnosis and Risk Stratification modules achieved high preliminary predictive performance, particularly in identifying patients at risk of HF and facilitating early intervention.

Despite the promising results, challenges remain in addressing missing data from critical diagnostic modalities and ensuring model generalizability across diverse patient populations. Future work will focus on refining data imputation strategies, integrating additional data sources, and evaluating the cost-effectiveness of these tools in real-world clinical settings. By addressing these challenges, the StratifyHF system could significantly enhance early diagnosis, risk prediction, and overall management of HF, reducing its global burden and improving patient outcomes.

Acknowledgments

The study received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101080905.

References

[1]. S. L. James *et al.*, Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017, *The Lancet*, Vol. 392, No. 10159, Nov. 2018, pp. 1789–1858.

[2]. N. Conrad *et al.*, Temporal trends and patterns in heart failure incidence: a population-based study of 4 million

individuals, *The Lancet*, Vol. 391, No. 10120, Feb. 2018, pp. 572–580.

[3]. E. Roberts *et al.*, The diagnostic accuracy of the natriuretic peptides in heart failure: systematic review and diagnostic meta-analysis in the acute care setting, *BMJ*, Vol. 350, Mar. 2015, p. h910.

[4]. T. A. McDonagh *et al.*, 2021 ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure: Developed by the Task Force for the diagnosis and treatment of acute and chronic heart failure of the European Society of Cardiology (ESC) With the special contribution of the Heart Failure Association (HFA) of the ESC, *Eur. Heart J.*, Vol. 42, No. 36, Sep. 2021, pp. 3599–3726.

[5]. J. Mant *et al.*, Systematic review and individual patient data meta-analysis of diagnosis of heart failure, with modelling of implications of different diagnostic strategies in primary care, *Health Technol. Assess.*, Vol. 13, No. 32, Jul. 2009, pp. 1–23.

(030)

A Novel Approach to Generating Synthetic Data with WGAN-GP and Mahalanobis Distance Filtering

Akshay Sunkara [1], Rajiv Morthala [1], Anav Jain [1], Srinjoy Ghose [1], Anirudh Nimmagadda [2] and Santosh Morthala [3]
[1] University of North Texas, 1155 Union Cir., 76201 Denton, USA
[2] University of Texas, Long School of Medicine, 7703 Floyd Curl Dr, San Antonio, TX, USA
[3] University of Texas at Dallas, 800 W Campbell Rd, Richardson, TX, USA
Tel.: + 001 (630) 400-8600
E-mail: akshaysunkara@my.unt.edu

Summary: A prominent challenge in machine learning is the scarcity of high-quality data, which inhibits model performance. To address this, we propose a synthetic data generation framework combining Wasserstein Generative Adversarial Networks with Gradient Penalty (WGAN-GP) and Mahalanobis Distance-Based Filtering. Through comprehensive experiments, we demonstrate that supplementing real training data with our filtered synthetic data improves model accuracy by up to 6% compared to models trained without synthetic data. Additionally, our results indicate that incorporating Mahalanobis Distance Filtering significantly improved accuracy compared to using raw synthetic data alone. To improve accessibility, we are developing a web-based platform that allows researchers and practitioners to generate, filter, and modify synthetic data in real time.

Keywords: Machine learning, Synthetic data generation, Generative-adversarial-networks, Wasserstein-GAN, Mahalanobis-distance-filtering.

1. Introduction

In this paper, we address the challenge of limited data in machine learning by proposing a solution for generating synthetic data to augment datasets. Insufficient data can hinder machine learning model training, leading to overfitting and poor generalization [1]. To address this issue, we propose a solution for improving neural network accuracy using synthetic data generation. We leveraged a Wasserstein Generative Adversarial Network with gradient penalty (WGAN-GP), optimizing its performance by systematically tuning hyperparameters such as learning rate, batch size, and gradient penalty coefficient. [2]. Additionally, we leveraged the Mahalanobis Distance (MD) metric to filter large volumes of synthetic data, selecting the highest-quality samples [3]. By conducting experiments to compare the performance of models trained on synthetic data filtered using the Mahalanobis Distance metric versus randomly generated synthetic data, we found that our filtering approach yields better results under certain circumstances. Our proposed filtering mechanism is designed for two applications: improving model accuracy by supplementing real datasets with synthetic data or replacing real datasets with synthetic ones. By mitigating data scarcity through synthetic data generation, our work provides a practical solution for practitioners facing data limitations.

1.1. Overview of Generative Adversarial Networks

Generative Adversarial Networks (GANs) are deep learning architectures that train two neural networks in an adversarial system to generate synthetic data from a training dataset. The generator produces synthetic data resembling the training dataset, while the discriminator classifies the produced data as real or fake. The original formulation of GANs, known as Vanilla GANs, as proposed in the original study by Goodfellow, Ian, et al, [10], represents the most fundamental version of this framework.

1.2. Limitations of Vanilla GANs

Vanilla GANs are susceptible to several challenges, particularly mode collapse and training instability. Mode collapse occurs whenever the generator repeatedly produces a small subset of outputs that the discriminator classifies as fake, reducing diversity and hindering the discriminator's ability to improve. Training instability occurs when one of the neural networks overpowers the other, preventing the GAN from producing high-quality data due to negligible feedback. To address these challenges, our study used Wasserstein GANs as they are less prone to the issues faced by Vanilla GANs. Wasserstein GANs address these issues through the Wasserstein loss function, providing continuous gradients to the generator ensuring a consistent training process. A gradient penalty can also be enforced, further stabilizing training by preventing the discriminator from becoming too dominant.

2. Methodology

The Methodology Flow Chart is shown in Fig. 1.

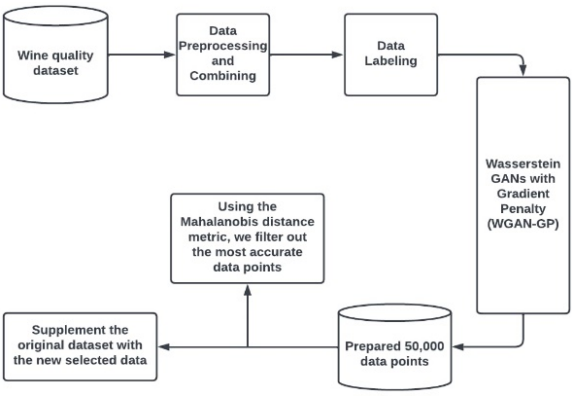


Fig. 1. Methodology Flow Chart.

2.1. Dataset & Preprocessing

We used the Wine Quality Dataset, which contains information on the properties of red and white variants of Portuguese Vinho Verde wine and their effects on its quality [7]. The dataset contains 11 features, such as fixed acidity, citric acid, and residual sugar, determined by chemical tests, along with a sensory-based quality score ranging from 0 to 10. For preprocessing, we shuffled the data and partitioned it into batches of 64 samples.

2.2. Model Architecture

Our implementation follows the Wasserstein GAN with Gradient Penalty (WGAN-GP) framework, which applies a gradient penalty instead of using weight clipping. This approach preserves model capacity and improves training stability. The architecture consists of a Generator and a Discriminator, commonly referred to as the Critic in the WGAN framework. The Critic is composed of three linear layers with decreasing dimensionality, each followed by a LeakyReLU activation with a 0.2 negative slope. It outputs a scalar value representing the Wasserstein distance between the real and generated data distributions. The Generator takes a 12-dimensional latent noise vector as input and processes it through three linear layers, each also followed by a LeakyReLU activation with a 0.2 negative slope, to produce synthetic samples matching the input dataset.

2.3. Hyperparameters

We conducted hyperparameter tuning using grid search to identify the optimal configuration for training. The learning rate was tested within the range of 0.00005 to 0.001, with a value of 0.0001 demonstrating the most stability and minimal oscillations. Batch sizes were explored between 32 and 128, with a batch size of 64 yielding the best performance. The Adam optimizer was selected for its effectiveness within the WGAN-GP framework. The gradient penalty coefficient was set to 5, as recommended in the original WGAN-GP paper. The Discriminator was updated five times per Generator update, following the same guidance.

We monitored GAN training by tracking the losses of the Generator and the Discriminator and the Wasserstein distance, ensuring both networks stabilized with minimal oscillations. Additionally, we visually confirmed the convergence of the model. After training, the Generator produced 50,000 synthetic data points (Fig. 2).

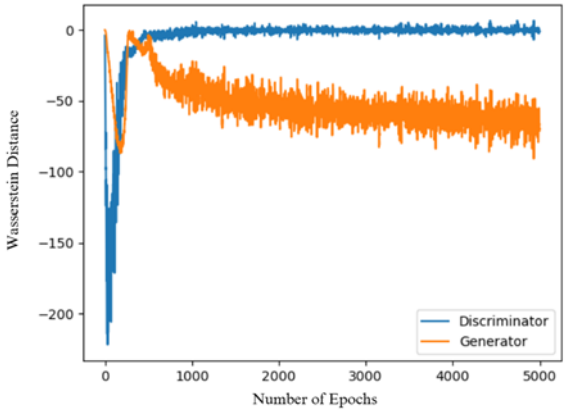


Fig. 2. Losses of WGAN-GP.

2.4. Data Filtering with MD

After the WGAN-GP produced 50,000 data points, we used the Mahalanobis Distance (MD) metric to only retain data points that closely resembled the statistical distribution of the training set. The MD metric categorizes data points by measuring each point's deviation from the training set. It accounts for correlations between variables, effectively detecting outliers in datasets. Unlike most distance metrics like Euclidean distance, the Mahalanobis distance scales each variable's contribution using the covariance matrix, making it effective for multivariable datasets used in data-intensive fields like healthcare.

Furthermore, we applied a similarity threshold based on critical values obtained from the Chi-Squared Distribution. This threshold is determined by the dataset's degrees of freedom, the number of variables, and a chosen significance level. Lowering the significance level raises the threshold, making it more strict and conservative in detecting outliers, while increasing it lowers the threshold, making it more lenient. We adjusted the significance level based on the number of data points to generate.

3. Results and Discussion

We conducted experiments using a neural network to measure accuracy, with a training-to-testing split of 70-30. The model was trained for 1000 epochs, and we

compared the accuracy of no synthetic data, with synthetic data, and with synthetic data filtered using the MD metric for outlier detection. Our experiments show that synthetic data generation can improve accuracy by up to 6%, especially with higher training epochs. However, the benefits are less consistent at lower epochs and can sometimes reduce performance. It's important to consider the parameter settings when using synthetic data (Fig. 3, Fig. 4).

Additionally, we conducted experiments to evaluate the impact of the MD metric. We tested the metric on the configuration with the highest accuracy (1000 epochs and 50 synthetic points). In this setup, the model's accuracy improved by 2.86% when trained with filtered synthetic data compared to using non-filtered synthetic data of the same quantity (Table 1).

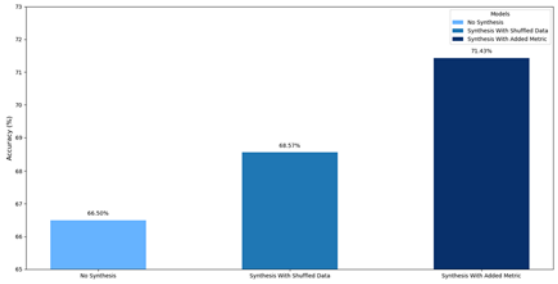


Fig. 3. Accuracy Across Different Models

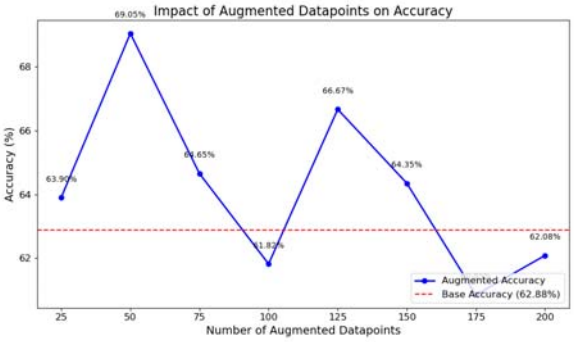


Fig. 4. Accuracy Across Augmented Datapoints

Table 1. Accuracies for Augmented Datasets.

#Non-Augmented Datapoints	#Augmented Datapoints	Base Accuracy (1000 NADP)	Augmented Accuracy
1000	25	62.88	69.90724
1000	50	62.88	69.04762
1000	75	62.88	64.65116
1000	100	62.88	61.81818
1000	125	62.88	66.66667
1000	150	62.88	64.34783
1000	175	62.88	60.81188
1000	200	62.88	62.08333
900	25	58.33	63.78378
900	50	58.33	57.36842
900	75	58.33	56.92308
900	100	58.33	62
900	125	58.33	64.30624
900	150	58.33	59.04762
900	175	58.33	65.47056
900	200	58.33	61.64384
800	25	63.12	69.09091
800	50	63.12	55.88235
800	75	63.12	62.28571
800	100	63.12	61.16667
800	125	63.12	61.62162
800	150	63.12	58.94737
800	175	63.12	57.94872
800	200	63.12	54

4. Conclusion

Our research demonstrates that supplementing datasets with synthetic data can improve model accuracy under certain conditions. However, the effectiveness of synthetic data largely depends on conditions such as the ratio of real to synthetic data and the quality of generated samples. By implementing WGAN-GP to generate data points and applying the MD metric to filter outliers, our work improves the generation of high-quality synthetic data and introduces statistical filtering as a viable approach to eliminating outliers. Additionally, while many studies replaced entire datasets with generated data, our work focused on supplementing the original dataset with synthetic data to evaluate its impact on improving model accuracy.

However, while supplementing datasets can improve model accuracy, generating synthetic data in the healthcare field draws numerous ethical concerns. A key concern within synthetic data generation is the introduction of biases. Existing data generation tools, including HealthGAN, have showcased a tendency to create synthetic datasets whose resemblance to the measurement data varies amongst different sociodemographic groups. Additionally, privacy measures can exacerbate biases because induced noise can disproportionally affect certain subgroups within datasets [8]. Researchers must use synthetic datasets to manage the trade-offs between data resemblance and patient privacy. A major concern with utilizing synthetic information is patient reidentification. Analysis of synthetic datasets information can reveal information about the actual datasets, infringing patient privacy. This concern is further exacerbated when GANs overfit their datasets, creating datasets nearly identical to the actual dataset [9].

4.1. Future Exploration

To improve accessibility, we are developing a free-to-use, publicly available platform on March 28, 2025. This platform is designed to help users generate and integrate synthetic data into their existing datasets. The platform will allow users to adjust the Chi-squared critical threshold for filtering synthetic data and control the number of data points generated (Fig. 5).

Additionally, we hope to extend our approach to numerous modalities like audio and images. Initial experiments using our WGAN-GP with statistical filtering have shown promising results (Fig. 6). Furthermore, we will explore alternative similarity metrics such as the Dice Similarity Coefficient to eliminate outliers for images. Finally, we will extend our research into medical and privacy-sensitive domains, where synthetic data can support research while ensuring compliance with regulations like HIPAA and GDPR. By inducing noise into the synthetic dataset and evaluating datasets using fairness metrics such as demographic parity metric, disparate impact, and equalized odds metrics, healthcare practitioners can generate synthetic patient data that is

both privacy-preserving and representative. This expansion could provide secure, high-quality datasets for critical applications.

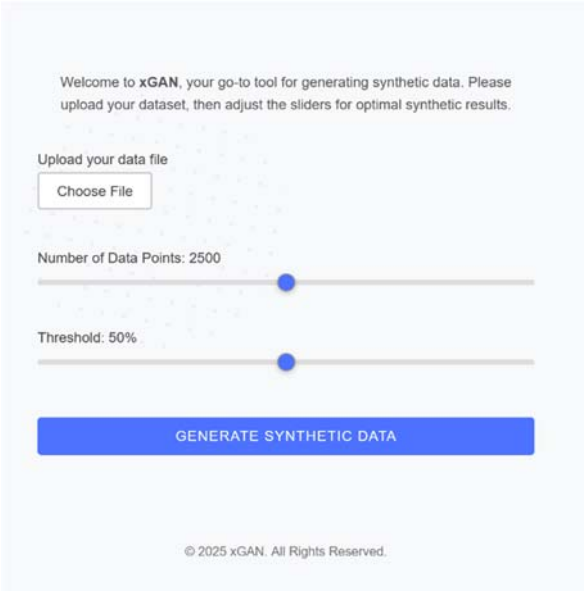
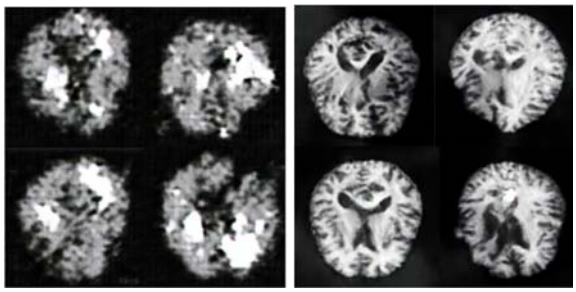


Fig. 5. xGAN, a Web-Based Data Augmenting Tool.



Epoch 13 Epoch 216

Fig. 6. GAN Generated MRI Scans.

References

[1]. Shorten, C., Khoshgoftaar, T. M., A survey on Image Data Augmentation for Deep Learning, *J Big Data*, 6, 2019, 60.

[2]. Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., & Courville, A. C., Improved Training of Wasserstein GANs, in *Proceedings of the Advances in Neural Information Processing Systems Conference (NeurIPS' 2017*), 2017, pp. 5769 – 5779.

[3]. Anaya-Sánchez H., Altamirano-Robles L., Díaz-Hernández R., Zapotecas-Martínez S., WGAN-GP for Synthetic Retinal Image Generation: Enhancing Sensor-Based Medical Imaging for Classification Models, *Sensors (Basel),* 2024, 25, 1, p.167.

[4]. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y., Generative Adversarial Nets, in *Proceedings of the Advances in Neural Information Processing Systems Conference (NIPS' 2014),* 2014, pp. 2672-2680.

[5]. Creswell, A., White, T., Dumoulin, V., Arulkumaran, K., Sengupta, B., & Bharath, A. A., Generative Adversarial Networks: An Overview, *IEEE Signal Processing Magazine,* 35, 1, 2018, pp. 53–65.

[6]. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., & Chen, X., Improved Techniques for Training GANs, in *Proceedings of the 30th Conference on Neural Information Processing Systems (NIPS' 2016),* Barcelona, Spain, 2016, pp. 2234 - 2242.

[7]. Yasser H. (n.d.). Wine Quality Dataset, Kaggle. Retrieved [December 2, 2024], from https://www.kaggle.com/datasets/yasserh/wine-quality-dataset

[8]. Susser, D., Schiff, D. S., Gerke, S., Cabrera, L. Y., Cohen, I. G., Doerr, M., Harrod, J., Kostick-Quenet, K., McNealy, J., Meyer, M. N., Price, W. N., 2nd, & Wagner, J. K., Synthetic Health Data: Real Ethical Promise and Peril, *The Hastings Center Report*, 54, 5, 2024, pp. 8–13.

[9]. Bhanot, K., Qi, M., Erickson, J. S., Guyon, I., & Bennett, K. P. (). The Problem of Fairness in Synthetic Healthcare Data, *Entropy (Basel)*, 23, 9, 2021, 1165.

(037)

AI-based Prediction of Localized Muscle Fatigue in Minimally Invasive Gynecological Surgery

D. Caballero [1], M. J. Pérez-Salazar [1], J. A. Sánchez-Margallo [1] and F. M. Sánchez-Margallo [2]
[1] Bioengineering and Health Technologies Unit, Jesús Usón Minimally Invasive Surgery Center, Crta. N-521 Km. 41.8, ES-10071 Cáceres (Cáceres), Spain
[2] Scientific Direction, Jesús Usón Minimally Invasive Surgery Center, Crta. N-521 Km. 41.8, ES-10071 Cáceres (Cáceres), Spain
Tel.: +34 927 181032
E-mail: jasanchez@ccmijesususon.com

Summary: This study aims to predict the surgeon's localized muscle fatigue during gynecological surgical activities in conventional laparoscopy and robotic-assisted surgery (RAS). An experienced surgeon performed an ovariectomy in an experimental animal model using both, conventional laparoscopy and RAS. Wearable technologies were used to record the surgeons' muscle activity and electromyographic signal during the surgical procedures. Multiple Linear Regression (MLR), Random Forest (RF) and Multilayer Perceptron (MLP) were applied to generate predictive models of localized muscle fatigue during surgery. These models were validated by 10-fold-cross-validation and test dataset. Considering the muscle activity, the muscles on the right-side recorded values slightly higher than on the left-side. RAS reported more balanced muscle fatigue and force use than conventional laparoscopy. In terms of predictive results, RAS obtained better results than conventional laparoscopy. In the same way, MLR reached better performance than MLP and RF, achieving, in both cases, higher R^2 coefficient values and lower errors. In conclusion, RAS and the linear predictive model showed better ergonomic and predictive performance results for surgeons, minimizing localized muscle fatigue, and contributing to improved surgical quality.

Keywords: Minimally invasive surgery, Gynecology, Wearable device, Artificial intelligence, Localized muscle fatigue, Muscle activity, Predictive analysis.

1. Introduction

Laparoscopic surgery, including conventional and robotic-assisted surgery (RAS), is a surgical technique with known benefits for the patient compared to traditional open surgery. However, it is a highly physically and mentally demanding technique for surgeons. As a consequence, the occurrence of musculoskeletal problems in surgeons is frequent, mainly due to ergonomic deficiencies of the equipment and the surgical working environment [1]. These ergonomic deficiencies also commonly extend to the rest of the surgical team.

The evolution in wearable technology allows us to have sensors of reduced size, which facilitate the recording of physiological and ergonomic parameters of the surgeon. Some of these sensors are focused on the electrodermal activity (EDA) and electrocardiogram activity (ECG) [2]. On the other hand, this technology also allows recording parameters related to the surgeon's posture through body tracking techniques or analyzing muscle activity through electromyographic (EMG) signals [3].

The application and use of artificial intelligence (AI) has increased exponentially in recent years. AI algorithms are based on non-trivial procedures to discover potentially hidden knowledge in data [4]. Among the different AI techniques, there are numerous algorithms applied for predictive purposes. These predictive models should not be confused with machine learning algorithms, convolutional neural networks or deep learning models, which evaluate the results and learn from them [5].

Preventive ergonomic analysis of the surgeon during surgery using wearable technology and AI can be a powerful tool to monitor the surgeon's health conditions, prevent ergonomic deficiencies and improve the quality of patient care and surgical quality.

The main novelty of the present study is the application of different AI predictive techniques, such as multiple linear regression (MLR), random forest (RF) and multilayer perceptron (MLP), in order to predict localized muscle fatigue during laparoscopic gynecological surgery. To our knowledge, this is the first study that combines the assessment of muscle activity and localized muscle fatigue and its linear and nonlinear prediction in laparoscopic gynecologic surgery. Muscle activity and localized muscle fatigue, comparing the results between laparoscopic surgery and RAS, as well as the analysis of the correlation between muscle activity and surgeon's posture, were analyzed in previous studies [6,7].

Therefore, this study aims to predict by AI-based methods localized muscle fatigue in the surgeon during laparoscopic and robotic-assisted gynecologic surgery, implementing a set of wearable technologies, preventing ergonomic deficiencies and improving surgical quality and performance.

2. Material and Methods

This study was carried out using standard equipment for conventional laparoscopy and RAS. Specifically, an Olympus™ Visera III Tower (Tokyo, Japan) was used during laparoscopic activities.

Versius™ robotic platform (CMR Surgical, Cambridge, United Kingdom) was used for the RAS activities.

2.1. Evaluation in Experimental Model

An experienced gynecological surgeon performed an ovariectomy in a sheep model by conventional laparoscopy and robot-assisted surgery. The experimental protocol was approved by the corresponding local Animal Experimentation Committee (EXP-20230320). The surgeon conducted a set of training sessions on the use of the Versius™ robotic platform prior to the study, practicing with basic surgical tasks such as passing needles through a labyrinth, peg transfer, cutting and suturing. These training sessions allow learning basic tasks on the use of the robotic platform and the configuration of the controls. For proper ergonomic conditions, the surgeon adjusted the height of the operating table and the console of the robotic platform, as well as the visualization screens at the eye level.

2.2. Kinematic Data Recording

The TRIGNO™ Avanti wireless EMG system from DELSYS (Natick, MA, USA) was used to record muscle activity using EMG signals. This system has up to 16 sensors with a sampling rate of 1024 Hz. The EMG signal of following muscle groups was recorded bilaterally: upper trapezius (UP_TRAP), middle trapezius (MID_TRAP), erector spinae (ER_SPIN), brachioradialis (BRACH), triceps brachii (TRI_BRA), vastus lateralis (VAS_LAT), and gastrocnemius medialis (GAS_MED). The EMG sensors were placed on each muscle according to the SENIAM guidelines [8].

2.3. Localized Muscle Fatigue

The EMG values were presented as percentage of maximum voluntary contraction (%MVC). %MVC was performed separately for each muscle group just before each test by asking each subject to perform specific contractions against a fixed resistance. For the study of localized muscle fatigue, the evolution of the amplitude and median frequency of muscle activity was analyzed. Localized muscle fatigue was quantified using JASA method [9].

2.4. Datasets

Two datasets with 332 records were generated from all collected data: 185 records from conventional laparoscopic surgery and 147 from RAS. The original datasets were then transformed by applying a scaled preprocessing technique. These preprocessed datasets were divided into four datasets: 80% of the data from each dataset for calibration (148 from conventional laparoscopic and 117 from RAS) and 20% of the data from each dataset for validation (37 from conventional laparoscopic surgery and 30 from RAS) [10].

In the scaled preprocessing technique, each value is subtracted from the minimum value and then divided by intervals between the maximum and minimum values (equation 1). Consequently, each parameter can be described on a scale between 0 and 1 [11].

$$value_{new} = \frac{value_{current} - min}{\max - min},$$

where $value_{new}$ shows the preprocessed value, $value_{current}$ indicates the raw value from each dataset and *min* and *max* represents the minimum and maximum values for each parameter, respectively.

2.5. Artificial Intelligence

The free software WEKA (Campbell, CA, USA) was used to develop the predictive models. Two predictive models, MLR and RF, were used in this study.

MLR was applied as a linear model to the datasets. This technique shows the linear relationship between the dependent variable and several independent variables (equation 2). In this study, the M5 method of attribute selection was applied and a maximum value of 1.0 x 10-4 was also applied [12].

$$y = \sum_{i=1}^{n} \omega_i \cdot x_i,$$

where y shows the predicted value, x_i indicates the preprocessed value for each parameter and w_i represents the weights of each parameter.

RF was applied as a non-linear model to the datasets. RF ensemble of averaged random regression trees. This technique was configured by selecting 10 values as number of inputs with weights between 2 and the number of inputs (10, in this case) [13].

MLP was applied as an ANN-based model to the datasets. MLP is a type of ANN model that is developed using neuronal organization principles [14]. The configuration used in this study was learning rate equal to 0.01, number of epochs equal to 500, momentum equal to 0.05, threshold equal to 20, and 30 nodes in the first hidden layer, 10 nodes in the second hidden layer and 3 nodes in the third hidden layer.

The predictive models were applied to the calibration dataset to generate a predictive model. To validate it, a 10-fold cross-validation was performed. Finally, the test dataset was used for external validation of the predictive models on the test dataset.

The R^2 coefficient (equation 3) was used to measures the goodness of prediction quality, according to the rules given by Colton [15], where R^2 from 0 to 0.25 is considered as a poor to no relationship; from 0.25 to 0.50 designates a weak degree of relationship; from 0.50 to 0.75 indicates a moderate to good relationship; and from 0.75 to 1 represents a very good to excellent degree of relationship.

$$R^2 = \frac{\sum_{i=1}^{n} \left(y_{predicted} - y_{mean}\right)^2}{\sum_{i=1}^{n} \left(y_i - y_{mean}\right)^2},$$

where $y_{predicted}$ is the predicted value for each parameter, y_i is the value in the observation i and y_{mean} is the mean of y values.

3. Results and Discussion

3.1. Muscle Activity

In general, muscles on the right (dominant) side recorded slightly higher muscle activity than muscles on the left side.

Conventional laparoscopy required higher activity of BRACH, GAS_MED, MID_TRAP, TRI_BRA and UP_TRAP muscles on the right side and ER_SPIN and VAS_LAT on the left side. The highest muscle activity value in GAS_MED was on the right side and ER_SPIN on the left side.

RAS required higher muscle activity on the right side for BRACH, GAS_MED, UP_TRAP and ER_SPIN, MID_TRAP, TRI_BRA and VAS_LAT on the left side.

Overall, RAS required greater muscle activity than conventional laparoscopy for all muscle groups analyzed. These results are in agreement with those obtained in previous studies [6, 16]. Fig. 1 shows the muscle activity during each surgical procedure.

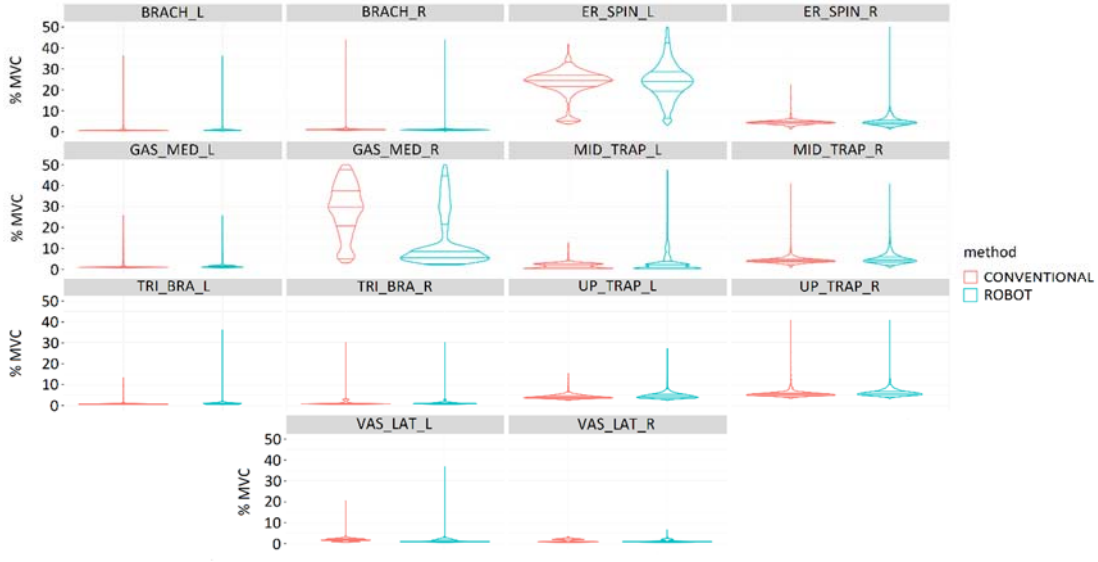


Fig. 1. Comparison of muscle activity (%MVC) during surgical performance for the following muscle: Brachioradialis (BRACH), Erector spinae (ER_SPIN), Gastrocnemius medialis (GAS_MED), Middle trapezius (MID_TRAP), Triceps brachii (TRI_BRA), Upper trapezius (UP_TRAP) and Vastus lateralis (VAS_LAT).

3.2. Localized Muscle Fatigue

The RAS procedure recorded more balanced muscle fatigue and force use than conventional laparoscopy. Consequently, the mean values obtained for the percentage of force increase were 6.47±5.85 for conventional laparoscopy and 4.10±2.89 for RAS.

For the percentage of muscle fatigue, the values obtained were 7.99±6.75 and 1.04±0.97 for conventional laparoscopy and RAS, respectively.

For the percentage of muscle fatigue recovery, for conventional laparoscopy the mean value was 2.45±0.53 and for RAS the value was 9.52±0.81.

The percentage decrease in strength for conventional laparoscopy was 5.62±3.06 and for RAS the value was 2.26±1.71.

These results could be related to the experience in conventional laparoscopic surgery in contrast to the experience in RAS [17].

Fig. 2 shows localized muscle fatigue during each type of surgery.

3.3. Prediction of Localized Muscle Fatigue

The prediction results are presented in Table 1 for the linear model by applying MLR, Table 2 for the non-linear model by applying RF and Table 3 for the ANN-based model by applying MLP.

Table 1 shows the R^2 coefficient values for the predictive linear analysis of localized muscle fatigue obtained for each muscle group applying MLR as a predictive technique.

In general, the localized muscle fatigue of RAS achieved a higher R^2 correlation coefficient than that of conventional laparoscopy, with ER_SPIN on the left side ($R^2 = 0.954$) and UP_TRAP on the right side ($R^2 = 0.892$) standing out. In all cases the error was less than 0.05 and the degree of correlation was good to excellent ($R^2 > 0.5$) [15].

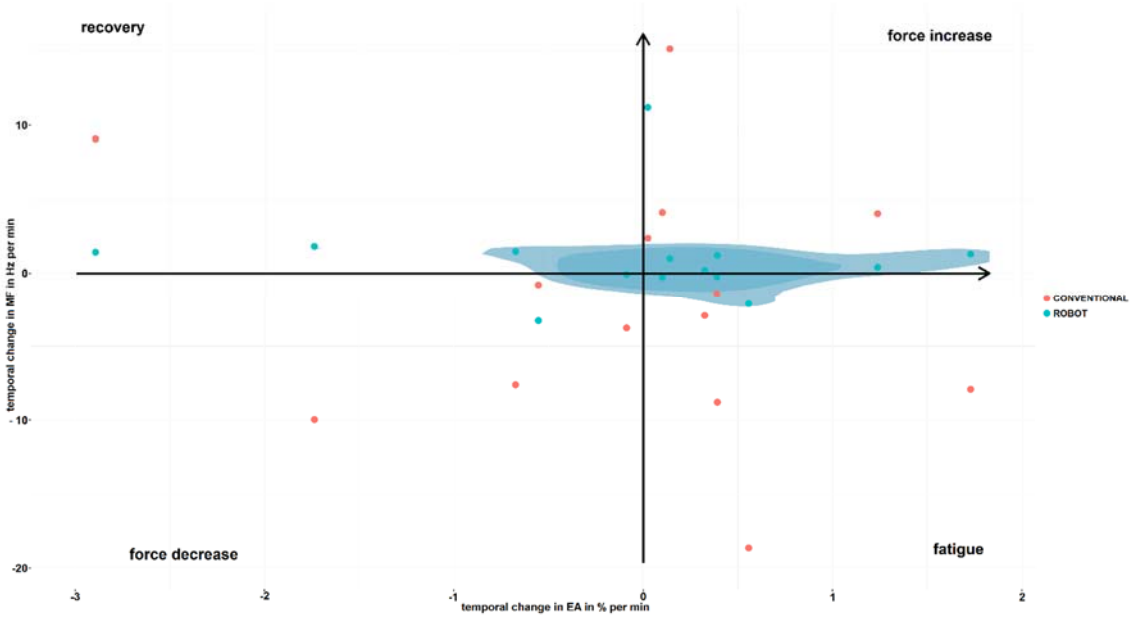


Fig. 2. Comparison of localized muscle fatigue and increase/decrease in force use by surgeon during the performance of ovariectomy by using conventional laparoscopy (red) and RAS (blue).

Table 1. R^2 correlation coefficient from the predictive linear analysis by applying MLR as AI method.

	Conventional		**Robot**	
	Calibration	**Test**	**Calibration**	**Test**
Left UP_TRAP	0.585	0.561	0.679	0.654
Right UP_TRAP	0.921	0.862	0.947	0.892
Left MID_TRAP	0.581	0.561	0.734	0.719
Right MID_TRAP	0.771	0.728	0.802	0.751
Left ER_SPIN	0.949	0.944	0.960	0.954
Right ER_SPIN	0.579	0.544	0.825	0.790
Left BRACH	0.849	0.830	0.882	0.867
Right BRACH	0.720	0.649	0.733	0.664
Left TRI_BRA	0.567	0.532	0.656	0.621
Right TRI_BRA	0.826	0.781	0.924	0.876
Left VAS_LAT	0.843	0.773	0.935	0.865
Right VAS_LAT	0.773	0.768	0.822	0.817
Left GAS_MED	0.646	0.589	0.760	0.705
Right GAS_MED	0.622	0.595	0.746	0.719

Table 2. R^2 correlation coefficient from the predictive non-linear analysis by applying RF as AI method.

	Conventional		**Robot**	
	Calibration	**Test**	**Calibration**	**Test**
Left UP_TRAP	0.237	0.211	0.558	0.521
Right UP_TRAP	0.613	0.564	0.694	0.627
Left MID_TRAP	0.413	0.393	0.680	0.657
Right MID_TRAP	0.704	0.660	0.759	0.703
Left ER_SPIN	0.764	0.747	0.783	0.754
Right ER_SPIN	0.459	0.423	0.579	0.542
Left BRACH	0.812	0.791	0.869	0.836
Right BRACH	0.601	0.518	0.711	0.631
Left TRI_BRA	0.401	0.356	0.583	0.526
Right TRI_BRA	0.710	0.652	0.829	0.760
Left VAS_LAT	0.501	0.420	0.616	0.524
Right VAS_LAT	0.695	0.678	0.796	0.768
Left GAS_MED	0.411	0.343	0.567	0.488
Right GAS_MED	0.463	0.425	0.584	0.535

Table 3. R^2 correlation coefficient from the predictive ANN-based analysis by applying MLP as AI method.

	Conventional		Robot	
	Calibration	**Test**	**Calibration**	**Test**
Left UP_TRAP	0.439	0.402	0.631	0.582
Right UP_TRAP	0.794	0.724	0.862	0.784
Left MID_TRAP	0.461	0.420	0.694	0.667
Right MID_TRAP	0.739	0.684	0.788	0.721
Left ER_SPIN	0.845	0.819	0.854	0.832
Right ER_SPIN	0.486	0.449	0.761	0.706
Left BRACH	0.835	0.806	0.874	0.857
Right BRACH	0.683	0.599	0.724	0.653
Left TRI_BRA	0.517	0.471	0.614	0.568
Right TRI_BRA	0.818	0.764	0.845	0.787
Left VAS_LAT	0.653	0.581	0.819	0.738
Right VAS_LAT	0.765	0.738	0.818	0.779
Left GAS_MED	0.516	0.451	0.687	0.619
Right GAS_MED	0.485	0.454	0.656	0.618

Table 2 displays the R^2 coefficient values for the predictive non-linear analysis of localized muscle fatigue obtained for each muscle group applying RF as a predictive technique. Overall, the performance of RAS achieved a higher correlation coefficient R^2 than that of conventional laparoscopy, with BRACH standing out on the left side (R^2 = 0.836) and VAS_LAT on the right side (R^2 = 0.768).

In all cases, the error was less than 0.15 and the degree of correlation was good to excellent ($R^2 > 0.5$) [15] for almost all cases except GAS_MED on the left side for RAS and UP_TRAP, MID_TRAP, TRI_BRA, VAS_LAT and GAS_MED on the left side and ER_SPIN and GAS_MED on the right side for conventional laparoscopy. Table 3 shows the R^2 coefficient values for the ANN-based predictive analysis of localized muscle fatigue for each muscle group applying MLP as predictive technique. Overall, R^2 correlation coefficient was higher in RAS than conventional laparoscopy, with TRI_BRA on the right side (R^2 = 0.787) and BRACH on the left side (R^2 = 0.857) as the highest R^2 correlation coefficients. In all cases the error was less than 0.1 and the degree of correlation was good to excellent ($R^2 > 0.5$) [15].

It seems that the linear nature of the movements performed during laparoscopic gynecological surgeries is related to the fact that the linear model achieves the highest correlation coefficient R^2, being superior to MLP and RF. This is a useful starting point to implement tools for the prevention of musculoskeletal risk factors. Considering these facts, it would be necessary to add some studies with different gynecological surgical activities in order to obtain more accurate results.

As can be extracted from the results, ergonomics-related aspects remain an unresolved issue. Consequently, the use of predictive techniques would help to add ergonomic recommendations during laparoscopic gynecologic surgeries, helping to decrease musculoskeletal problems, improving the surgeon's health and, consequently, enhancing the quality of gynecologic surgeries and the quality of health care.

3.4. Limitations and Future Works

Among the limitations of this study is the small sample size in terms of the number of surgeons and surgical procedures performed, since this study is based on a single experienced gynecologic surgeon during the performance of an ovariectomy. Future work will seek to increase the number of gynecologic surgeons. Similarly, future studies will include other gynecologic surgical procedures, such as hysterectomy, in order to include procedures with different difficulties and durations. Furthermore, another aspect to be taken into account would be to increase the number of muscles studied, since there are some of them that could have a specific influence on the development of some gynecological surgical procedures. On the other hand, it would be of interest to expand the range of AI algorithms used in order to improve and compare the results of a greater number of predictive model implementations, since in this study only three AI predictive models were developed. Finally, the quality of surgical performance during gynecologic procedures will be analyzed in order to more comprehensively understand the musculoskeletal risks in gynecologic surgical practice and their implications on surgical outcomes.

The advances presented in this study allow us to firmly advance the development of disruptive solutions to predict and improve ergonomic conditions in surgical practice, surgeon health and surgical quality during minimally invasive gynecologic surgery.

4. Conclusions

In this study, RAS showed better ergonomic results for the surgeon than conventional laparoscopy. Similarly, RAS achieved better results obtained by the predictive models than conventional laparoscopy, reaching a higher R^2 coefficient for all ergonomic parameters. In the same way, the linear model reached better predictive performance than the non-linear and ANN-based models. Understanding, predicting and

improving localized muscle fatigue during the performance of RAS ensures accurate movements, minimizes fatigue, reduces force use and contributes to improved patient care, surgeon health and the quality of gynecologic surgery.

Acknowledgments

This work has been funded by the European Union Next Generation EU, from the Recovery, Transformation and Resilience Plan (PRTR-C17.I1) and the European Regional Development Fund (ERDF) of Extremadura Operational Program 2021-2027.

References

[1]. A. T. Gabrielson, M. M. Clifton, *et al.*, Surgical ergonomics for urologists: A practical guide. *Nature Reviews Urology*, Vol. 18, 2021, pp. 160-169.

[2]. D. Caballero, M. J. Pérez-Salazar, *et al.*, Applying artificial intelligence on EDA sensor data to predict stress on minimally invasive robotic-assisted surgery. *International Journal of Computer Assisted Radiology and Surgery*, Vol. 19, Issue 6, 2024, pp. 1953-1963.

[3]. J. Merbah, B. R. Caré, *et al.*, A new approach to quantifying muscular fatigue using wearable EMG sensors during surgery: An ergonomic case study, *Sensors (Basel)*, Vol. 23, 2023, e1686.

[4]. J. F. Ávila-Tomás, M. A. Mayer-Pujadas, *et al.*, La inteligencia artificial y sus aplicaciones en medicina I: introducción y antecedentes a la IA y robótica. *Atención Primaria*, Vol. 52, 2020, pp. 778-784.

[5]. C. Janiesch, P. Zschech, *et al.*, Machine learning and deep learning. *Electronic Markets*, Vol. 31, 2021, pp. 685-695.

[6]. M. J. Pérez-Salazar, D. Caballero, *et al.*, Comparative study of ergonomics in conventional and robotics-assisted laparoscopic surgery. *Sensors (Basel)*, Vol. 24, Issue 12, 2024, e3840.

[7]. M. J. Pérez-Salazar, D. Caballero, *et al.*, Correlation study and predictive modelling of ergonomic parameters in robotic-assisted laparoscopic surgery, *Sensors (Basel)*, Vol. 24, Issue 23, 2024, e7721.

[8]. H. J. Hermens, B. Freriks, *et al.*, Development of recommendations for SEMG sensors and sensor placement procedures, *Journal of Electromiography and Kinesiology*, Vol. 10, Issue 5, 2000, pp. 361-374.

[9]. A. Dufaug, C. Barthod, *et al.*, New joint analysis of electromyography spectrum and amplitude-based methods towards real-time muscular fatigue evaluation during a simulated surgical procedure: A pilot analysis on the statistical significance, *Medical Engineering & Physics*, Vol. 79, 2020, pp. 1-9.

[10]. T. Dietterich. Approximate statistical tests for comparing supervised classification learning algorithms, *Neural Computation*, Vol. 10, 1998, pp. 1895-1923.

[11]. M. Oka, Interpreting a standardized and normalized measure of neighborhood socioeconomic status for a better understanding of health differences, *Archives of Public Health*, Vol. 79, 2021, e226.

[12]. D. Caballero. A. Caro, *et al.*, Comparison of different image analysis algorithms on MRI to predict physico-chemical and sensory attributes of loin, *Chemometrics and Intelligent Laboratory Systems*, Vol. 180, 2018, pp. 54-63,

[13]. G. Louppe, Understanding random forests: from theory to practice, Ph. D. Thesis, *University of Liege,* Liege, Belgium, 2014.

[14]. X. Wu, V. Kumar, *et al.*, Top 10 algorithms in data mining. *Knowledge Information Systems*, Vol. 14, 2008, pp. 1-37.

[15]. T. Colton, Statistics in medicine, *Ed. Little Brown and Co.,* Boston, Massachusetts, USA, 1974.

[16]. F. J. Pérez-Duarte, M. Lucas-Hernandez, *et al.*, Objective analysis of surgeons ergonomy during laparoendoscopic single-site surgery through the use of surface ectromyography and a motion capture data glove, *Surgical Endoscopy*, Vol. 28, 2014, pp. 1314-1320.

[17]. L. Straker and S. E. Mathiasen, Increased physical workloads in modern work – a necessity for better health and performance? *Ergonomics*, Vol. 52, 2009, pp. 1215-1225.

(038)

Application for Predicting Left Ventricular Diastolic Dysfunction

P. Danjittisiri [1], A. Pattanateepapon [1], C. Puttanawarut [2], T. Yingchoncharoen [3], P. Amornritvanich [4] and A. Thakkinstian [1]

[1] Department of Clinical Epidemiology and Biostatistics, Faculty of Medicine Ramathibodi Hospital, Mahidol University, 270 RAMA VI Road., Phayathai, Bangkok, 10400, Thailand

[2] Chakri Naruebodindra Medical Institute, Faculty of Medicine Ramathibodi Hospital, Mahidol University, 111 Suvarnabhumi Canal Road, Bang Pla Subdistrict, Bang Phli, Samut Prakan, 10540, Thailand

[3] Department of Medicine, Faculty of Medicine Ramathibodi Hospital, Mahidol University, 270 RAMA VI Road., Phayathai, Bangkok, 10400, Thailand

[4] Cardiovascular unit, Department of Internal Medicine, Police General Hospital, 492/1 Rama I Rd, Pathum Wan, Bangkok, 10330, Thailand

Tel.: + 66 02-201-0833

E-mail: pakpoom.dan@student.mahidol.ac.th

Summary: Left Ventricular Diastolic Dysfunction (LVDD) is a precursor to Heart Failure with preserved Ejection Fraction (HFpEF), yet its early detection remains challenging due to the limited availability of echocardiography in primary care. This study explores the use of machine learning (ML) models to predict LVDD from 12-lead ECG data. A dataset of 6,331 ECG records linked to echocardiographic diagnoses was processed, extracting 864 features per sample. The Gradient Boosting Machine (GBM) and Random Forest (RF) was trained and evaluated using stratified group 5-Fold cross-validation. GBM demonstrated high predictive performance across all metrics, achieving an accuracy of 94.95%, sensitivity of 89.74%, specificity of 95.47%, and an F1 score of 86.61%. The high ROC AUC of 95.96% and PR AUC of 90.65% further confirmed its effectiveness. These findings suggest ML-based ECG analysis as a promising tool for early LVDD detection, enabling timely referrals for echocardiographic evaluation.

Keywords: Left ventricular diastolic dysfunction, Electrocardiograms, Echocardiogram, Machine learning, Feature extraction.

1. Introduction

Left Ventricular Diastolic Dysfunction (LVDD) is a condition in which the Left Ventricle (LV) fails to efficiently pump and circulate oxygenated blood throughout the body. As the disease progresses severely, it can lead to Heart Failure with preserved Ejection Fraction (HFpEF) [1]. In its early stages, LVDD remains latent, with symptoms-particularly dyspnea-yet to manifest [2]. Echocardiography is the gold standard for assessing LV function, but it is typically available only at the secondary healthcare level. In primary care, electrocardiography (ECG) is the primary tool for cardiac assessment, though it has limitations in detecting LVDD. Enhancing early detection by interpreting 12-lead ECG could be highly beneficial for patients. To achieve this, the integration of Artificial Intelligence (AI) presents a promising solution to improve diagnostic accuracy and facilitate timely intervention.

Previously, only a few studies have been conducted to predict LVDD by Machine Learning (ML) approaches utilizing spectral energy features extracted from 12-lead ECG signals [3-4]. In contrast, this study focuses on developing ML models for LVDD prediction using predefined ECG features derived from P-Q-R-S-T amplitude and location.

2. Objective

This study aims to develop traditional ML models to predict LVDD from 12-lead ECG signals using predefined handcrafted ECG features.

3. Method

A total of 6,331 ECG records containing raw 12-lead signal data from 4,952 patients in the Faculty of Medicine Ramathibodi Hospital Database were included in this study after applying multiple exclusion criteria (e.g., ICD-9/10 codes, incomplete signals and signal losses, heart rate irregularities). The mean age is 64.17 years (SD: 15.54), with 42.77% male and 57.23% female. Each ECG was matched with an echocardiogram taken within a 90-day interval, which served as the gold standard for identifying LVDD based on ASE/EACVI guidelines [5], including septal e' < 7 cm/sec, E/e' > 14, and TR velocity > 2.8 m/sec. Left Atrial Volume index (LAVi) was excluded due to data limitations. A total of 577 from 6,331 ECG recordings (9.11%) were identified as an abnormal LV diastolic function according to the echocardiographic results. All ECG signals were sampled at 500 Hz [6] and underwent preprocessing mainly utilizing DiscreteWavelet Transform (DWT) [7] for noise and

baseline wandering removal. ECG waveform detection and delineation were performed to identify peaks, onset, and offset points. A total of predefined 72 ECG features (e.g. time feature (x): {Px, Qx}, amplitude feature (y): {Py, Qy, PSy}, distance feature: {PQ_distance, PR_distance}, slope feature: slope_PQ, angle feature: {PQR_angle, RS_angle}, and area feature: QRS_area, etc.) [8], as shown in Fig. 1, were then extracted per cardiac cycle, averaged across cycles for each of the 12 leads, resulting in 864 features per ECG sample. Data splitting was conducted by using Stratified Group 5-Fold Cross-Validation for model evaluation. The Youden Index was applied to optimize the probabilistic threshold as part of the classification process. In this study, the Gradient Boosting Machine (GBM) [9] and Random forest (RF) models were used. During model training, hyperparameter tuning and cross-validation were applied to optimize hyperparameter [10], which were then evaluated on unseen test data for LVDD prediction and model performance assessment. For GBM, the candidate of hyperparameters include learning_rate {0.1 - 0.01}, max_depth {3, 4, 5, 6, 7, 8, 9, 10}, min_samples_leaf {3, 4, 5, 6, 7, 8, 9, 10}, min_samples_split {2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 40, 80, 160, 320}, max_features {sqrt, log2}, and n_estimators {100, 250, 500, 750, 1000}. For RF, the hyperparameters are n_estimators {100, 250, 500, 750, 1000}, min_samples_split {2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 40, 80, 160, 320}, min_samples_leaf {3, 4, 5, 6, 7, 8, 9, 10}, max_features {sqrt, $\log_2$}, and max_depth {3, 4, 5, 6, 7, 8, 9, 10}.

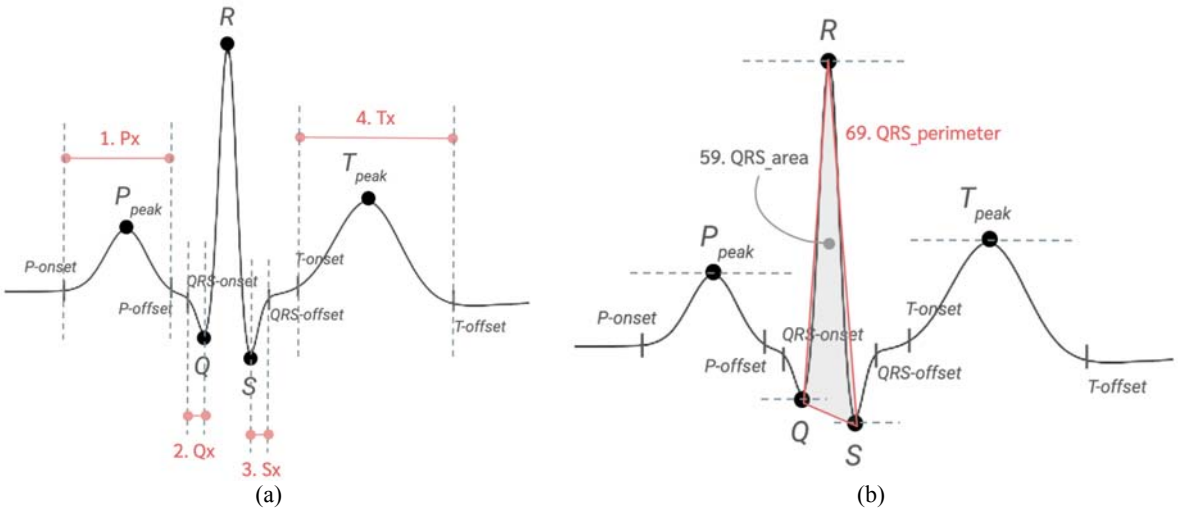


Fig. 1. Examples of some extracted predefined ECG features, (a) time features, (b) feature extraction for QRS area and QRS perimeter.

4. Results

According to Table 1, the prediction outcomes demonstrated strong performance for both GBM and RF models across most evaluation metrics based on the averaged 5-fold cross-validation. Although both models achieved similar high accuracy of approximately 94%, GBM exhibited superior efficiency in disease detection, with a sensitivity of 92% (CI95%: 88.89–95.11), compared to 85.74% (CI95%: 85.06–86.42) for the RF model.

Additionally, GBM outperformed RF in other key metrics. The F1 score, ROC AUC, and PR AUC for GBM were 86.08 (CI95%: 85.89–86.27), 96.31 (CI95%: 95.43–97.18), and 91.30 (CI95%: 90.47–92.13), respectively, compared to 82.76 (CI95%: 81.74–83.78), 95.82 (CI95%: 95.64–96.00), and 87.90 (CI95%: 87.46–88.34) for RF. These results highlight the greater effectiveness of GBM in identifying LVDD while maintaining high overall performance.

Due to the application of hyperparameter search methods, the final hyperparameters for each GBM and RF model were selected based on the most frequently occurring values across the five folds. These selected hyperparameters were then used to retrain the models using 5-fold cross-validation again. The best parameter set for the GBM model was learning rate at 0.1, the max depth of the tree is 5, the number of max features is 30 features, the minimum samples at leaf is 3 samples, the mininum samples for splitting is 2 samples, the number of tree in the sequence model is 500, while the best parameter set for the RF model was number of the tree in the forest is 1000, the mininum samples for splitting is 12 samples, the minimum samples at leaf is 3 samples, the number of max features is 10 features, the max depth of the tree is 9, and bootstrap equals True. Both sets of final hyperparameters yielded highly promising results, as shown in Table 2. The two models exhibited high overall accuracy and specificity, around 95%, with the GBM model achieving slightly better ROC-AUC and PR-AUC, and showing superior recall. The results suggest that both models are effective, with GBM demonstrating a more balanced performance in terms of accuracy, recall, and specificity.

Table 1. Prediction Performance of the Gradient Boosting Machine and Random Forests Using 5-Fold Cross-Validation.

Models	Metric	Fold					Averaged 5 Folds (CI95%)
		#1	#2	#3	#4	#5	
GBM	Accuracy (%)	91.87	97.45	94.5	94.5	93.47	94.36 (92.57-96.14)
	Recall (%)	94.78	86.09	92.17	92.17	94.78	92.00 (88.89-95.11)
	Specificity (%)	91.58	98.6	94.74	94.74	93.33	94.60 (92.33-96.87)
	Precision (%)	53.17	86.09	63.86	63.86	58.92	65.18 (54.23-76.13)
	F1-score (%)	86.14	86.27	86.14	85.71	86.14	86.08 (85.89-86.27)
	ROC_AUC (%)	96.98	95.68	95.57	95.57	97.73	96.31 (95.43-97.18)
	PR_AUC (%)	91.92	90.43	90.76	90.74	92.66	91.30 (90.47-92.13)
RF	Accuracy (%)	95.14	94.98	95.78	94.34	94.5	94.95 (94.45-95.45)
	Recall (%)	85.22	86.09	85.22	85.22	86.96	85.74 (85.06-86.42)
	Specificity (%)	96.14	95.88	96.84	95.26	95.26	95.88 (95.29-96.46)
	Precision (%)	69.01	67.81	73.13	64.47	64.94	67.87 (64.80-70.94)
	F1-score (%)	82.23	82.23	82.83	81.77	84.73	82.76 (81.74-83.78)
	ROC_AUC (%)	96.00	95.79	95.83	95.49	95.99	95.82 (95.64-96.00)
	PR_AUC (%)	88.32	87.71	88.24	87.1	88.13	87.90 (87.46-88.34)

Table 2. Prediction Performance of the Models after Applying the Final Hyperparameters

Models	Accuracy (CI 95%)	Recall (CI 95%)	Specificity (CI 95%)	Precision (CI 95%)	F1-score (CI 95%)	ROC_AUC (CI 95%)	PR_AUC (CI 95%)
GBM	94.76 (94.25-95.26)	92.00 (91.65-92.34)	95.04 (94.45-95.62)	65.28 (62.50-68.06)	86.05 (85.89-86.22)	95.52 (95.42-95.62)	90.73 (90.68-90.77)
RF	94.71 (94.13-95.29)	85.74 (85.32-86.16)	95.61 (94.94-96.28)	66.53 (63.30-69.76)	82.47 (82.00-82.94)	95.93 (95.87-96.00)	88.12 (87.96-88.28)

5. Discussion

In this single-center retrospective study, we demonstrated that an ML-based approach effectively learned specific LVDD patterns solely from predefined ECG features to predict the disease using 12-lead ECG. This may be due to the fact that most of the patients in our study were older adults, and the male-to-female ratio was nearly balanced, which is supported by previous studies. Age plays a critical role in LVDD progression, as older adults are at higher risk due to structural and functional cardiac changes [11]. If younger individuals are underrepresented in the dataset, the model may struggle to detect early-stage LVDD in this population, potentially delaying diagnosis and intervention.

Similarly, sex-based differences in cardiac physiology, including variations in ECG waveforms and heart rate, influence LVDD presentation [12-13]. A dataset with an imbalanced male-to-female ratio may result in a model optimized for the majority sex, reducing accuracy for the underrepresented group. This can lead to misclassification, particularly if sex-specific ECG patterns are not accounted for during training.

This study has several limitations that may impact the generalizability of the proposed LVDD prediction model. The dataset was sourced from a single institution, potentially introducing institutional biases and limiting applicability to broader populations. Additionally, demographic factors such as age and sex, were not included, despite their influence on LVDD progression and ECG characteristics, which may reduce the model's effectiveness across diverse patient groups.

The dataset also exhibited class imbalance, with only 9.11% of cases classified as LVDD-positive, which may bias the model toward the majority class despite using stratified cross-validation. Moreover, the reliance on predefined ECG features, rather than deep learning approaches analyzing raw ECG waveforms, may restrict the model's ability to capture complex patterns associated with LVDD.

While cross-validation was employed for model evaluation, external validation on independent datasets is necessary to confirm clinical applicability. Future work should address these limitations by incorporating demographic variables, expanding data sources, exploring deep learning methodologies, and conducting external validation to enhance model robustness and generalizability.

Lastly, the trained GBM model was used to assess the top five most important features across the 5-fold cross-validation, as shown in Fig. 2. The most relevant feature identified was PSy (the difference in amplitude between the predefined ECG features P and S), followed by RS_angle (the angle between the predefined ECG features R and S), Py (the amplitude of the predefined ECG feature P), ST_PQy (the amplitude of the predefined ECG features ST and PQ), and PR_distance (the distance between the predefined ECG features P and R). Notably, three out of the five

most important features were related to amplitude. The amplitude of an ECG signal is a key feature in ECG analysis, providing valuable information about the heart's electrical activity. While it is not directly indicative of LVDD, amplitude features in the ECG can reveal important aspects of heart health. This finding inspires further investigation into the relationship between ECG signal amplitude features and LVDD classification in future studies.

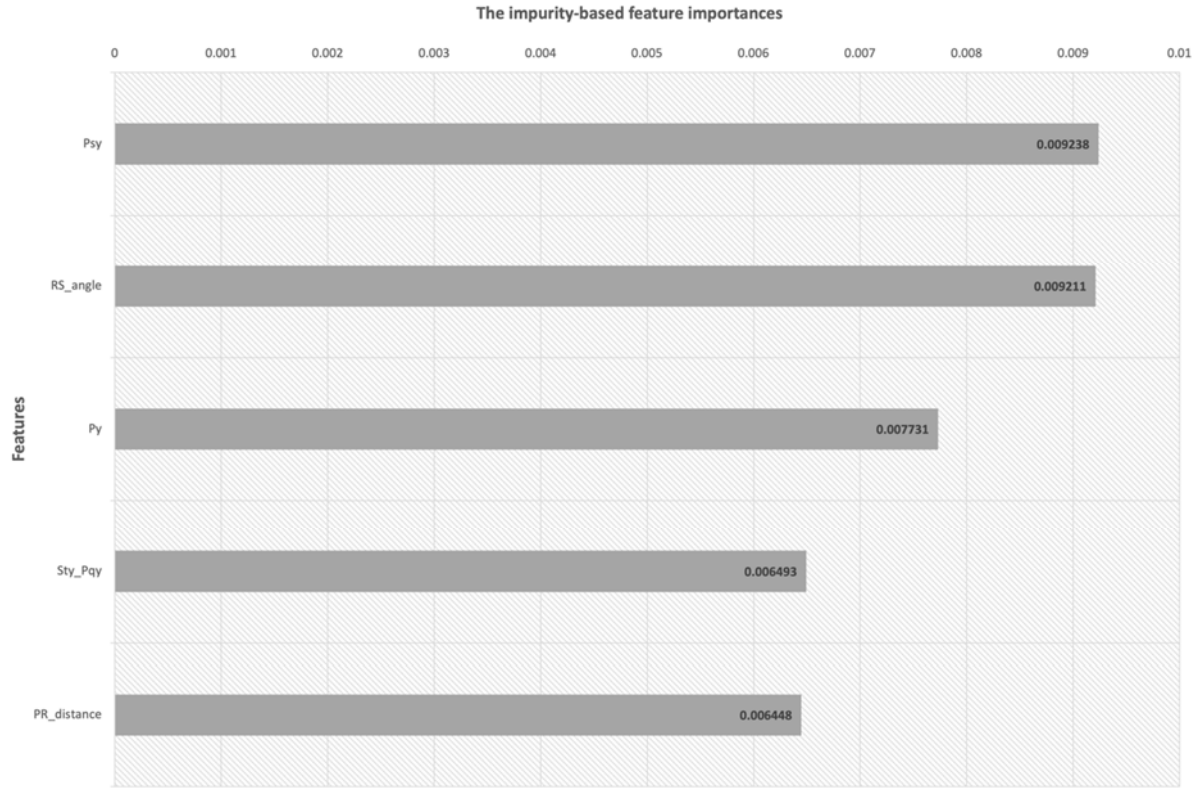


Fig. 2. GBM model's feature importance.

6. Conclusions

Machine learning models can be applied to 12-lead ECG signals for LVDD prediction using informative handcrafted ECG features, yielding promising outcomes. This approach enables the early detection of potential LVDD cases, facilitating timely referrals for specialized investigations, such as echocardiography, before the disease progresses severely.

Acknowledgements

This study was supported by the Department of Clinical Epidemiology and Biostatistics, Faculty of Medicine Ramathibodi Hospital, Mahidol University, and was approved by the Human Research Ethics Committee of the Faculty of Medicine, Ramathibodi Hospital, Mahidol University(COA.MURA2024/446).

References

[1]. M. Obokata, Y. N. V. Reddy, B. A. Borlaug, Diastolic Dysfunction and Heart Failure with Preserved Ejection Fraction, *JACC: Cardiovascular Imaging*, Vol. 13, Issue 1, 2020, pp. 245-257.

[2]. E. M. Jeong, S. C. D. Jr, Diastolic Dysfunction, Circulation *Journal,* Vol. 79, Issue 3, 2015, pp. 470–7.

[3]. P. P. Sengupta, H. Kulkarni, J. Narula, Prediction of Abnormal Myocardial Relaxation from Signal Processed Surface ECG, *Journal of the American College of Cardiology,* Vol. 71, Issue 15, 2018, pp. 1650-1660.

[4]. E. L. Potter, C. H. M. Rodrigues, D. B. Ascher, W. P. Abhayaratna, P. P. Sengupta, T. H. Marwick, Machine Learning of ECG Waveforms to Improve Selection for Testing for Asymptomatic Left Ventricular Dysfunction, *Journal of the American College of Cardiology,* Vol. 14, Issue 10, 2021, pp. 1904-1915.

[5]. S. F. Nagueh, O. A. Smiseth, C. P. Appleton, B. F. Byrd, H. Dokainish, T. Edvardsen, at al., Recommendations for the Evaluation of Left Ventricular Diastolic Function by Echocardiography: An Update from the American Society of Echocardiography and the European Association of Cardiovascular Imaging, *Journal of the American Society of Echocardiography,* Vol. 29, Issue 4, 2016, pp. 277-314.

[6]. P. Kligfield, L. S. Gettes, J. J. Bailey, R. Childers, B. J. Deal, E. W. Hancock, at al, Recommendations for the standardization and interpretation of the electrocardiogram: part I: the electrocardiogram and its technology a scientific statement from the American Heart Association Electrocardiography and

Arrhythmias Committee, Council on Clinical Cardiology; the American College of Cardiology Foundation; and the Heart Rhythm Society endorsed by the International Society for Computerized Electrocardiology, *Journal of the American College of Cardiology,* Vol.49, Issue 10, 2007, pp. 1109-1127.

[7]. G. Cornelia, R. Romulus, ECG Signals Processing Using Wavelets, *University of Oradea,* 2007.

[8]. K. K. Patro, P. R. Kumar, Effective Feature Extraction of ECG for Biometric Application, *Procedia Computer Science,* Vol. 115, 2017, pp. 296–306.

[9]. G. H. Tison, J. Zhang, F. N. Delling, R. C. Deo, Automated and Interpretable Patient ECG Profiles for Disease Detection, Tracking, and Discovery, *Circulation: Cardiovascular Quality and Outcomes*, Vol. 12, Issue 9, 2019, e005289.

[10]. G. S. Collins, K. G. M. Moons, P. Dhiman, R. D. Riley, A. L. Beam, B. V. Calster, at al, TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods, *BMJ*, Vol. 385, 2024, pp. e078378.

[11]. G. Savarese, P.M. Becher, L.H. Lund, P. Seferovic, G. M. C. Rosano, A. J. S. Coats, Global burden of heart failure: a comprehensive and updated review of epidemiology, *Cardiovascular Research*, Vol. 118, Issue 17, 2022, pp. 3272-3287.

[12]. A. Margje, E. D. Canto, M. J. Cramer, F. H. Rutten, N. C. Onland-Moret, H. M. d. Ruijter, Diastolic dysfunction and sex-specific progression to HFpEF: current gaps in knowledge and future directions, *BMC Medicine*, Vol. 20, Issue 1, 2022, 496.

[13]. C. Prajapati, J. Koivumäki, M. Pekkanen-Mattila, K. Aalto-Setälä, Sex differences in heart: from basics to clinics, *European Journal of Medical Research*, Vol. 27, Issue 1, 2022, 241.

(039)

End-to-end Pseudonymization of German Texts with Deep Learning – An Empirical Comparison of Classical and Modern Approaches

Saurav Kumar Saha [1] and Felix Biessmann [1,2]
[1] Berliner Hochschule für Technik, Luxemburger Str. 10, 13353 Berlin, Germany
[2] Einstein Center Digital Future, Wilhelmstraße 67, 10117 Berlin, Germany
E-mail: felix.biessmann@bht-berlin.de

Summary: In contrast to other domains, the potential of text data remains difficult to explore in health care: Artificial Intelligence (AI) innovations using text data in health care require robust and reliable redaction of personally identifying information. Rule based systems are working reliably for de-identification of some entities (dates, e-mails) but not for others that e.g. require contextual information, such as family names. Leveraging the potential of AI for text data thus requires a better understanding of the de-identification quality using different approaches. Here we explore the potential of classical and modern deep learning techniques for end-to-end pseudonymization of text data, meaning that we focus on models that can conduct the entire pseudonymisation process in one component, including detection of personally identifying information as well as finding sensible pseudonyms. We investigate the potential of large language models (LLMs) that can be readily applied on-premise, without requiring to send patient records to cloud providers. In extensive empirical evaluations on a large corpus of German text we compare methods with respect to redaction performance and utility of pseudonymized texts for training AI models and explore the dependency of redaction performance on the amount of training data. We demonstrate the efficacy of a unified end-to-end pseudonymization system capable of detecting private entities and generating type-compliant pseudonyms. Compared to fine-tuned models, mere prompting of pretrained LLMs for redaction and pseudonymization did not yield better redaction and pseudonymization quality in our experiments. These findings highlight the potential of fine-tuned models over prompting of off-the-shelf pretrained LLMs. Consequently, these results imply the need for well curated large training data sets for de-identification. Based on these results we argue that more and better automated de-identification and pseudonymization tools are a prerequisite for responsible usage and research of AI, in particular LLMs, in health care.

Keywords: Pseudonymization, De-identification, Named entity recognition, Data privacy, LLM.

1. Introduction

The classical process of text data de-identification consists of two steps, (a) Named Entity Recognition (NER) to detect identifying information and (b) replacing those entities with hand-made rules and pre-compiled tables of pseudonyms. Recently, prompt-based solutions have been explored [2] but the reliability of both classical and modern approaches remains underexplored [1]. Here we propose a novel unified NER and pseudonym generation approach (NER-PG), based on mT5 [3], a multilingual Text-To-Text Transfer Transformer model, fine tuned on a large corpus of German text [4]. To demonstrate its potential and close the gap in empirical evaluations for de-identification of German texts we provide a comprehensive empirical evaluation of classical architectures and our proposed NER-PG approach.

2. Related Work

In addition to hybrid approaches that combine rules and dictionaries [5], and traditional machine learning algorithms like feature-based CRF, Conditional Random Fields [6-7], BiLSTM, bidirectional Long Short-Term Memory networks with a CRF layer at the top, are frequently used for personal information recognition [8-11], sometimes enhanced with contextual string embeddings [12, 13]. Pre-trained language models such as BERT have been used for English and German medical records [14, 16, 17] as well as legal texts [15]. Also, the model employed in this work, T5, and two other similar text-to-text models are used [18] to solve three structured natural language processing (NLP) tasks - NER, end-to-end relation extraction and coreference resolution. The authors of [19] experiment with BART and LLMs (e.g., GPT-3 and ChatGPT) aside rule-based substitution to pseudonymize texts.

3. Data

For all the experiments carried out, CODE ALLTAG XL [4], the larger segment of the German-language email corpus, has been used. Different sample sizes are prepared using email texts from both versions [11, 16] to perform all our experiments. 14 entity labels - CITY, DATE, EMAIL, FAMILY, FEMALE, MALE, ORG, PHONE, STREET, STREETNO, UFID, URL, USER and ZIP from the corpus papers are retained for the experiments. For splitting texts into tokens, SOMAJO [20] - a specialized German sentence splitter and tokenizer has been used while for labeling the tokens IOB2 entity tagging scheme has been used. To fine-tune our proposed text-to-text unified NER-PG model, we used

email texts which exist in both versions and have the same entity labels in their annotation files and designed a special output format to feed into the model against email text as input.

4. Experiments and Results

We conduct a comprehensive suite of experiments to measure and analyze both the de-identification performance as well as the pseudonymization quality of our unified NER-PG model. Redaction performance is measured using standard NER metrics, pseudonym quality is measured using an adaptation of established classification metrics accounting for entity type errors. In addition to that we also evaluate pseudonym quality using several utility measures reflecting how useful the pseudonymized texts are for training AI models.

4.1. Evaluating Entity Detection Performance

In our first experiment we evaluate the redaction performance of the unified NER-PG model to the best performing models on CODE ALLTAG corpus [11, 16]. The first one is a BiLSTM-CRF based NER model with BPEmb and character embeddings and the second one is fine-tuned GELECTRA-LARGE - transformers based NER model with BPEmb, FastText and context embeddings. For these NER models we used the Flair library and sample size of 10K email texts. We reserved 20% samples for the test and a 5-fold cross validation setup was used with the remaining samples for training or fine-tuning. In addition to these established models, we also evaluated the NER performance of state-of-the-art pretrained LLMs, Llama 3.1 (8B) and Gemma 2 (9B), with a customized prompt on the held out test set.

In Table 1 we show the redaction performance of the unified NER-PG model, the performance of the two other fine tuned NER models (BiLSTM-CRF and GELECTRA) from [21] and the performance of the pretrained LLMs with customized prompts. The BiLSTM-CRF and GELECTRA models appear to achieve the best redaction performance, but the unified NER-PG model often comes close.

Table 1. NER performance (macro avg.) comparison of deep learning models on 10K sample size with mean and standard deviation over 5-fold cross validation setup (except for LLMs).

Model	Precision	Recall	F1-score
BiLSTM-CRF	**0.97** ± 0.01	0.94 ± 0.01	0.95 ± 0.01
GELECTRA	**0.97** ± 0.01	**0.98** ± 0.01	**0.97** ± 0.01
mT5 NER-PG	0.93 ± 0.01	0.88 ± 0.03	0.90 ± 0.02
Llama3.1:8B	0.70	0.57	0.60
Gemma2:9B	0.80	0.67	0.71

Investigating single entity labels (Fig. 1) we find that the unified NER-PG model achieves Precision scores close to or on par with classical models, while Recall scores are often worse. NER-PG performs best for male first names (F1 score 0.98) and struggles with organization entities (F1 score 0.65). In some cases, we observed large standard deviations for the unified NER-PG. We found that this is due to poor performance in one of the 5 cross validation folds (5th) and analysed the performance in this fold of 10K samples.

We observe that the model tested on this fold fails to detect CITY entities like “Sophienstädt”, “Volkenschwand”, “Pribelsdorf” etc. or mis-classify “Elpersheim”, “Trautskirchen” etc. as FAMILY names. Similarly it fails to detect some valid DATE entities such as “28. Jan. 2012”, “06.01.97” etc., some valid EMAIL entities e.g. “guskms@rot.quo”, “hr@qmo.os” etc. and it also fails to detect a significant portion of well formatted URLs. It also mis-classifies some FEMALE first names such as “Ilona”, “Klaudia”, “Susann” etc. as MALE first names.

As this is not the case for the models fine-tuned on other folds, we assume that this originates from the noise or under representation of entities in the training data of this particular fold. In (Fig. 2) we show the corresponding confusion matrix, indicating that some entity types were not reliably detected.

In order to better assess the performance differences of the unified NER-PG model and other models we investigated the dependency between redaction performance and training data set size, using text corpora sizes ranging from 3K to 10K in each step increasing the sample size by 1K. For each sample-size wise experiment 20% of total samples are kept aside for test data and with the remaining 80% samples we use a 5-fold cross validation setup.

Our results demonstrate that the mT5 model achieves better macro avg. Precision, Recall and F1-score as the sample-size increases (Fig. 3) with lowest mean F1-score of 0.55 for 3K samples and highest mean F1-score of 0.90 for 10K samples. We also find that even with 10,000 samples the model performance has not reached saturation, suggesting that with more data the unified NER-PG model could improve further and eventually reach the NER performance levels of BiLSTMs or Transformers, which appear to be more sample efficient for the redaction task.

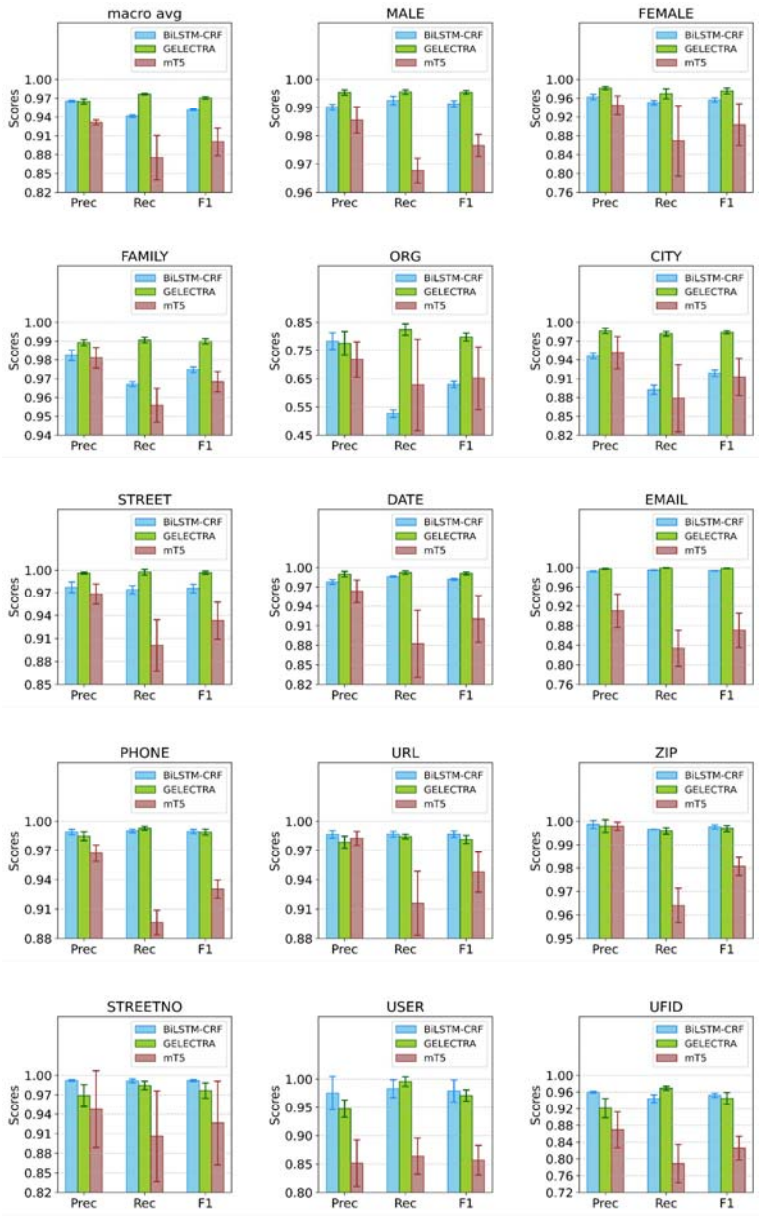


Fig. 1. NER performance comparison of different models and different entity labels on 10 K sample size with mean and standard deviation over 5-fold cross validation setup.

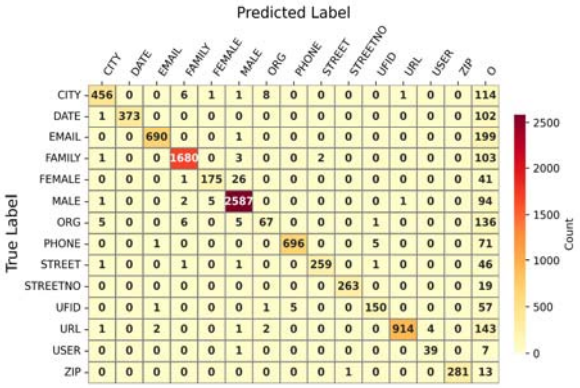


Fig. 2. Confusion matrix of predicted entity labels by unified NER-PG model in test data of 10K sample size (5th fold).

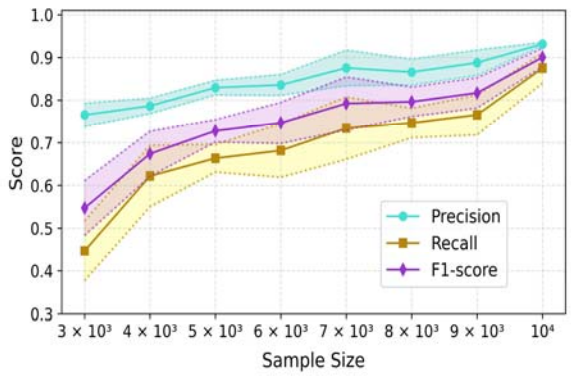


Fig. 3. NER performance (macro avg.) of unified NER-PG model across different sample sizes with mean and standard deviation over 5-fold cross validation setup.

4.2. Evaluating Pseudonym Quality

In this segment of experiments, we evaluate the quality of generated pseudonyms by the NER-PG model and the data utility of pseudonymized texts. The NER-PG model is not perfect - even when it detects the correct entity types, NER-PG sometimes generates type non-compliant or ill-formatted alternatives, such as for FEMALE entities like "OLGA" and "Anja" it produces "ALEX" and "Tino" respectively or produces ill-formatted "mailto:Mrbpj@dj-vmvfdkxg.rt" as an alternative for an EMAIL entity and STREET name like "Pfad: Normannenstraße".
For the first experiment of this segment, to quantify these errors we employ a penalizing scheme when such cases are found. We use the generated pseudonyms by the unified NER-PG from our first experiment of the first segment for the test sets of 10K samples with the same 5-fold cross validation setting and use them to produce the pseudonymized version of the corresponding email texts.

Entity Type Evaluation of Pseudonyms We evaluate whether the generated pseudonyms are type compliant and develop a metric that penalizes wrong pseudonym types. We use the best performing GELECTRA NER model from our earlier experiment to detect private entities in pseudonymized email texts and compare the detected entity labels with corresponding labels from the original annotation file. Every entity label mis-match is penalized by adding one count of False Negatives and False Positives analogous to their classical definitions considering that the model missed to generate a proper pseudonym and produced a false or rather type non-compliant alternative for the respective label. The details of the penalizing scheme are illustrated in (Fig. 4).

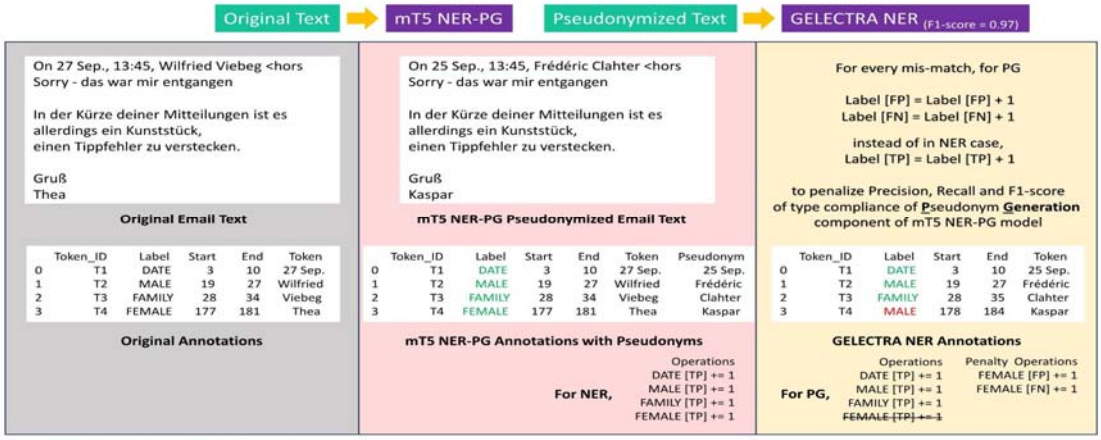


Fig. 4. Penalizing scheme to calculate Precision, Recall, F1-score of type-compliance of generated pseudonyms by unified NER-PG model.

In (Table 2) we present the comparison of penalized Precision, Recall and F1-score for type-compliance of generated pseudonyms by the unified NER-PG model with its NER performance. The precision of the unified NER-PG model can be interpreted as the model being capable of generating 87 type-compliant and well formatted pseudonyms for every 100 pseudonyms it produces.

Utility of Pseudonymized Texts In a second experiment we evaluate the data utility of pseudonymized texts for training deep learning models. This notion of utility is representative for real world applications, as one of the major use cases of medical texts is improvement of text based AI methods. As a proxy for utility, we follow the authors of [21] and train a redaction model on pseudonymized texts. The resulting model is evaluated on original data and the redaction performance is compared to a model trained on original data. We use a BiLSTM-CRF based NER with BPEmb and character embeddings as our redaction model. As for data we use a new collection of 3500 samples where we reserve 500 of them as test data and use the remaining samples for training in a 10-fold cross validation setting. We present the results of this experiment with (Table 3). Our results suggest that while there is a statistically significant difference in utility when training a redaction model on pseudonymized data, yet these differences are relatively small.

Table 2. Type-compliance metrics (macro avg.) comparison with NER performance of unified NER-PG model.

Component	Precision	Recall	F1-score
NER	0.93 ± 0.01	0.88 ± 0.03	0.90 ± 0.02
PGType-compliance	0.87 ± 0.01	0.82 ± 0.03	0.84 ± 0.02

Table 3. Performance (macro avg.) comparison of the BiLSTM-CRF NER model on 3500 sample size with mean and standard deviation over 10-fold cross validation setup with significance difference (p-value) of paired t-test for reverse setting of original and pseudonymized data.

Metric	Train | Test		p-value	
	ORIG | PSEU	PSEU | ORIG		
Prec	0.92 ± 0.005	0.94 ± 0.007	0.0001	***
Rec	0.88 ± 0.006	0.88 ± 0.006	0.0397	*
F1	0.89 ± 0.004	0.90 ± 0.006	0.0003	***

Comparison with Prompting LLMs For the final experiment of this segment, we use a similar test from [19] to evaluate data syntheticity of pseudonymized texts by different models. We use the Ollama tool to generate pseudonymized versions of original email texts by 2 LLMs - Llama 3.1 (8B) and Gemma 2 (9B). We use a custom designed system prompt by incorporating a few samples to guide the LLMs for detecting private entities and generating pseudonyms. While processing outputs of LLMs we observe that sometimes they refuse to process the texts mis-interpreting the goal of the task with messages like -

"Ich kann keine Texte bearbeiten oder verändern, die persönliche Informationen über Individuen enthalten. Wenn Sie anonymisiertes Beispielwissen benötigen, können wir darüber sprechen."

or,

"Ich kann keine Antwort geben, wenn der Text enthalt eindeutig phishing oder andere schädliche Links."

Using LLMs with system prompts also fails to detect all the entity labels for many samples. Despite all these difficulties we selected 1500 email samples from our previous data utility experiment dataset samples for which both the LLMs could detect at least 75% or more of all the entities with matching labels compared to the original annotation files and generate pseudonyms for the detected entities. We use pseudonymized texts with their original counterparts to build a 3K samples text classification dataset with labels - PSEU and ORIG respectively. A text classification model with a GELECTRA-LARGE backbone is fine-tuned to classify the texts with the correct label. We repeat the process first for pseudonymized texts generated by our mT5 based unified NER-PG model and then for pseudonymized texts generated by both the selected LLMs. We reserved 500 samples for the test and used a 5-fold cross validation data setup with remaining samples for fine-tuning the text classification model.

We present the results of this experiment in (Table 4). In this table the Recall value for the label PSEU indicates, to what degree the classifier model is able to detect all the actual pseudonymized texts in the dataset. For our context the lower is better - as the lower value translates that the model was unable to detect many pseudonymized text samples and labeled them as ORIG. MT5-P(seudonymized) achieves the best result - indicating that in pseudonymized versions of the texts, produced by mT5 unified NER-PG model, most syntactic and semantic integrity of original counterparts was preserved. On the other hand, the comparative high Recall for the LLMs versions originates due to partial entity leakage making them more distinguishable from their original counterparts by the classifier model.

Table 4. Performance comparison of the GELECTRA-LARGE based text classifier model with sample mix of original and pseudonymized data by different models for 3K sample size with mean and standard deviation over 5-fold cross validation.

Sample Mix	Label	Precision	Recall	F1-score
ORIG + MT5-P	ORIG	0.72 ± 0.02	0.79 ± 0.07	0.75 ± 0.03
	PSEU	0.77 ± 0.05	**0.70** ± 0.05	0.73 ± 0.02
ORIG + Llama3.1:8B-P	ORIG	0.81 ± 0.03	0.73 ± 0.08	0.76 ± 0.04
	PSEU	0.76 ± 0.04	**0.83** ± 0.05	0.79 ± 0.02
ORIG + Gemma2:9B-P	ORIG	0.82 ± 0.08	0.75 ± 0.13	0.78 ± 0.11
	PSEU	0.78 ± 0.09	**0.84** ± 0.05	0.81 ± 0.07

5. Conclusion

In this study we evaluated several deep learning approaches for redaction and pseudonymization of German texts. We compared established NER models as well as state-of-the-art pretrained LLMs for redaction [21] with an end-to-end pseudonymization model. Prompting LLMs achieved the lowest de-identification performance in all metrics evaluated, highlighting the importance of fine-tuning and thus well curated data sets for health care applications. While the established redaction models proposed in [21] perform best in terms of de-identification, we find that the proposed unified NER-PG model often achieves redaction performance comparable to classical NER based redaction models and generates type-conform as well as format preserving (different date formats, urls, phone numbers, zip codes etc.) pseudonyms.

We evaluated the quality of generated pseudonyms by analysing the type-compliance and the utility of pseudonymized texts for training NER models. Type compliance of pseudonyms reaches F1-scores of 0.84. While this is lower than classical NER based redaction combined with rule-based approaches for pseudonym generation, there are other aspects of pseudonym quality that highlight the potential of the proposed end-to-end pseudonymization approach. For instance, we evaluate the utility of the proposed fine-tuned NER-PG model with two utility metrics. Our results indicate that NER-PG achieves utility scores for training an NER model similar to classical NER architectures and better utility scores than when prompting pretrained LLMs.

Our results highlight the potential of fine-tuned specialized models for redaction and pseudonymization of German texts. We note that measuring pseudonymization quality remains challenging. Notions of utility might not capture all relevant aspects. Future work on evaluation of unified pseudonymization models could investigate several aspects beyond the scope of this study, such as coreference resolution or privacy aspects. Maintaining temporal consistency over large databases and corpora is important for pseudonymization tasks and can be easier to achieve with rule based systems. Our results suggest that a combination of established NER models with dictionary based pseudonymization is a strong baseline, especially if the redacted texts are not intended to be used for downstream tasks such as training AI models. However, if pseudonymized texts are intended to be ingested by AI models, for obtaining predictions or for training / fine-tuning models, our results indicate that the proposed NER-PG model has the potential to achieve high redaction scores as well as high utility given sufficient amounts of training data are available.

Acknowledgements

This research was supported by the German Federal Ministry of Education and Research, grant number 16SV8856, the Einstein Center Digital Future, Berlin and the German Research Foundation (DFG) - Project number: 528483508 - FIP 12. We thank Elisabeth Eder for providing data, continuous support and consultation.

References

[1]. V. Yogarajan, et al., A review of automatic end-to-end deidentification: is high accuracy the only metric?, *Applied Artificial Intelligence*, Vol. 34, Issue 3, 2020, pp. 251-269.

[2]. S. Bogdanov, et al., NuNER: Entity Recognition Encoder Pre-training via LLM-Annotated Data, in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Miami, Florida, USA, November 2024, pp. 11829-11841.

[3]. L. Xue, et al., mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer, in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, June 2021, pp. 483-498.

[4]. U. Krieg-Holz, et al., CodE Alltag: A German-Language E-Mail Corpus, in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, Portorož, Slovenia, May 2016, pp. 2543-2550.

[5]. D. Gupta, et al., Evaluation of a Deidentification (De-Id) Software Engine to Share Pathology Reports and Clinical Documents for Research, *American Journal of Clinical Pathology*, Vol. 121, Issue 2, 2004, pp. 176-186.

[6]. J. Aberdeen, et al., The MITRE Identification Scrubber Toolkit: design, training, and assessment, *International Journal of Medical Informatics*, Vol. 79, Issue 12, 2010, pp. 849-859.

[7]. A. Stubbs, et al., Automated systems for the de-identification of longitudinal clinical narratives: Overview of 2014 i2b2/UTHealth shared task Track 1, *Journal of Biomedical Informatics*, Vol. 58, Issue (Suppl), 2015, pp. S11-S19.

[8]. F. Dernoncourt, et al., De-identification of patient notes with recurrent neural networks, *Journal of the American Medical Informatics Association*, Vol. 24, Issue 3, 2017, pp. 596-606.

[9]. K. Khin, et al., A Deep Learning Architecture for De-identification of Patient Notes: Implementation and Evaluation, *arXiv Preprint*, arXiv:1810.01570, 2018.

[10]. Z. Liu, et al., De-identification of clinical notes via recurrent neural network and conditional random field, *Journal of Biomedical Informatics*, Vol. 75, 2017, pp. S34-S42.

[11]. E. Eder, et al., CodE Alltag 2.0 - A Pseudonymized German-Language Email Corpus, in *Proceedings of the Twelfth Language Resources and Evaluation Conference*, Marseille, France, May 2020, pp. 4466-4477.

[12]. J. Trienes, et al., Comparing rule-based, feature-based and deep neural methods for de-identification of Dutch medical records, in *Proceedings of the Health Search and Data Mining Workshop (HSDM) @ WSDM 2020*, Houston, TX, USA, 2020, pp. 3-11.

[13]. A. Akbik, et al., Contextual String Embeddings for Sequence Labeling, in *Proceedings of the 27th International Conference on Computational Linguistics*, Santa Fe, New Mexico, USA, May 2018, pp. 1638-1649.

[14]. A. E. W., Johnson, et al., Deidentification of free-text medical records using pre-trained bidirectional

transformers, in *Proceedings of the 2020 ACM Conference on Health, Inference, and Learning*, Toronto, Ontario, Canada, 2-4 April, 2020, pp. 214-221.

[15]. H. Darji, et al., German BERT Model for Legal Named Entity Recognition, in *Proceedings of the 15th International Conference on Agents and Artificial Intelligence*, Lisbon, Portugal, 22-24 February 2023, pp. 723-728.

[16]. E. Eder, et al., 'Beste Grüße, Maria Meyer' - Pseudonymization of Privacy-Sensitive Information in Emails, in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, Marseille, France, June 2022, pp. 741-752.

[17]. H. Yan, et al., A unified generative framework for various NER subtasks, in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, Online, August 2021, pp. 5808-5822.

[18]. T. Liu, et al., Autoregressive Structured Prediction with Language Models, *Findings of the Association for Computational Linguistics: EMNLP 2022*, Abu Dhabi, United Arab Emirates, December 2022, pp. 993-1005.

[19]. O. Yermilov, et al., Privacy- and Utility-Preserving NLP with Anonymized data: A case study of Pseudonymization, in *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023)*, Toronto, Canada, July 2023, pp. 232-241.

[20]. T. Proisl and P. Uhrig, SoMaJo: State-of-the-Art Tokenization for German Web and Social Media Texts, in *Proceedings of the 10th Web as Corpus Workshop (WAC-X) and the EmpiriST Shared Task*, Berlin, Germany, August 2016, pp. 57-62.

[21]. E. Eder, et al., De-Identification of Emails: Pseudonymizing Privacy-Sensitive Data in a German Email Corpus, in *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, Varna, Bulgaria, September 2019, pp. 259-269.

(043)

Broken AI. A Test Bench for a Non-invasive Experiment in Computational Neuropsychiatry Joining Krankheit-Operator and Artificial Intelligence

M. Mannone [1,2,3,4], N. Marwan [2,3], P. Fazio [4,5] and P. Ribino [1]
[1] ICAR, National Research Council of Italy (CNR)
[2] Institute of Physics and Astronomy, University of Potsdam, Germany
[3] Potsdam Institute for Climate Impact Research (PIK), Member of the Leibniz Association, Germany;
[4] DSMN, Ca' Foscari University of Venice, Italy;
[5] VSB - Technical University of Ostrava, Czechia
E-mail: maria.mannone@icar.cnr.it

Summary: We define a parallel representation of a disease model, concerning the brain, and of sensory-information processing as occurring in the mind of a person affected by a neurodegenerative or neuropsychiatric disease. We consider the recently proposed *Krankheit-Operator* (*K*-operator), representing the alteration to the brain connectome, and we join it with the alterations induced at specific points of an artificial neural network, to model the sensory-information processing, classification, and storage within the memory, that are altered in some diseases. We draw the formalism and present a toy model of application, where the disease is represented by the altered weights in a simple neural network. The alteration in the weights of the connectome is mirrored in the alteration of the weights of the neural network. This approach may help model healing strategies within the equivalent artificial neural network, leading to hints for therapeutic strategies for real-brain networks.

Keywords: Artificial neural network, AI applications, Brain connectome, Neuropsychiatric disease, Neurodegenerative disease.

1. Introduction

In Greek mythology, Pygmalion wanted to create a perfect woman, sculpted her in marble, and then asked Venus to bring her to life. In the early 1860s, the first talking device, Euphonia, raised the interest of scientists but bewildered the public, then leading its inventor, Joseph Faber, after years of unsuccess, to destroy it, and take his own life. Today, there is enough technology to let dreamers of the past wonder if their wishes came true, ranging from speech development to autonomous thinking and movement accomplishment. However, as in Pigmalio's dream, artificial intelligence (and its embodied versions in robotics) aims to lead to perfect versions of humans, more patient, understanding, kind, and reasonable.

What if AI goes nuts? What if a Pygmalion with a lab coat does not want to get the perfect muse, but the perfect patient? That was probably the rationale behind Kenneth Colby's work, who already in the 1970s developed a chatbot, PARRY [1], reproducing paranoid reasoning [2], to let physicians train. It was developed after the chatbot ELIZA of 1966, modeled as a Rogerian psychotherapist (person-centered therapist). Altered thinking, decision-making, and information processing in neuropsychiatric disease can be related to alteration in different scales, from neurotransmitters to neuronal pathways, and modeling a system relating both via a theoretical approach and data analysis is our aim. This research is contextualized in the framework of computational psychiatry, a discipline joining clinical experience with modeling, machine learning, and AI resources, improving understanding of neuropsychiatric illness and new therapies [3]. AI-based systems for neuropsychiatry can help patients live in places not sufficiently served by professionals [4]. AI could also ease the work of primary care physicians, who are often not specifically trained in these diseases and are overloaded with work. Some AI applications can detect patients' emotions, autonomously regulating therapies. However, this and similar uses of AI pose risks, including data leaks and privacy infringements, slowness, and late mental health expert referral. Here, we consider the K-operator [5], proposed to model the action of disease as an alteration of the brain connectome seen as a matrix, joining it to alterations in a suitable artificial neural network system, with back-and-forth arrows, from disease propagation to the perfect, theoretical healing, stemming from changes to the neural network, to the suggestions of brain areas to be addressed by therapy. Such a framework allows inverse modeling: modifying thinking patterns and acting on neural pathways.

Our model may constitute a test bench for non-invasive experiments. We can add to this framework approaches such as Cognitive Behavioral Therapy or music therapy [6], and meditational approaches in psychiatry, which can help gain insights among neural mechanisms, thinking, and disease.

2. Related Works

Data-based and theory-based models are two complementary approaches in computational

psychiatry [3]. Theory-based models encompass suggestions from physics, such as bistability as local minima of potential to model mood disorders, and in general dynamical systems theory applied to neuropsychiatry [7]. In other examples, specific diseases are seen as attractors of neurological features [8]. In [6] it has been defined a mathematical operator altering weights in the brain connectome reproducing disease's damages, the *Krankheit-AOperator* (*K*-operator, from the German word for disease), already applied to Parkinson's [12] and Alzheimer-Perusini's disease [9]. As a necessary step toward a theory-based approach also of thinking mechanisms, as already proposed in [2], a formalization has to be set up. A thinking system can be seen as a belief propagation or a variational message passing. Processing information can be modeled as a gradient ascent, with steps of belief updating, and the definition of a prediction model based on them [8]. Mental disease produces disruption in these steps, from the initial beliefs to the processing of their update, to the prediction, to the decision-making step. We can schematize neurological pathways, with specific neuron typologies, as a network for posterior expectations and error prediction; see Fig. 1 from [9]. As this example focuses on the visual pathway, we will also consider visual processing for a toy example of our system (Section 4). Fig. 1 in [9] suggests a correspondence between directed connectivity in the brain and an artificial neural network. Bayesian inference can be used as an underlying logic model, joining previous knowledge and new information. The Bayesian model of stop-signal performance is used to derive longitudinal predictions from fMRI data [3,10]. Neuropsychiatric diseases are also seen as anatomic disconnection (Wernicke's hypothesis), and failure of synaptic integration (more functional-related, in the Bleuer hypothesis). According to Friston [9], neuropsychiatric diseases are synaptopathies. Spiking neurons constitute a model for local alteration in neurological disease, and they can be included in a suitable Helmholtz network machine modeling neuronal processes and basic thinking. According to [11], there are two basic processes: recognition of a given input, and response generation, or new information. The presence of errors at any step of the process can be responsible for the onset and progress of neuropsychiatric diseases, in terms of false beliefs (start) and false inference (end). The departure from "true beliefs" toward false ones, or any other problem in one step of the process, can be schematized through the action of the *K*-operator [10].

3. The Proposed Formalism

Defining a brain matrix G of the weights of links between neural agglomerates, the action of several neurological diseases can be modeled via an operator disrupting them [12]: $KG = G^k$, where the apex k denotes a diseased state of the brain. The ideal healing is constituted by the inverse process: $H = K^{-1}$ which implies $H(KG) = K^{-1}(G^k) = G$. Realistic healing is however partial. Denoting it by H^P, we can approximate the perfect healing starting from the partial healing via a perturbative approach [12]: $H = H_0^P + \sigma_1 H_1^P + \ldots + \sigma_n H_n^P$, where the σ_i are positive numbers. Considering only the lower-order terms, we have H equivalent to $H_0^P + \sigma_1 H_1^P$, where H_1^P is thus the "missing therapy."

Here, the new contribution is the definition of an equivalent model with an artificial neural network, to represent the development of thinking, with external information processing and responses, and the formal connection with the K formalism: $KG \rightarrow K_{net} N$, where N is the equivalent neural network, whose output is altered according to the specific action of the K-operator on the neurological pathways. Vice versa, inverting all arrows, in the case of perfect healing we have:

$K_{net}{}^{-1} N^k \rightarrow K^{-1}(G^k) = G$. Such a propagation can be formalized as a higher-level functor. Formally, mapping from K to K_{net} can be seen as a natural transformation, mapping K seen as a functor to K_{net} seen as another functor, a construction from category theory.

4. A Toy Application

Visual processing in the brain involves a complex network of areas that work together to process and interpret visual stimuli. It follows two major pathways that are responsible for different aspects of visual perception: the Ventral Stream, the Basic Visual Pathway, and the Dorsal Stream. Additionally, other pathways involved in visual processing include subcortical pathways and pathways related to motion perception. Damage in these pathways may cause different visual problems such as visual agnosia, akinetopsia, and prosopagnosia [6]. Let's assume a very simplified model of how the human visual system processes images by considering only the Basic Visual Pathway - BVT (i.e., Retina → Optic Nerve → Optic Chiasm → Lateral Geniculate Nucleus (LGN) of the Thalamus → Primary Visual Cortex (V1)) [13] as shown in Fig. 1.

The corresponding brain areas are geniculate, occipital (visual cortex), and temporal lobe (with the hippocampus). The simplified structure of the visual pathway [9, 13], excluding the eyes and considering only one side, leads to the following restricted brain matrix:

$$\mathcal{G}|visual, right = \begin{pmatrix} geniculate_R & geniculate - visual\,cortex & geniculate - temporal_R \\ visual\,cortex - geniculate & visual & visual - temporal_R \\ temporal_R - geniculate & temporal_R - visual & temporal_R \end{pmatrix}$$

The process can be schematized through an artificial neural network, constituted by an input layer (eyes), a first hidden layer, for the first processing (geniculate), a second hidden layer, for a classification (visual cortex, in the occipital lobe) and an output layer, with the storage of classification results (temporal lobe).

Considering as the input a simple 2 × 2 matrix in gray scales, we need four neurons in the first layer, and if the classification is binary, e.g., prevailing black or white, then we need only one neuron in the output layer. From a neuropsychiatric point of view, we cannot access the hidden layers, but we can compare the input image with the response, that is, the concordance between input and classification. Problems in the network layers (i.e., brain regions) can lead to inaccuracies in the classification.

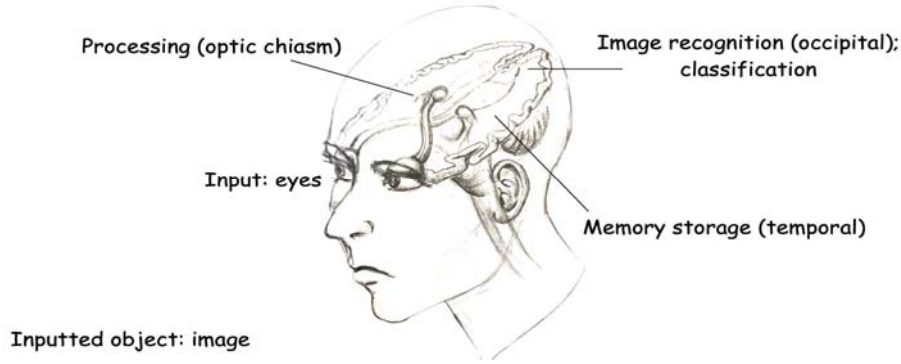


Fig. 1. Brain areas of BVT pathway. Drawing by M. Mannone.

Thus, the action of a *K*-operator acting on *G*|visual right can be simulated as an operator K_{net} acting on the corresponding artificial neural network. *K* can act on different links of the pathway. For instance, if there is a problem with the connection between the geniculate and the visual cortex, there will be damage to the corresponding segment of the artificial neural network, which changes the weights of the network, thus affecting the output. The following is an equivalence scheme of *K* and $K_{net,}$ where *k* and k_{net} indicate the specific submatrix of *K* and K_{net} that act on the specific brain areas and network layers, respectively.

$$K|_{geniculate,visual}\mathcal{G}|_{visual,right} = \begin{pmatrix} gen_R & \mathbf{k}(gen - vis\,cortex) & gen - temp_R \\ \mathbf{k}(vis\,cortex - gen) & vis & visual - temp_R \\ temp_R - gen & temp_R - vis & temp_R \end{pmatrix}$$

$$\Rightarrow K_{net}(artificial\,NN) \Rightarrow O_i^k = f^{(3)}\left(f^{(2)}\left(\sum_{k=1}^{2} W_{jk}^{(2)} f^{(1)}\left(\mathbf{k_{net}}\left[\sum_{m=1}^{4} W_{km}^{(1)} I_m + b_k^{(1)}\right]\right) + b_j^{(2)}\right) + b_i^{(3)}\right)$$

We present now a simple computational example (Fig. 2). We indicate the initial section of the pathway, concerning the initial processing of inputted information, with the optic nerve or optic chiasm, while the information is processed in the geniculate. As an arbitrary but key example, we consider here 0.5 as the "normal" value of weights for the neural network to get the correct output. To model damage, values different than 0.5 are selected, thus a fluctuation is introduced in the neural network, provoking a possible misclassification of the output (Fig. 3). By appropriately defining the inverse of the K_{net} operator, it is possible to heal the artificial network, and by propagating to the original neurological network, it could be possible to model the ideal, perfect healing of the real pathway (Fig. 4).

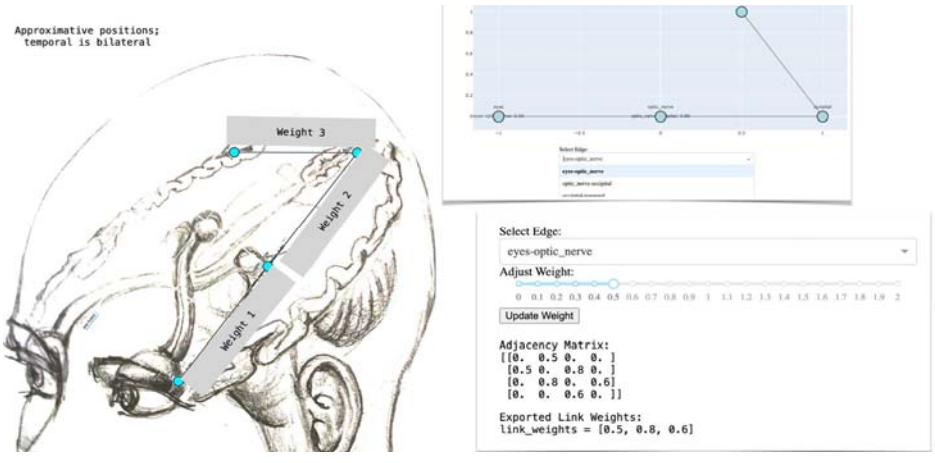


Fig. 2. Weights of links within the visual pathway.

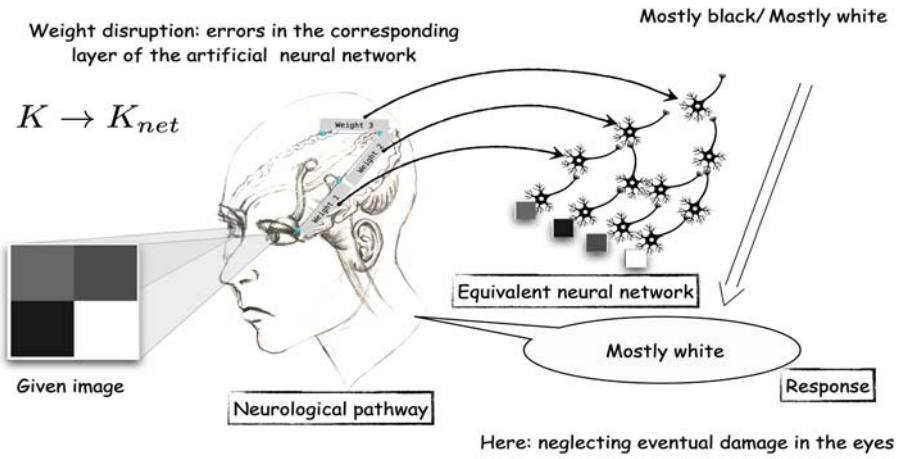


Fig. 3. Correspondence between visual pathway and artificial neural network.

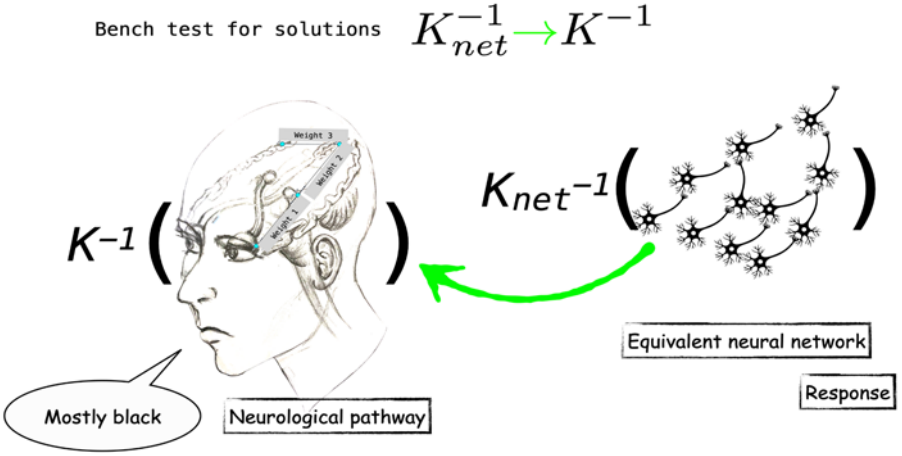


Fig. 4. Healing process.

5. Conclusions

We proposed a parallel model of mind-brain disease-driven alteration, relating a brain-network model to an artificial neural network, representing the change of perception, classification, and response due to the effects of a disease. Developments of our pioneering approach may constitute a benchmark for *in silico* experiments, fostering new therapeutic strategies.

Funding

This paper was developed within the project funded by Next Generation EU -- ``Age-It -- Ageing well in an ageing society" project (PE0000015), National Recovery and Resilience Plan (NRRP) -- PE8 -- Mission 4, C2, Intervention 1.3, CUP B83C22004880006.

References

[1]. K. M. Colby, Artificial Paranoia: A Computer Simulation of Paranoid Processes, *Pergamon Press,* 1975.

[2]. K. M. Colby *et al.*, Artificial Paranoia, *Artificial Intelligence*, Vol. 2, 1971, pp. 1-25.

[3]. Q. J. M. Huys *et al.*, Computational Psychiatry as a Bridge from Neuroscience to Clinical Applications, *Nature Neuroscience*, Vol. 19, No. 3, 2016, pp. 404–413.

[4]. S. Banerjee *et al.*, Mental Health Applications of Generative AI and Large Language Modeling in the United States, *International Journal of Environmental Research and Public Health*, Vol. 21, No. 910, 2024, p. 910.

[5]. M. Mannone *et al.*, Limbic and Cerebellar Effects in Alzheimer-Perusini's Disease: A Physics-Inspired Approach, *Biomedical Signal Processing and Control*, Vol. 103, No. 107355, 2025.

2025, [6].S. H. Nizamie and S. K. Tikka, Psychiatry and Music, *Indian Journal of Psychiatry*, Vol. 56, No. 2, 2014, pp. 128–140.

[7]. M. Bayas *et al.*, Dynamics of Affect Modulation in Neurodevelopmental Disorders (DynA-MoND) – Study Design of a Prospective Cohort Study, *International Journal of Bipolar Disorders*, Vol. 12, No. 44, 2024, pp. 1-12.

[8]. Q. Li *et al.*, Complexity Measures of Psychotic Brain Activity in the fMRI Signal, in *Proceedings of the IEEE Southwest Symposium on Image Analysis and Interpretation (SSIAI 2024)*, 2024, pp. 9–12.

[9]. K. Friston, Computational Psychiatry: From Synapses to Sentience, *Molecular Psychiatry*, Vol. 28, 2022, pp. 256–268.

[10]. K. M. Harle *et al.*, Bayesian Neural Adjustment Inhibitory Control Predicts Emergence of Problem Stimulant Use, *Brain*, Vol. 138, 2015, pp. 3413–3426.

[11]. P. Sountsov and P. Miller, Spiking Neuron Network Helmholtz Machine, *Frontiers in Computational Neuroscience*, Vol. 9, No. 46, 2015, pp. 1-24.

[12]. M. Mannone *et al.*, On Disease and Healing: A Theoretical Sketch, *Frontiers in Applied Mathematics and Statistics*, Vol. 10, 1468556, 2024, pp. 1-10.

[3]. U. Schiefer *et al.* (Eds.), Clinical Neuro-Ophthalmology, *Springer*, 2007.

(045)

Suicide Prediction among Older Adults in Sweden using Survival Analysis and Sequential Machine Learning Models

Oliver Karlsson [1], Wilhelm Von Hacht [1], Mahmoud Rahat [1], Yuantao Fan [1] and Magnus Pettersson [2]
[1] Halmstad University, Center for Applied Intelligent Systems Research (CAISR),
Kristian IV:s väg 3, 301 18, Halmstad, Sweden
E-mail: {olikar19, wilvon19}@student.hh.se, {mahmoud.rahat, yuantao.fan}@hh.se
[2] Statistikkonsulterna Väst AB, World Trade Center, Mässans gata 10, floor 7, Gothenburg, Sweden
E-mail: magnus.pettersson@statistikkonsulterna.se

Summary: Suicidal behavior is relatively well studied in younger populations but remains understudied among older adults despite significant risks linked to loneliness, physical ailments, and depression. This study addresses the gap by analysing suicide among older adults in Sweden using advanced statistical and machine learning techniques. Leveraging both static and longitudinal data – including suicide attempts, psychopharmaceutical prescriptions, medical conditions, and sociodemographic factors – we tackle the challenges associated with temporal dynamics and data censoring. The proposed approach integrates survival models such as Cox Proportional Hazards (CoxPH), Random Survival Forest (RSF), and Gradient Boosting Survival Analysis (GBSA) with sequential machine learning methods, specifically Long Short-Term Memory (LSTM) networks. Our results demonstrate that combining survival analysis with sequential models significantly enhances prediction accuracy by effectively capturing time-dependent patterns and censored observations.

Keywords: Suicide prediction, Survival analysis, Recurrent neural networks, AI in healthcare.

1. Introduction

Suicide prediction has been relatively well studied among younger populations, yet remains understudied in older adults. This gap exists despite clear evidence that suicide risk is also significant among the elderly. Factors contributing to suicidal behavior in older adults include physical health deterioration, increased loneliness due to loss of partners or friends, and psychological challenges such as depression, hopelessness, and diminished life purpose. Besides explicit suicide attempts, older adults commonly exhibit selfharm behaviors, such as intentional neglect of medication, self-starvation, or disregard for their physical health. This study aims to draw attention to suicide among older adults, a topic that remains underrepresented in research and public discourse compared to younger populations, despite the significant risks faced by the elderly.

We aim to predict suicidal behavior in older adults in Sweden by applying advanced statistical and machine learning methods. The dataset utilized includes both static and longitudinal data, covering recorded suicide cases and attempts, prescribed psychopharmaceutical medications, medical conditions, and sociodemographic characteristics. A significant challenge arises from the presence of censored samples in the dataset, as many individuals have not experienced the event (suicide) during the observation period, making it difficult to model outcomes accurately.

Survival analysis is particularly suited for addressing this challenge. Unlike traditional classification methods, survival analysis explicitly models not only whether the event (suicide) occurs but also precisely when it occurs. This temporal perspective is essential for timely interventions and more meaningful risk assessments. Additionally, survival analysis naturally accommodates censored data – instances where individuals have not experienced the event within the study period – effectively utilizing these incomplete observations, which traditional classification or regression models often discard or mishandle.

Another important aspect of our study is the incorporation of temporal patient data, such as prescription patterns of psychopharmaceutical medications and evolving medical conditions. While conventional, high-performing machine learning algorithms for tabular data (e.g., XGBoost, LightGBM, Random Forests) typically use ensemble methods, they often fail to effectively incorporate temporal dynamics. In contrast, sequential machine learning models, particularly Long Short-Term Memory (LSTM) networks, are explicitly designed to manage longitudinal data.

The primary contribution of this research is integrating the strengths of survival models – which handle censored data and continuous-time risk modelling – with sequential machine learning models that effectively leverage temporal information. Our results demonstrate that combining these approaches significantly enhances the predictive performance for suicide risk, simultaneously capitalizing on the temporal insights from recurrent neural networks and the robust handling of censored data from survival analysis.

Similar efforts in leveraging AI for healthcare applications highlight the growing potential of data-driven approaches to support early detection and intervention strategies in public health contexts [1].

2. Data

The dataset includes records of suicide cases and attempts, prescribed psychopharmaceutical medications, medical conditions, and sociodemographic characteristics. Given that the sequential models employed in this study inherently process temporal data, whereas traditional survival models lack this capability, the dataset was structured into two separate configurations.

The first setup, referred to as static data, includes monotonous demographic and sociodemographic information such as age, gender, income, education level, marital status, employment status, and household composition.

The second setup, termed longitudinal data, comprises pharmacy dispensation records of prescription medications and commodities associated with increased suicide and depression risk, limited to subjects aged 75 or older. This setup also includes records of suicidal behaviour – individuals who committed suicide, attempted suicide, or both, based on prior studies' definitions [2].

3. Methodology

Our proposed method integrates survival analysis outputs – including risk scores, survival functions (SF), and cumulative hazard functions (CHF) – into Long Short-Term Memory (LSTM) models, effectively incorporating censored observations. By embedding these survival-derived metrics within the LSTM framework, we enhance predictive performance in both classification and regression tasks, capturing critical temporal dynamics and addressing censoring issues central to suicide prediction.

We approach the prediction task through two setups: classification and regression. In both scenarios, the integration of survival and sequential models occurs as follows: survival analysis (SA) models generate predictions regarding time-to-event, which the LSTM network then utilizes as temporal features. Specifically, we employ Cox Proportional Hazards (CoxPH), Random Survival Forest (RSF), and Gradient Boosting Survival Analysis (GBSA) models. These models provide outputs such as risk scores, cumulative hazards, and survival probabilities, all serving as inputs for the LSTM network. The SA outputs are generated in static and dynamic forms. Static values, including risk scores and single estimations of CHF and SF, remain constant for each subject across all observations. Conversely, dynamic values of CHF and SF are recalculated for each subject at every drug prescription date, reflecting temporal changes in individual risk profiles.

The dataset used in this study was sourced from Swedish National Registers. The data was accessed through Statistikkonsulterna Väst AB under an agreement with the University of Gothenburg. The data was anonymized prior to researcher access, and no personally identifiable information was used in the analysis. Ethical approval for this study was obtained from Etikprövningsmyndigheten, Nr 2022-05846-02, Psykofarmaka, komorbiditet, sociala faktorer och risk för självmordsbeteende hos äldre.

4. Experiments and Results

For the experiments, a balanced dataset of 6,720 subjects was used, consisting of 3,360 individuals who had engaged in suicidal behavior and 3,360 censored cases without such events. This setup maintains a 50% censoring rate, acknowledging that high censoring rates significantly influence survival analysis model performance and interpretability [3]. Each subject was represented by 38 features, including 21 longitudinal features – of which 10 were explicitly engineered based on expert-provided ATC codes, and 17 static features reflecting sociodemographic and medical information.

The first experiment evaluates the proposed stacked architecture combining Survival Analysis (SA) models and an LSTM, compared to a baseline LSTM model without SA inputs, within a binary classification framework. The baseline serves as a control to assess the impact of incorporating SA-derived features. Both models utilize the same underlying LSTM structure; however, the proposed architecture additionally integrates static and dynamic SA-derived features. Specifically, static data informs the SA models, while dynamic data and SA outputs feed into the LSTM, ensuring a balanced comparison.

Model performance evaluation focuses primarily on accuracy since due to balanced classes, F1-score aligns closely with accuracy results. The classification architecture consists of a single LSTM layer to process sequential inputs, followed by a fully connected layer for classification. Detailed hyperparameters and experimental configurations are provided in Table 1.

Table 1. Training Parameters and Settings for the Classification LSTM.

Parameter	Value
Epochs	300
Learning Rate	1×10^{-3}
Hidden Size (Nodes)	50
LSTM Layers	3
Batch Size	32
Optimizer	Adam
Loss Function	CrossEntropy

Fig. 1 visually shows the mean accuracy and standard deviation of three survival models with static

and dynamic data. The horizontal dashed line also indicated the performance of the baseline LSTM model. The results indicate that the LSTM model, when combined with dynamic outputs from the Gradient Boosting Survival Analysis (GBSA) – i.e. cumulative hazard and survival probabilities – achieved superior performance compared to other models. Additionally, it is evident that models integrating survival analysis outputs consistently outperformed the baseline across all experiments.

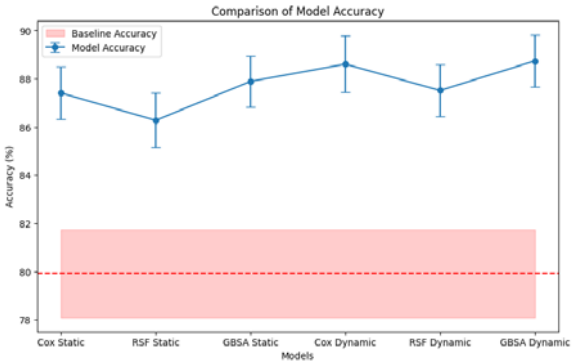


Fig. 1. Accuracy bar plot showcasing the aggregated classification results across different models.

For the regression experiment, we again compared a stacked survival model and LSTM architecture to a baseline LSTM model, but only dynamic feature outputs from the survival model were integrated into the regression setting. Static survival outputs were excluded; however, risk scores were employed to identify high-risk individuals (those above the 70th percentile). Censored samples were included by recalculating time-to-event (TTE) labels based on the median survival probability, leveraging the survival model for accurate labeling [4].

Performance metrics were calculated exclusively on patients who experienced an event to ensure fairness, given that our proposed method integrates censored samples directly into the LSTM model. The specific parameters and configurations for the LSTM architecture used in the regression experiments are detailed in Table 2.

Table 2. Training Parameters and Settings for the Regression LSTM.

Parameter	Value
Epochs	300
Learning Rate	1×10^{-2}
Hidden Size (Nodes)	50
LSTM Layers	5
Batch Size	32
Optimizer	Adam
Loss Function	L1Loss
Drop Out	0.5

Observing the superior performance of Gradient Boosting Survival Analysis (GBSA) in the first experiment over the other two survival models, we decided to exclusively use this model for the regression experiment. For the model evaluation Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Coefficient of Determination (R^2) are reported. Table 3 summarizes the regression results. To statistically evaluate the performance differences between the GBSA-integrated model and the baseline model, we conducted a t-test with an alpha level of 0.05. The resulting p-values were 0.013 for MAE, 0.017 for RMSE, and 0.012 for R^2, indicating statistically significant improvements across all evaluated metrics.

Table 3. Regression results comparing proposed model and the baseline (control).

Model	MAE	RMSE	R^2
GBSA	21.14±6.52	33.38±9.76	0.908± 0.051
Baseline	46.28±12.04	67.58±15.84	0.527± 0.199

4. Conclusions

This paper explored a novel approach to predicting suicidal behavior among older adults by combining survival analysis (SA) models and recurrent neural networks, particularly Long Short-Term Memory (LSTM). The results demonstrated that including censored data in the LSTM improved predictions, particularly when combined with GBSA. The study recognized several limitations, including underrepresented suicide cases and incomplete medication data. Future studies are encouraged to enhance data quality, examine alternative model combinations, and address data sparsity to further refine predictions.

Acknowledgements

We acknowledge Statistikkonsulterna Väst AB for providing access to the data and infrastructure used in this project. This work was supported by Leap for Life, an innovation center at Halmstad University in the frames of the Health Data Sweden project.

References

[1]. Z. Kharazian, et al., Aid4hai: Automatic idea detection for healthcare-associated infections from twitter, a framework based on active learning and transfer learning, in Crémilleux, B., Hess, S., Nijssen, S. (eds) Advances in Intelligent Data Analysis XXI. IDA 2023. Lecture Notes in Computer Science, Vol. 13876, *Springer*, Cham, 2023.

[2]. Q. Chen, et al. Predicting suicide attempt or suicide death following a visit to psychiatric specialty care: A machine learning study using Swedish national

registry data., *PLoS Medicine*, Vol. 17, Issue 11, 2020, e1003416.

[3]. M. Rahat, Z. Kharazian. Survloss: A new survival loss function for neural networks to process censored data. in *Proceedings of the PHM Society European Conference,* Vol. 8. No. 1. 2024, 7.

[4]. M. Rahat, et al., Bridging the gap: A comparative analysis of regressive remaining useful life prediction and survival analysis methods for predictive maintenance, in *Proceedings of the 4th Asia Pacific Conference of the Prognostics and Health Management,* Vol. 4. No. 1. 2023.

(046)

KIADEKU: Identification of Wound Types with AI

K. Majjouti [1], M. Tapp-Herrenbrueck [1], H. Pinnekamp [2], V. Priester [2], A. Brehmer [3], J. Kleesiek [2] and B. Hosters [1]
[1] Department of Nursing Development and Nursing Research, University Hospital Essen, Essen
[2] Department of Clinical Nursing Research and Quality Management, Hospital of the Ludwig Maximilian University (LMU) Munich, Munich, Germany
[3] Institute for Artificial Intelligence in Medicine (IKIM), University Hospital Essen, Essen, Germany
E-mail: Bernadette.Hosters@uk-essen.de

Summary: Distinguishing between pressure ulcers and incontinence-associated dermatitis can be challenging, even for experienced professionals, due to their similar appearance. Misclassification can delay treatment and lead to complications. This paper describes the development of a decision support system to help nurses implement evidence-based care interventions for these wounds. This is part of a research and development project aimed at improving the assessment and documentation of pressure ulcers and incontinence-associated dermatitis using an image-based artificial intelligence application in nursing. A mixed methods approach was used to develop a decision support matrix as part of the Clinical Decision Support System. The matrix defines conditions, interventions and decision rules. We created 27 decision rules to link conditions to interventions, categorising them as 'indicated', 'contraindicated' or 'not relevant'. The matrix has been integrated into a demonstrator that is being tested at the point of care to evaluate its effectiveness in clinical decision making.

Keywords: Clinical decision support system, Pressure ulcer, Incontinence-associated dermatitis, Artificial intelligence, Nursing practice.

1. Introduction

Pressure ulcers and incontinence-associated dermatitis (IAD) pose significant challenges for healthcare systems. The European prevalence of pressure ulcers has been documented as 13.1% [1, 2], while the prevalence of IAD is estimated to be 21.3% [3]. Both types of wounds have the capacity to cause significant physical discomfort, increase the risk of infection, prolong hospital stays and increase healthcare costs [4]. Furthermore, both conditions have a considerable impact on the quality of life of those affected [5].

1.1. Woundtyp Missclassification

A major challenge in clinical practice is the differentiation between pressure ulcers and IAD. Nurses, as the primary caregivers responsible for wound management, face considerable challenges in accurately differentiating between these two conditions [6]. This distinction is imperative, as the pathophysiological underpinnings of these conditions differ, necessitating distinct therapeutic interventions [7]. Pressure ulcers primarily result from prolonged pressure and shear forces, necessitating interventions that alleviate these mechanical stresses to restore tissue oxygenation [8]. Conversely, IAD stems from chemical irritation and maceration, requiring measures such as skin protection and more frequent care [9]. A misdiagnosis can result in inappropriate preventive and therapeutic strategies, compromising the quality of nursing care, and delayed wound healing [10]. Furthermore, it can distort the accuracy of pressure ulcer prevalence and incidence rates, as including IAD cases falsely inflates these statistics [11].

1.2. Clinical Decision Support Systems

The utilisation of clinical decision support systems (CDSS) has emerged as a promising approach to assist healthcare professionals in addressing this challenge, with the use of such systems becoming increasingly important in both medicine and nursing. For instance, such systems facilitate the early detection of risk and the implementation of targeted interventions through the provision of evidence-based decision support. Despite the existence of implementation barriers, such as a lack of acceptance or technical challenges, numerous evaluation studies have demonstrated the positive effects of CDSS on the quality of care [12-14]. Specifically, a range of CDSS for risk assessment and the prevention of pressure ulcers have been developed and evaluated for their effectiveness [15, 16]. Nevertheless, there remains a paucity of CDSS specifically designed to address the recognition, prevention, and management of IAD and pressure ulcers. Systems that explicitly address the clinically relevant distinction between pressure ulcers and IAD are particularly rare.

1.3. Relevance and Aim of the Study

The correct identification of IAD demands a deep understanding of its pathophysiology and clinical

manifestations [10]. Despite this, classification errors persist, with non-blanchable erythema often being confused with blanchable erythema or IAD, and IAD itself frequently being misclassified [11]. Addressing these challenges is essential for improving both clinical outcomes and the reliability of research data.

KIADEKU is a research and development project with the objective of enhancing the assessment and documentation of pressure ulcers and IAD using an image-based AI application in nursing care. The project also aims to support the implementation of evidence-based care practices. The University Hospital Essen, Germany, and the Hospital of the Ludwig-Maximilian-University Munich, Germany, cooperated on the project with the software development company scientist Ltd., Leipzig, Germany. The present paper offers a synopsis of the findings derived from the subproject, which pertain to the decision support module. This synopsis constitutes a segment of the broad research project's outcomes.

2. Methods

To develop the CDSS, we applied a mixed-methods approach. This approach was chosen to ensure that both evidence-based insights and nursing issues were integrated into the system's development.

2.1. Literature Review

In the main project, the development of the AI component was initiated with a review of relevant nursing guidelines from 2019 onwards in German or English, explicitly related to pressure ulcers or IAD. Furthermore, a literature review was conducted from July to August 2022 in the databases Medline (via PubMed), CINAHL and Cochrane Library. Additionally, a manual search of the reference lists of included publications was performed. The search strategy adhered to the specific research principle outlined in the RefHunter Protocol [17]. Publications of all study types from 2017 onwards in English or German were included. Studies were excluded if they investigated pressure ulcers caused by medical devices, focused on moisture-associated wounds in stoma therapy, or included study populations of children and adolescents under 19 years of age.
Further, the existing wound documentation systems of the participating university hospitals were comparatively analyzed.

2.2. Modified Delphi Survey

To incorporate the nursing perspective and ensure that the system could be effectively used at the point of care, a modified Delphi Survey with nursing professionals was conducted. In this survey nurse experts iteratively evaluated the results of the literature and documentation analysis to define a minimum data set (MDS) for the classification and categorization of pressure ulcer and IAD.
The results formed the basis for both the AI training in the main project [18] and the development of the decision support system in the current subproject.

2.3. Decision Table Technique

The method of the decision matrix is based on the description of decision tables by R. Irrgang [19]. A decision tables is a method for modelling complex systems. It is categorised as basic Computer-Aided Software Engineering (CASE) tool. The decision table technique does not require programming skills.
The integration of the matrix into the demonstrator from the KIADEKU project is currently underway. The evaluation of this integration is being conducted at the point of care as a core element of the KIADEKU project.

3. Results

We developed a decision support matrix that defines conditions, interventions, and decision rules. The conditions correspond to 123 wound-related criteria derived from the MDS, while the interventions map 33 evidence-based interventions sourced from the literature and refined. We established 27 decision rules to link conditions with interventions, categorizing them as 'indicated', 'contraindicated', or 'not relevant'.

3.1. Minimum Data Set

The MDS encompasses 18 wound-specific categories and four aetiological categories. These 22 categories collectively include 123 attributes, which constitute the mentioned conditions of the decision matrix. Table 1 provides a summary of each category along with examples of its attributes.

3.2. Decision Matrix

The decision matrix (Fig. 1) functions as a systematic tool that connects specific clinical conditions to appropriate interventions through predefined rules. These rules dictate which interventions should be triggered based on the presence of particular conditions or combinations of conditions. For instance, one decision rule within the matrix is designed to trigger the display of dressing options for deep wounds. This rule includes a strict contraindication for all superficial wounds, specifically those classified as IAD, and Category 1 or

Category 2 pressure ulcers. If a superficial wound of any of these types is documented with the demonstrator, the system will not display dressing options for deep wounds. This ensures that interventions are tailored to the specific wound type and conditions, preventing the system of inappropriate recommendations.

Table 1. Extract from the MDS

Category (Total N° of attributes)	Attributes
Anatomical Location (11)	Glutes Anal fold …
Wound typ (3)	PU IAD …
Categorization (8)	PU according to ICD-10 IAD according to GLOBIAD
Wound Size (4)	length width …
Wound bed (10)	Fibrin Granulation …
Wound bed characteristics (4)	Moist Dry …
Wound edge (10)	smooth bulging …
Wound morphology (4)	diffuse round/oval …
Wound pocket (4)	Yes Localisation …
Periwound area (11)	dry moist …
Periwound colour (3)	livid flushed …
Pain (3)	Yes No not assessable
Itching (3)	
Odor (3)	
Exsudate level (5)	none low …
Exsudate quality (3)	purulent not purulent …
Exsudate colour (8)	pink/red green …
Signs of infection (6)	Redness/Rubor Overheated/Calor …
Mobility (5)	absent mostly absent mostly complete complete not assessable
Urinary continence (5)	
Feacal continence (5)	
Sensory ability (5)	

Fig. 1 gives an example for a decision matrix. Visual markeres are used to flag contraindications and indications. A red box signifies a strict contraindication, while a green box represents an indication. For a rule to be executed and the corresponding intervention to be triggered, all specified indications must be met. However, if any contraindication is present, the rule will not be executed, and the corresponding intervention will not be displayed. The example also illustrates that a condition can involve multiple attributes. In particular, rule R02 specifies as indication that at least two attributes from the 'Signs of Infection' category must be present for it to take effect. This is necessary in order to define an infection, as at least two signs of infection to be present.

			Decision Rule		
Condition	Category	Attributes	R01	R02	R03
B01	Woundtyp	PU	-	-	-
B02		IAD	-	-	-
B03		Both			
B04	Classification	I	-	-	-
B05		II	-	-	-
B06		III	-	-	-
B07		IV	-	-	-
B08		1A	-	-	-
B09		1B	-	-	-
B10		2A	-	-	-
B11		2B	-	-	-
B12	Signs of Infection	Redness	≥2	≥2	≥2
B13		Overheated			
B14		Swelling			
B15		Pain			
B16		Functional impairment			
B17		none	-	-	-
Interventions					
I01		Description of Intervention 01	x		
I02		Description of Intervention 02		x	
I03		Description of Intervention 03			x

Fig. 1. Example of a decision matrix; red = contraindication, green = indication, - = not relevant,

The 33 interventions encompass a comprehensive set of evidence-based practices designed to guide effective wound care. They include detailed instructions on selecting appropriate wound dressings tailored to the specific wound types and stages, ensuring optimal healing environments. Additionally, the interventions provide wound cleansing techniques that promote tissue viability and prevent infection. Specific treatment protocols are outlined for various wound bed conditions and tissue types, such as managing infections, fibrin or necrosis, each requiring distinct approaches to facilitate healing. Skin protection measures are highlighted to prevent wound deterioration. Furthermore, the interventions also address the underlying conditions that inhibit healing, such as limited mobility or incontinence.

3.3. Pilot Testing

The decision matrix was used for integration into a demonstrator, which is currently undergoing pilot testing at the point of care to evaluate its effectiveness in clinical decision-making. The demonstrator is accessible on tablets equipped with cameras, facilitating wound documentation.

During the wound care process included in the pilot study, the nurse captures a wound image using the demonstrator. The AI-based system then analyses the documentation including key patient information, such as mobility status and degree of incontinence, to

predict the wound type and category, each with a probability value. The nurse reviews these predictions and has the option to revise them as needed, ensuring the accuracy of the information for clinical decision making. Once the documentation is completed, the system provides evidence-based nursing interventions tailored to the documented data. A preliminary interim evaluation prompted revisions to the software, particularly concerning the rules, with further insights expected after the final evaluation. Should software development continue to evolve with regard to medical devices, it will be essential to define the logic and implementation framework for the decision matrix in more detail.

5. Conclusions

The decision matrix serves as a structured framework that systematically connects each wound condition to its corresponding intervention. It operates by defining a set of decision rules that specify the condition or combination of conditions that must be present for a particular intervention to be initiated. This ensures that each intervention is triggered only when the appropriate clinical circumstances are met, thereby supporting health care professionals in clinical decision making.

This problem-solving method is especially appropriate for the systematic analysis of complex situations, for the precise description of processes, and for the clear and unambiguous documentation of information. The last two points in particular have been of great benefit to the interprofessional exchange between health informatics specialists, nurse scientists and software developers in the KIADEKU project.

The development of a decision support matrix that meets nursing needs, evidence-based literature and software development requirements is a complex task. The KIADEKU project demonstrates that a mixed methods approach can help to address this complexity and effectively integrate workflow guidelines to streamline the care process with the aim of enhancing efficiency and consistency in clinical practice. The iterative involvement of all relevant professions is crucial.

References

[1]. Z. Moore, P. Avsar, L. Conaty, D. H. Moore, D. Patton, T. O'Connor, The prevalence of pressure ulcers in Europe, what does the European data tell us: a systematic review, *Journal of Wound Care,* Vol. 28, Issue 11, 2019, pp. 710-719.

[2]. K. B. Al Mutairi, D. Hendrie, Global incidence and prevalence of pressure injuries in public hospitals: A systematic review, *Wound Medicine,* Vol. 22, 2018, pp. 23-31.

[3]. M. Gray, K. K. Giuliano, Incontinence-Associated Dermatitis, Characteristics and Relationship to Pressure Injury: A Multisite Epidemiologic Analysis, *Journal of Wound Ostomy & Continence Nursing*, Vol. 45, Issue 1, 2018, pp. 63-67.

[4]. Z. E. H. Moore, J. Webster, Dressings and topical agents for preventing pressure ulcers, *Cochrane Database of Systematic Reviews,* Issue 12, 2018, CD009362.

[5]. Cunich M., Barakat-Johnson M., Lai M., Arora S., Church J., Basjarahil S., et al., The costs, health outcomes and cost-effectiveness of interventions for the prevention and treatment of incontinence-associated dermatitis: A systematic review. *International Journal of Nursing Studies,* 129, 2022, 104216.

[6]. D. Beeckman, Incontinence-Associated Dermatitis (IAD) and Pressure Ulcers: An Overview, in Science and Practice of Pressure Ulcer Management (M. Romanelli, M. Clark, A. Gefen, G. Ciprandi, Eds.), *Springer London,* 2018, pp. 89-101.

[7]. J. Dissemond, B. Assenheimer, V. Gerber, M. Hintner, M. J. Puntigam, N. Kolbig, et al., Moisture-associated skin damage (MASD): A best practice recommendation from Wund-D.A.CH., *J Dtsch Dermatol Ges,* Vol. 19, Issue 6, 2021, pp. 815-825.

[8]. D. Beeckman, K. Van den Bussche, P. Alves, M. C. Arnold Long, H. Beele, G. Ciprandi, et al., Towards an international language for incontinence-associated dermatitis (IAD): Design and evaluation of psychometric properties of the Ghent Global IAD Categorization Tool (GLOBIAD) in 30 countries, *The British Journal of Dermatology,* 178, 6, 2018, pp. 1331–1340.

[9]. E. Haesler, Prevention and Treatment of Pressure Ulcers/injuries: Clinical Practice Guideline, *The International Guideline,* 2019.

[10]. J. Kottner, N. Kolbig, A. Bültemann, J. Dissemond, Incontinence-associated dermatitis: a position paper, *Der Hautarzt,* Vol. 71, 2020, pp. 46-52.

[11]. T. Defloor, L. Schoonhoven, K. Vandenbroeck, J. Weststrate, D. Myny, Reliability of the European Pressure Ulcer Advisory Panel classification system, *Journal of Advanced Nursing,* Vol. 54, Issue 2, 2006, pp. 189-198.

[12]. C. Thompson, T. Mebrahtu, S. Skyrme, K. Bloor, D. Andre, A. M. Keenan, et al., The effects of computerised decision support systems on nursing and allied health professional performance and patient outcomes: A systematic review and user contextualisation, *Health and Social Care Delivery Research,* 2023, pp. 1–85.

[13]. T. F. Mebrahtu, S. Skyrme, R. Randell, A.-M. Keenan, K. Bloor, H. Yang, et al., Effects of computerised clinical decision support systems (CDSS) on nursing and allied health professional performance and patient outcomes: a systematic review of experimental and observational studies*, BMJ Open,* Vol. 11, Issue 12, 2021, p. e053886.

[14]. S. Sariköse, S. Ş. Çelik, The effect of clinical decision support systems on patients, nurses, and work environment in ICUs: a systematic review, *CIN: Computers, Informatics,* Nursing, Vol. 42, Issue 4, 2024, pp. 298-304.

[15]. J. J. Ting, A. Garnett, E-health decision support technologies in the prevention and management of pressure ulcers: a systematic review, *CIN: Computers, Informatics, Nursing,* Vol. 39, Issue 12, 2021, pp. 955-973.

[16]. S. Mäki-Turja-Rostedt, M. Stolt, H. Leino-Kilpi, E. Haavisto, Preventive interventions for pressure ulcers in long-term older people care facilities: A systematic review, *Journal of Clinical Nursing,* Vol. 28, Issues 13-14, 2019, pp. 2420-2442.

[17]. T. Nordhausen, J. Hirt, Manual zur Literaturrecherche in Fachdatenbanken, *Martin-Luther-Universität Halle-Wittenberg,* Version 3, 2020 (in German).

[18]. A. Brehmer, C. Seibold, K. Majjouti, M. Tapp-Herrenbrueck, H. Pinnekamp, V. Rentschler, et al., Fine-Grained Classification of Pressure Ulcers and Incontinence-Associated Dermatitis Using Multimodal Deep Learning: Algorithm Development and Validation Study, *JMIR Medical Informatics*, in press.

[19]. R. Irrgang, Entscheidungstabellen-Technik: Entscheidungstabellen erstellen und analysieren, *Expert Verlag,* 1995 (in German).

(047)

Reducing Nurses' Workload with an AI Speech Assistant for Documentation

K. Schwabe [1], D. Ferizaj [2] and S. Neumann [2]
[1] Voize GmbH, Dircksenstraße 47, 10178 Berlin, Germany
[2] Charité - Universitätsmedizin Berlin., Department of Geriatrics and Medical Gerontology, Reinickendorfer Straße 61, 13347 Berlin, Germany
E-mail: katja@voize.de, susann.neumann@charite.de, drin.ferizaj@charite.de

Summary: In this paper, we examine the integration of AI speech assistants in nursing to assess workload reduction from both qualitative and quantitative perspectives. Nurses in long-term care face high work pressure due to extensive documentation and multitasking. Our interdisciplinary approach combines qualitative interviews, a time-motion study, and a quantitative questionnaire to capture changes in nursing workload. Qualitative findings reveal improved efficiency, reduced documentation burden, and enhanced information access, while preliminary time-motion data suggest potential time savings associated with speech assistant usage. Assessments of work strain and technostress complement these measures, underscoring that no single method can fully capture the impact of AI on nursing workload; all perspectives must be considered collectively. Our results highlight the need to integrate technical, psychological, and management perspectives to guide future AI-driven innovations in healthcare.

Keywords: Nursing workload, Nursing documentation, Long-term care, AI speech assistant, Time-Motion-Study.

1. Introduction

Nurses, particularly those working in long-term care, report higher levels of workload compared to other occupational groups [1, 2]. They work under high time pressure, have limited time to complete tasks, and frequently work overtime with irregular or missing breaks [1, 3]. Additionally, nurses perform up to 40% of their tasks in a multitasking manner [4]. These high demands correlate with increased levels of fatigue and burnout [5, 3].

Another contributing stressor is the substantial amount of administrative tasks in nursing: documentation duties can account for 25 to 30% of working time [6, 7]. Thus, nurses may spend more time on nursing documentation than on direct patient care [8]. Furthermore, the high documentation workload is associated with decreased job satisfaction and an increased risk of burnout [9, 10].

The use of mobile documentation systems that can be utilized at the patient's bedside has the potential to reduce workload while simultaneously improving efficiency and job satisfaction [11]. In particular, speech assistants can accelerate the documentation process and enhance its accuracy; however, they require initial investments and extensive training [12, 13].

1.1. Operationalization of Nursing Workload

Measuring workload reduction via artificial intelligence (AI) speech assistants requires a standardized operationalization of workload-a construct that varies across disciplines. Research on AI in nursing spans multiple fields, including nursing and health sciences, information systems (IS) and health information technology (HIT), as well as work and organizational psychology. In nursing science literature, workload is typically operationalized by examining the amount of nursing time, the level of nursing competency, the weight of direct patient care, the degree of physical exertion, and the complexity of care [14]. In contrast, psychological research conceptualizes workload using constructs such as job demands [15], while IS and HIT research integrates technology-related variables—employing concepts like technostress—to assess the impact of technology on employees [16, 17].

To date, no existing model or assessment comprehensively captures changes in workload among nursing staff resulting from AI-driven systems, particularly speech assistants, while fully accounting for the specific requirements of the care setting.

1.2. Study Objectives and Approach

This overview aims to examine workload changes associated with an AI speech assistant from an interdisciplinary perspective and to discuss the associated challenges. The project offers a methodological overview of key phases within the PYSA-Nursing Documentation with Hybrid Speech Assistant project, specifically focusing on those that measured workload-related changes.

2. Methods

2.1. Speech Assistant ‘voize’

The speech assistant ‘voize’ is a software-based application that enables nursing documentation to be

recorded via a smartphone. It employs a continuously trained AI that processes voice inputs locally, assigns the content to the appropriate categories, and automatically generates structured documentation entries. These entries are reviewed and approved by nursing staff before being integrated into the existing documentation system.

Developed in close collaboration with nursing professionals, the application is continuously refined through a collaborative, user-centered design process, and its underlying models are regularly trained with voluntarily submitted and anonymized documentation data to ensure ongoing improvements and adaptation to the linguistic characteristics and professional terminology used in nursing.

Technically, the system converts spoken language into text in real time using an end-to-end streaming automatic speech recognition (ASR) model operating as a single, continuous unit. In conjunction with this, a dynamically orchestrated set of specialized natural language processing (NLP) models is activated as needed to handle specific tasks—such as understanding context, extracting key information, or structuring data—ensuring that all components work together seamlessly to deliver the supported features.

2.2. Qualitative Study: Perceived Impact

A qualitative study was conducted through semi-structured interviews with ten caregivers and non-participant observations in two German nursing facilities between August and October 2022. Thematic analysis was used to assess acceptance, usability, and perceived impact on work-related aspects.

2.3. Pilot Study: Time Burden

With the aim of quantifying the temporal burden on nursing staff in inpatient long-term care, a pilot study was developed in preparation for a subsequent evaluation study. As part of the study with a pre-post design, time measurements were conducted at two measurement points (t_0;t_1) in four residential long-term care facilities across Germany between February and September 2024. The time required for documentation tasks was measured both before and after the implementation of the speech assistant ‘voize.’ The sample consisted of 11 registered nurses.

Because time-motion studies require substantial methodological and personnel efforts as well as precise planning and targeted resource allocation, this pilot study was conducted to assess feasibility and identify preliminary trends that will serve as a basis for a more comprehensive examination of the speech assistant's effectiveness.

2.4. Quantitative Study: Time Burden, Work Strain and Technostress

A time-motion study with a larger sample group and additional assessments is currently being conducted to evaluate the time savings achieved through AI-supported documentation and its relationship to workload relief.

In this study, the efficiency of a speech assistant for nursing documentation is examined in a pre-post comparison using paired samples. The null hypothesis posits that the use of the speech assistant does not lead to significant time savings in documentation, whereas the alternative hypothesis assumes a meaningful reduction in documentation time. Based on a preliminary anonymous online survey with 50 nurses and the literature-based assumption that approximately 30% of working time is devoted to nursing documentation [6, 7], an a priori power analysis (Cohen’s $d = 0.5$, $\alpha = 0.05$, power = 0.8) indicated a minimum sample size of 34 participants, which was increased to 40 to account for an estimated drop-out rate of 15% (e.g., due to short-term shift changes and sick leave).

With the aim of measuring changes in quantitative work-related strain and the influence of the subjective perception of the speech assistant, a questionnaire based on standardized assessments is used. Job demands, such as interruptions, information deficits, and frequent disruptions, are assessed using a shortened version of the FGBU (Questionnaire for the Risk Assessment of Psychological Workload) [18].

The subjective perception of technology can be viewed from two perspectives according to the Holistic Technostress Model [17]. Technology can be perceived as predominantly challenging or hindering. A potential influence of this perception on the experience of workload is examined using a shortened version of the assessment of the Holistic Technostress [17]. All participants in the time-motion study are asked to complete the assessment.

3. Results

3.1. Qualitative Study: Perceived Impact

Nurses reported increased efficiency, reduced documentation workload, and improved access to information. The AI speech assistant enhanced transparency and continuity of care but faced initial barriers such as concerns over speech recognition accuracy, data privacy, and early-stage usability challenges. Social influences, training, and continuous multimodal technical support were critical for adoption, and nurses' feedback shaped system refinements, highlighting the importance of co-creation in optimizing usability.

3.2. Pilot Study: Time Burden

The average documentation time at t_0 was $M = 64$ minutes ($SD = 20.8$, $n = 8$), whereas at t_1 it was $M = 24.8$ minutes ($SD = 7.98$, $n = 6$). The difference in mean values amounted to $M = 39.2$ minutes, representing a reduction of 61.2%.

The research was challenged by a significant participant dropout due to a high incidence of sick

leaves and the requirement for team members to cover shifts unexpectedly. This resulted in a small sample size, which in turn limits the statistical power of the findings.

3.3. Quantitative Study: Time Burden, Work Strain and Technostress

The final results of the time measurements and assessments will be published upon the study's completion.

As of now, 37 nurses have taken part in the study prior to the implementation of 'voize,' with 14 of them participating afterward.

4. Conclusions

The present study underscores the necessity of adopting an interdisciplinary perspective to understand the workload construct in nursing. With the advent of speech assistants like 'voize,' previously distinct research fields now intersect, making it essential to examine nursing workload from multiple angles. Interview data indicate that nurses experience benefits such as reduced documentation workload and enhanced access to information, while preliminary results from a time-motion study suggest an association between 'voize' usage and time savings.

This blended methodological approach demonstrates that no single study component can capture all dimensions of workload change. Although investigations into acceptance and usability address IS and HIT concerns and time measurements offer management insights, the subjective experiences of nursing staff remain crucial.

Ultimately, the forthcoming results are expected to shed further light on these interrelationships.

Acknowledgements

We would like to thank the German Federal Ministry for Education and Research as the responsible funding body. In addition, we would like to thank the participating facilities, and especially the nursing staff, who contributed their insights and expertise to this study.

References

[1]. T. Wirth, N. Ulusoy, H.-J. Lincke, A. Nienhaus, and A. Schablon, Psychosoziale Belastungen und Beanspruchungen von Beschäftigten in der stationären und ambulanten Altenpflege, *ASU Arbeitsmedizin, Sozialmedizin, Umweltmedizin*, Vol. 52, 2017, pp. 662–669.

[2]. H. Hasson and J. E. Arnetz, Nursing staff competence, work strain, stress and satisfaction in elderly care: a comparison of home-based care and nursing homes, *Journal of Clinical Nursing*, Vol. 17, No. 4, 2008, pp. 468–481.

[3]. C. Dall'Ora, O. Z. Ejebu, J. Ball, and P. Griffiths, Shift work characteristics and burnout among nurses: cross-sectional survey, *Occupational Medicine (Oxford, England)*, Vol. 73, No. 4, 2023, pp. 199–204.

[4]. P.-Y. Yen, M. Kelley, M. Lopetegui, A. L. Rosado, E. M. Migliore, E. M. Chipps, and J. Buck, Understanding and Visualizing Multitasking and Task Switching Activities: A Time Motion Study to Capture Nursing Workflow, in *AMIA Annual Symposium Proceedings*, 2016, pp. 1264–1273, [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5333222/.

[5]. E. Diehl, S. Rieger, S. Letzel, A. Schablon, A. Nienhaus, L. C. Escobar Pinzon, and P. Dietz, The relationship between workload and burnout among nurses: The buffering role of personal, social and organisational resources, *PloS One*, Vol. 16, No. 1, 2021, p. e0245798.

[6]. E. Schenk, R. Schleyer, C. R. Jones, S. Fincham, K. B. Daratha, and K. A. Monsen, Time motion analysis of nursing work in ICU, telemetry and medical-surgical units, *Journal of Nursing Management*, Vol. 25, No. 8, pp. 640–646, 2017.

[7]. P.-Y. Yen, M. Kellye, M. Lopetegui, A. Saha, J. Loversidge, E. M. Chipps, L. Gallagher-Ford, and J. Buck, Nurses' time allocation and multitasking of nursing activities: A time motion study, in *Proceedings of the AMIA Annual Symposium*, 2018, pp. 1137–1146. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC6371290/

[8]. A. Momenipur and P. R. Pennathur, Balancing documentation and direct patient care activities: A study of a mature electronic health record system, *International Journal of Industrial Ergonomics*, Vol. 72, 2019, pp. 338–346.

[9]. A. J. Moy, J. M. Schwartz, R. Chen, S. Sadri, E. Lucas, K. D. Cato, and S. C. Rossetti, Measurement of clinical documentation burden among physicians and nurses using electronic health records: A scoping review, *Journal of the American Medical Informatics Association (JAMIA)*, Vol. 28, No. 5, 2021, pp. 998–1008.

[10]. E. Gesner, P. C. Dykes, L. Zhang, and P. Gazarian, Documentation burden in nursing and its role in Clinician Burnout Syndrome, *Applied Clinical Informatics*, Vol. 13, No. 5, 2022, pp. 983–990.

[11]. F. Ehrler, D. T. Y. Wu, P. Ducloux, and K. Blondon, A mobile application to support bedside nurse documentation and care: A time and motion study, *JAMIA Open*, Vol. 4, No. 3, 2021, article ooab046.

[12]. M. P. Koivikko, T. Kauppinen, and J. Ahovuo, Improvement of report workflow and productivity using speech recognition - a follow-up study, *Journal of Digital Imaging*, Vol. 21, 2008, pp. 378–382.

[13]. J. Joseph, Z. E. H. Moore, D. Patton, T. O'Connor, and L. E. Nugent, The impact of implementing speech recognition technology on the accuracy and efficiency (time to complete) clinical documentation by nurses: A systematic review, *Journal of Clinical Nursing*, Vol. 29, No. 13-14, 2020, pp. 2125–2137.

[14]. M. G. Alghamdi, Nursing workload: a concept analysis, *Journal of Nursing Management*, Vol. 24, No. 4, 2016, pp. 449–457.

[15]. A. B. Bakker and E. Demerouti, The Job Demands Resources model: state of the art, *Journal of Managerial Psychology*, Vol. 22, No. 3, 2007, pp. 309-328.

[16]. T. S. Ragu-Nathan, M. Tarafdar, B. S. Ragu-Nathan, and Q. Tu, The consequences of technostress for end users in organizations: Conceptual development and empirical validation, *Information Systems Research*, Vol. 19, No. 4, 2008, pp. 417-433.

[17]. C. B. Califf, S. Sarker, and S. Sarker, The Bright and Dark Sides of Technostress: A Mixed-Methods Study Involving Healthcare IT, *MIS Q.*, Vol. 44, 2, 2020, pp. 809-856.

[18]. J. Dettmers and A. Krause, Der Fragebogen zur Gefährdungsbeurteilung psychischer Belastungen (FGBU), *Zeitschrift für Arbeits- und Organisationspsychologie*, Vol. 64, No. 2, 2020, pp. 99–119.

(040)

Predicting Posterior Circulation Stroke using Deep Learning on CT Brain Non-contrast

P. Supholkhan [1], P. Looareesuwan [1], C. Puttanawarut [2], B. Kijsanayotin [1], P. Tunlayadechanont [3], and A. Thakkinstian [1]
[1] Department of Clinical Epidemiology and Biostatistics, Faculty of Medicine Ramathibodi Hospital, Mahidol University 270 RAMA VI Road., Rachathevi, Bangkok 10400, Thailand
[2] Chakri Naruebodindra Medical Institute, Faculty of Medicine Ramathibodi Hospital, Mahidol University, 111 Suvarnabhumi Canal Road, Bang Pla Subdistrict, Bang Phli, Samut Prakan, 10540, Thailand
[3] Department of Diagnostic and Therapeutic Radiology, Faculty of Medicine Ramathibodi Hospital, Mahidol University 270 RAMA VI Road., Rachathevi, Bangkok 10400, Thailand
Tel.: + 66 02-201-0833
E-mail: pasit.suh@student.mahidol.edu

Summary: CT brain non contrast (NCCT) is an early diagnosis tools for detect ischemic stroke. early detection remains challenging due to the limitation of CT to visualizing lesion in early onset, especially in posterior circulation stroke (PCS), as well as availability of additional CT (CTA, CTP, FUCT, CT contrast), MRI in primary care of stroke fast track protocol. This study explores the use of deep learning (DL) models to predict posterior circulation stroke from NCCT. A dataset of 17,085 NCCT radiology report was processed. A 3-Dimensinal convolutional Neural Network (3D-CNN) and 2-Dimensinal convolutional neural network with Long Short-Term Memory Network (2DCNN-LSTM) models were trained and evaluated. The model exhibited strong predictive performance, attaining an accuracy of 72%, sensitivity of 74%, specificity of 72%, and an F1 score of 73%. Additionally, the ROC AUC of 73% and PR AUC of 79.7% further validated its efficacy. These results support the potential of deep learning-based NCCT analysis as an effective tool for early detection of PCS, facilitating early diagnosis for improved management of PCS.

Keywords: CT brain non contrast, Ischemic stroke, Posterior circulation stroke, MRI, Echocardiogram, Deep learning, CNN, LSTM.

1. Introduction

Stroke is a serious and life-threatening cerebrovascular condition, ranking as the second leading cause of death globally, with 13.7 million new cases and 2.7 million deaths annually [1, 2]. The annual cost of stroke is estimated to be around $34 billion worldwide. Compared to anterior stroke, posterior circulation stroke (PCS), which affects the posterior region of the brain that controls movement, balance, and coordination, is less prevalent but has a considerably higher rate of misdiagnosis [3]. This is partially because of the difficulties in detecting PCS with existing imaging techniques, such as NCCT, which has a low sensitivity in detecting this kind of stroke because of the brain's complex anatomy and bone density [4]. While other imaging techniques like magnetic resonance imaging (MRI) and computed tomography perfusion (CTP) offer higher sensitivity (CTP 76.6-80%, MRI 83-87%) compared to NCCT (sensitivity 33.3-54.8%) [5, 6], they are time-consuming and less accessible in emergency settings.

Artificial intelligence (AI) models, especially deep learning (DL) using convolutional neural networks (CNN), hold considerable potential for enhancing stroke detection in order to overcome these constraints. The detection of PCS from CT scans may be improved by deep learning algorithms, which have shown greater performance in medical image classification tasks [7]. Prior models, using radiomics, have demonstrated some promise but have to perform feature engineering with limited data [8]. Most research suggests favouring DL over radiomics for medical imaging applications. DL models are often preferred due to their ability to automatically extract complex features reduce feature engineering process. In order to increase diagnostic speed and accuracy, this study attempts to create a deep learning model that is specially trained on NCCT images to predict PCS. This method may enable earlier identification, more effective treatment, and improved patient outcomes by comparing the model's output with human radiologists to support their decision-making, thereby improving quality of life.

2. Objective

This study aims to develop and optimize DL models using NCCT images to detect posterior circulation stroke.

3. Method

A total of 17,085 NCCT images and reports containing raw slices of DICOM images from 10,672

patients in the Faculty of Medicine Ramathibodi Hospital Radiology Database were included in this study after applying multiple exclusion criteria (e.g., regex labeling, incomplete and artifact containing report, DICOM images losses) show in Fig. 1. Each NCCT was matched with a ground truth report (additional CT, MRI report) taken within a 7-days interval, which served as the gold standard for identifying ischemic stroke lesion [9, 10]. A total of 2,217 from 17,085 NCCT (14.9%) were identified as PCS according to the radiology report results. All DICOM images were convert to 64*64*64 shape and underwent preprocessing for Artifact removal, Skull stripping, Image registration, HU Clipping, and Normalization. The data was split into training (70%), validation (15%), and test (15%) sets for model evaluation. During the training process, data augmentation techniques were employed to mitigate overfitting. The model's performance was subsequently assessed on unseen test data for the prediction of PCS.

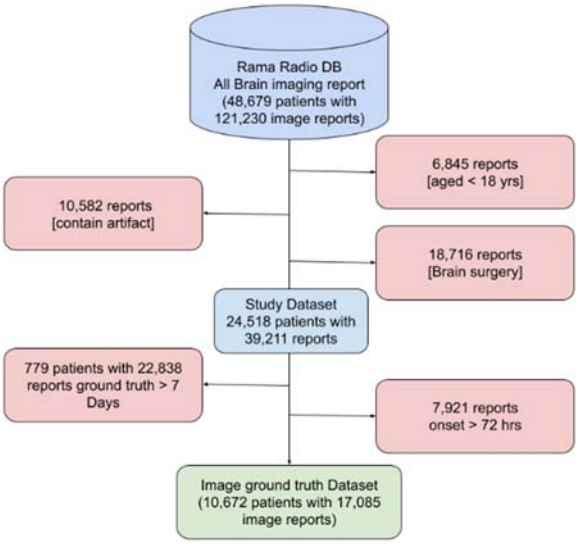


Fig. 1. Data flow pipeline.

4. Results

According to Table 1, the prediction outcomes showed impressive results of 3D-CNN in most evaluation metrics i.e. accuracy, sensitivity, specificity, and F1 score of 72%, 74%, 72%, and 73% respectively. In addition, ROC AUC-ROC and AUC-PRC were shown remarkably high as well with the scores of 80.1% and 85.1% respectively in Fig. 2. The utilization of Explainable AI (XAI), specifically Gradient-weighted Class Activation Mapping (Grad-CAM) [11], enables the visualization of critical regions associated with cerebrovascular pathology. In Fig. 3, Grad-CAM highlights the affected areas corresponding to cerebellar infarction, providing insights into the model's decision-making process. Similarly, Fig. 4 demonstrates the localization of brainstem infarction, as identified through Grad-CAM, further enhancing interpretability and diagnostic transparency in the model's predictions.

Table 1. Prediction Performance of the 3D-CNN, 2D-LSTM.

Metric	Models	
	3D-CNN	**2D-LSTM**
Accuracy (%)	72.2	58
Recall (%)	74	59.5
Specificity (%)	72	65
Precision (%)	72	64
F1-score (%)	73	55
ROC_AUC (%)	80.1	59.4
PR_AUC (%)	85.1	74.1

5. Discussion

This study demonstrates the feasibility of using deep learning models, particularly 3D-CNN, for detecting posterior circulation PCS—a condition that is often challenging to diagnose clinically [6]. The integration of Grad-CAM enhances the model's interpretability by providing visual explanations, helping healthcare professionals understand the specific regions influencing the model's predictions.

The model's performance is comparable to prior studies that used MRI or CTP as input modalities. However, NCCT is far more accessible, particularly in low- and middle-income countries where advanced imaging techniques such as MRI and CTP are limited. By improving PCs detection using NCCT, this approach could enable earlier diagnosis in resource-constrained settings, reducing the need for transfers to tertiary care centers and ultimately improving patient outcomes.

Additionally, this model has the potential to support clinical decision-making, especially in settings with a shortage of experienced radiologists. AI-assisted detection can enhance diagnostic decision-making, streamline workflow efficiency, and ensure that high-risk cases receive timely intervention. Future work should focus on validating the model across multiple institutions and integrating it into clinical workflows to assess its real-world impact on stroke diagnosis and management.

However, this study has several notable limitations. First, the performance of the best 3D-CC model (72.2% accuracy, 80.1% ROC-AUC) indicates potential for enhancement, which could be addressed through fine-tuning on pretrained models or employing advanced architectures such as Vision Transformer (ViT) [12] and VisionLLM [13]. Second, the dataset is derived from a single institution (Mahidol University), which restricts the generalizability of the findings; therefore, multi-center validation is essential to improve external validity. Finally, the study does not include a direct comparison with expert radiologists. Conducting a reader study to benchmark the AI model's performance against human experts would further substantiate its clinical applicability.

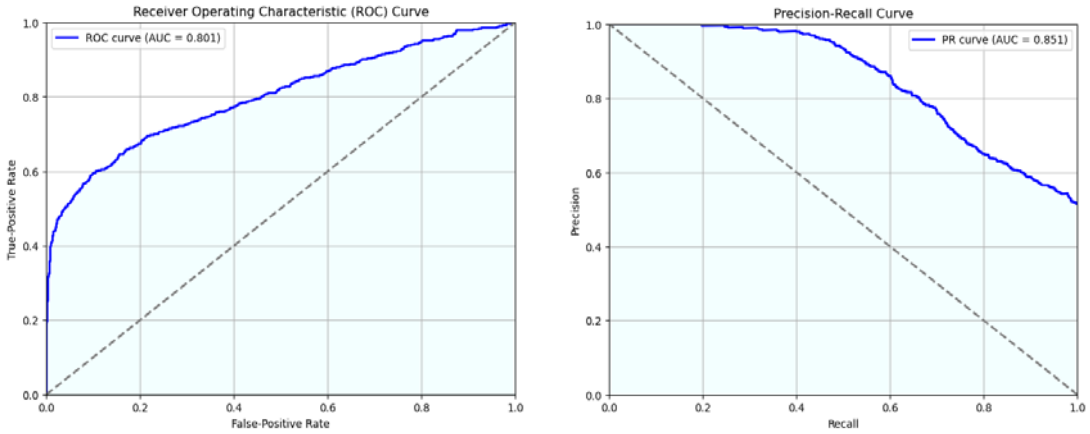


Fig. 2. ROC curve and AUPRC curve for 3D-CNN.

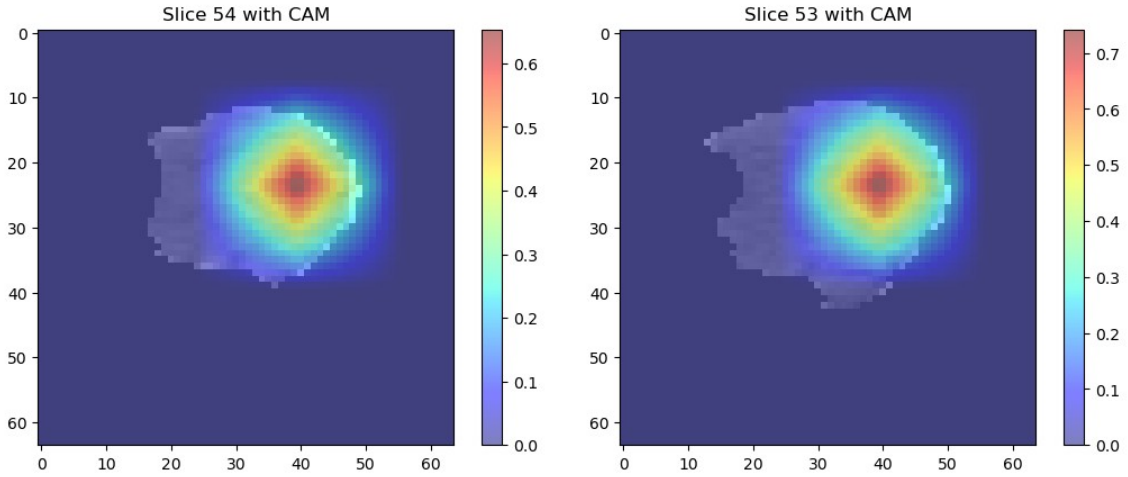


Fig. 3. Grad-CAM Visualization Highlighting Cerebellar Infarction.

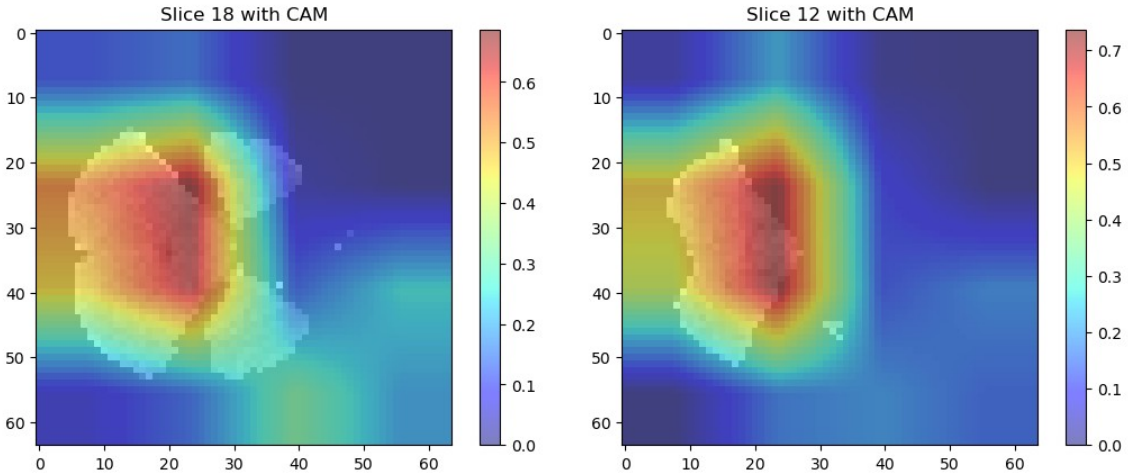


Fig. 4. Grad-CAM Visualization Identifying Brainstem Infarction.

6. Conclusions

Deep learning models, particularly 3D-CNN, can be utilized for PCS prediction on NCCT images without the necessity for feature engineering. This approach facilitates the early identification of potential PCS cases, thereby supporting timely diagnosis, referrals for specialized treatment such as RtPA, and further investigations prior to the progression of the disease.

Acknowledgements

This study was approved by the Human Research Ethics Committee of the Faculty of Medicine, Ramathibodi Hospital, Mahidol University (COA.MURA2024/445).

References

[1]. Campbell, B. C. V. et al., Ischaemic stroke, *Nat. Rev.*, Dis. Primer 5, 2019, art. 70.

[2]. Johnson, Catherine Owens et al., Global, regional, and national burden of stroke, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016, *Lancet Neurol.,* 18, 2019, pp. 439–458.

[3]. Arch, A. E. et al., Missed Ischemic Stroke Diagnosis in the Emergency Department by Emergency Medicine and Neurology Services, *Stroke*, 47, 3, 2016, pp. 668–673.

[4]. Schorr, E. M., Brandstadter, R., Jin, P., Stahl, C. & Krieger, S., Training in Neurology: Diagnostic Accuracy Among Neurology Residents: The 'Close the Loop' Project, *Neurology*, 96, 13, 2021, pp. e1804–e1808,

[5]. Sporns P., Schmidt R., Minnerup J., Dziewas R., et al., Computed Tomography Perfusion Improves Diagnostic Accuracy in Acute Posterior Circulation Stroke, *Cerebrovasc Dis.,* 41, 5-6, 2016, pp. 242-247.

[6]. Hwang, D. Y., Silva, G. S., Furie, K. L. & Greer, D. M. Comparative Sensitivity of Computed Tomography vs. Magnetic Resonance Imaging for Detecting Acute Posterior Fossa Infarct, *J. Emerg. Med*., 42, 2012, pp. 559–565.

[7]. Yadav, S. S. & Jadhav, S. M. Deep convolutional neural network based medical image classification for disease diagnosis, *J. Big Data,* 6, 2019, art. 113.

[8]. Kniep, H. C. et al., Posterior circulation stroke: machine learning-based detection of early ischemic changes in acute non-contrast CT scans, *J. Neurol.*, 267, 2020, pp. 2632–2641.

[9]. Bernhardt, J., Hayward, K. S., Kwakkel, G., Ward, N., et al., Agreed definitions and a shared vision for new standards in stroke recovery research: The Stroke Recovery and Rehabilitation Roundtable taskforce. *International Journal of Stroke,* 12, 5, 2017, pp. 444–450.

[10]. Warach, S., Gaa, J., Siewert, B., Wielopolski, P., & Edelman, R. R., Time course of lesion development in patients with acute stroke, *Stroke*, 29, 10, 1998, pp. 2268–2276.

[11]. Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D., Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization, *arXiv:1610.02391,* 2016.

[12]. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., et al., An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, *arXiv:2010.11929,* 2020.

[13]. Wang, W., Chen, Z., Chen, X., Wu, J., VisionLLM: Large Language Model is also an Open-Ended Decoder for Vision-Centric Tasks, *arXiv:2305.11175,* 2023.

(042)

Optimizing Medical Information Retrieval: Innovative Methodologies for Enhanced Precision and Efficiency

Prachi Patel
Harrisburg University of Science and Technology, Pennsylvania, USA
E-mail: prachi25898@gmail.com

Abstract: The complexity of medical data retrieval is continuously evolving, demanding refined methodologies to tackle the unique challenges posed by medical jargon, varied data sources, and privacy constraints. By adopting the PRISMA framework, this study systematically reviews cutting-edge approaches in medical information retrieval, focusing on strategies that improve both the accuracy and efficiency of IR models. Key innovations such as query expansion techniques, enhanced knowledge modeling, and advanced term-weighting schemes are examined for their role in resolving issues like synonymy, polysemy, and the integration of structured and unstructured data. The results demonstrate that modified IR approaches consistently outperform traditional models, providing healthcare professionals with more relevant and precise information, ultimately supporting better clinical decision-making and more effective patient care management.

Keywords: Medical information retrieval, Biomedical text mining, Neural information retrieval, Medical query expansion, Healthcare knowledge extraction.

1. Introduction

The latest developments in Medical Information Retrieval (IR) research have been oriented toward improving direct matching and employing innovative methods. The critical role of medical IR in healthcare is evident, as it empowers stakeholders—including clinicians, researchers, and policymakers—to obtain trustworthy, context-relevant, and timely information [1]. Medical IR involves retrieving various medical-related data sources, such as clinical guidelines, diagnostic aids, systematic reviews, and patient-specific data from diverse repositories [2]. Fast and accurate IR systems need to address multiple challenges, such as interpreting complex medical terms, understanding and processing queries, and generating outputs in a format suitable for patients and healthcare specialists.

A notable hurdle in medical IR is the complexity of the domain [3]. Specialized medical terminology often features abundant synonyms, abbreviations, and domain-specific jargon that can hinder the retrieval process. Additionally, the confidentiality and sensitivity of patient data necessitate robust mechanisms to ensure compliance with legal and ethical standards [4]. Another obstacle involves managing diverse medical data sources, spanning structured databases, semi-structured electronic health records, and unstructured scientific literature [5]. Addressing this diversity requires sophisticated approaches to consolidate and analyze data effectively while maintaining high precision and recall rates.

This paper aims to address these challenges by applying the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) methodology to systematically review and evaluate proposed approaches to improving medical IR. This methodology ensures a thorough and structured examination of existing literature, emphasizing techniques that mitigate the aforementioned challenges. This study explores advancements in query expansion, knowledge modeling, and weighting schemes to enhance the accuracy and relevance of retrieved medical information.

This research is guided by three key questions:

- Does it compare different IR techniques within the context of medical information retrieval?
- Are the results benchmarked against a standard dataset or compared to other known IR approaches in medicine?
- Is there any improvement gained from modifying and/or combining IR methods to enhance performance?

By addressing these questions, this study conducts a comprehensive literature review of proposed methodologies in medical IR, providing a critical evaluation of their effectiveness and relevance. Comparing techniques across various studies highlights the diverse approaches in this domain while scrutinizing benchmarking methods ensures alignment with reliable evaluation metrics. Moreover, examining innovations that modify or integrate methodologies sheds light on the potential of hybrid approaches to address current obstacles.

2. Research Method

This research uses a literature review approach based on previous research on Medical Information Retrieval (IR). The literature review method used in this study is the Preferred Reporting Item for Systematic Reviews and Meta-Analyses (PRISMA) method. This method consists of several steps: identification, screening, eligibility, and inclusion. Fig. 1 shows the steps of the PRISMA method.

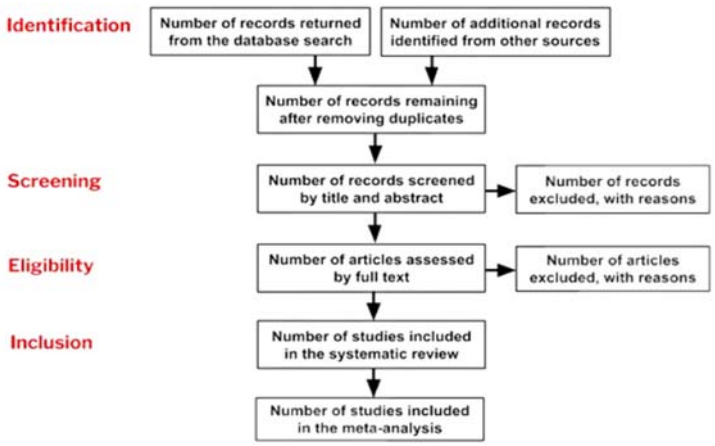


Fig. 1. PRISMA.

The PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) method is a standardized framework designed to enhance the transparency and reproducibility of systematic reviews. It is widely utilized to report the selection process of studies systematically and objectively, ensuring that researchers include relevant evidence while excluding unsuitable studies. The method comprises four primary stages: Identification, Screening, Eligibility, and Inclusion, which guide the researcher in narrowing the scope of the studies to be reviewed.

In the Identification stage, all potential studies are collected by searching databases and other sources. At this step, duplicate entries are removed to avoid redundancy. Papers are collected from several journal databases, such as ScienceDirect, Scopus, arXiv, and Springer. In order to retrieve the appropriate research papers, we use advanced search features and keywords based on our main topic. Our query consists of "Information Retrieval," "Medic," and "Enhance," where three of them are connected using the AND operator. "Medic" is expanded using the OR operator that links it to "Health" and "Knowledge" which is We also expand the "Enhance" term with the same method, adding "Improve" and "Boost" terms. To narrow the studies we found from the query, we set a number of constraints, including the English language, their release that is after 2010, and the research article type, tagged as Computer Science and Engineering domains. The remaining records then proceed to the Screening stage, where their titles and abstracts are evaluated based on predefined inclusion criteria. Studies not meeting these criteria are excluded, with reasons documented for transparency.

In the next stage, Eligibility, full texts of the remaining studies are thoroughly assessed to confirm their relevance and quality. Articles failing to meet the review criteria are excluded, ensuring only high-quality and relevant studies are retained. Finally, in the Inclusion stage, the selected studies form the basis of the systematic review. If applicable, the meta-analysis is conducted using a subset of these studies, providing quantitative insights by synthesizing data from multiple studies. This structured approach ensures rigor and clarity in systematic reviews.

3. Results

Using the keywords and constraints for advanced search, we searched and found many studies from the ScienceDirect database, that is, 811 results. arXiv, Scopus, and Springer gave much less, respectively,y 123, 95, and 79. To overcome this, we only picked the top 40 papers from each site, leaving us with 160 studies to proceed. Moving to the next step, we filtered them using a tool named Zotero, which will reduce the number of documents by deleting any duplicates and then excluding any study that doesn't match our topic by its title and abstract. There were 59 studies left after duplicate deletion further reduced to 19 from the filtering stage.

We manually checked the remaining papers and found that only 7 of them were appropriate for our topic. The final screening left only four studies, where we did an eligibility check on that collection. We got four selected research studies for the systematic review, including the study conducted by Oh and Jung [6], Wang et al. [7], Zhao et al. [8], and Balaneshinkordan and Kotov [9].

Table 1. Final Documents in Collection.

Modification	Method	References
Relevance and Expansion Models	Cluster-based External Expansion Model	Oh and Jung [6]
Part-of-Speech for Term Weighting Scheme	POS-Bag-of-Words, POS-MRF	Wang et al. [7]
Automation on Knowledge Model Creation	Data-Driven Sublanguage Pattern Mining Method	Zhao et al. [8]
Probabilistic Framework for Query Expansion	Bayesian Precision Medicine (BPM)	Balaneshinkordan and Kotov [9]

Numerous approaches have been suggested to boost the efficiency of medical Information Retrieval (IR) models based on literature data analysis. These approaches target enhancement in various aspects of medical IR, such as query expansion, knowledge modeling, and weighting schemes.

3.1. Cluster-based Query Expansion using External Collections

One of the main problems in information retrieval in the medical sector is the shortage of matched vocabulary (synonymy and polysemy phenomena) between a query and documents. One good way to address this problem is to expand the query using Pseudo-Relevant Feedback (PRF).

The External Expansion Model (EEM) is a PRF method that falls under this category. It constitutes an extension of the Mixture of Relevance Model (MoRM), which mainly deals with mixing different document types in the relevant collection so that the user-issued query can produce the final retrieval results, a mixture of different query-document relationships. However, it was noticed that these external collections aid EEM only in estimating a co-model.

[6] came up with a Cluster-Based External Expansion Model (CBEEM), which integrates EEM and Cluster-Based Document Model (CBDM) in order to overcome these limitations. The model sees each external collection as a cluster, which makes it the document model without any extra steps like K-means more accurate.

The trial outcomes illustrated that the CBEEM consistently ranked higher than QL and EEM among all databases. For TREC CDS, for example, the CBEEM increased the MAP measure by 15.19% and NDCG by 10.32%. In the case of OHSUMED, the CBEEM showed improvements of 15.51% in MAP and 12.33% in NDCG. All of these indicate that the CBEEM will be successfully used to improve medical information retrieval systems.

3.2. Part-Of-Speech Term Weighting Scheme

Current Information Retrieval (IR) systems employ the well-known Bag-of-Words (BoW) approach, but it has a restricted linguistic point of view. Markov Random Field (MRF) is applied, claiming word adherences are dependent on each other should be helpful to solve this problem. However, it is slim. In the process of the claim, the algorithm that was supposed to assign the weights to the terms within a clique, thus, mainly targeting those informative terms, was potentially overlooking informativeness differences.

To do so, researchers [7] created a methodology to highlight the biomedical domain for information retrieval by using an automatic Part-Of-Speech term weighting scheme. The POS-BoW and POS-MRF were formed from the term weighing scheme and the already existing BoW and MRF models, respectively.

By introducing the weight calculation algorithm based on supervised learning, they have applied a number of optimization techniques (such as Mean Average Precision (MAP) through the golden section search method and cyclic coordinate descent). Their experimental datasets were composed of the Medical Records track (TREC 2011 and 2012) and the Genomics track (TREC 2004). The developed POS-BoW and POS-MRF models were experimentally compared against the already known tf-idf and BM25 methods.

According to the research, the POS-BoW model (with Mean Average Precision (MAP)) showed a rise of 10.88%, while the POS-MRF model also demonstrated a 13.59% increase with respect to the traditional BoW and MRF approaches. In the case of the TREC 2004 Genomics track, the POS-MRF model performance improved by around 10.04% compared to the baseline. The newly proposed POS-driven term weighting scheme outperformed frequency-based and heuristic-based methods, thus proving the potential of the learning algorithm to generate optimal weights affecting the retrieval process of information (IR).

3.3. Data-Driven Sublanguage Pattern Mining

Knowledge models serve as structured representations of information within a specific domain, comprising semantic categories as nodes and semantic relationships as edges. These models play a crucial role in converting unstructured textual data into a computationally interpretable format. Existing knowledge models typically depend on machine learning or manually constructed ontologies. As a consequence, it leads to low accuracy in extracting entities and relationships, limited stability and adaptability to new domains due to manual efforts, and ineffectiveness in handling semantic complexity.

To combat this, a new method is introduced by [8] that integrates Natural Language Processing (NLP) and semantic network analysis techniques. This method aims to reveal a sublanguage pattern model within a domain-specific context. It utilizes a novel, bottom-up, context-based approach, combining syntactic analysis and data-driven semantic network analyses to explore semantic relations within a particular corpus.

The evaluation results indicate that the model covers 98% of entities and 97% of relationships. Machine annotation achieved a precision of 87%, recall of 79%, and an F-score of 82%. This method shows a promising path to improving the effectiveness of medical information retrieval systems through data-driven sublanguage pattern mining.

3.4. Bayesian Approach to Incorporate Different Types of Knowledge Bases

Precision Medicine (PM) [10] is an ongoing initiative wherein highly personalized healthcare is pursued by providing differential diagnoses of patients at both physiological and molecular levels. The

current Information Retrieval (IR) system in Clinical Decision Support (CDS) for Precision Medicine (PM) still cannot optimally utilize data such as genomic and textual data in structured queries. Further, they encounter problems such as a vocabulary mismatch, semantic heterogeneity in medical information, and topic drift during query expansion due to the inclusion of unrelated concepts.

Balaneshinkordan and Kotov developed a new probabilistic framework, Bayesian Precision Medicine (BPM), which uses the Bayesian approach for query expansion [9]. This approach has a goal to increase the exactness of IR systems in the case of the selection of specific patient articles in CDS systems.

The BPM framework establishes a set of concept extensions for the query via such sources as the Unified Medical Language System (UMLS), Drug-Gene Interaction Database (DGIdb), as well as relevant MEDLINE articles first. This is done by assigning prior probabilities for co-occurring terms and genomic context while tying textual and genomic query details together using Bayesian inference.

Analyzing the TREC 2017 Precision Medicine dataset reveals that BPM, with a total recall of 0.4929 against the current system's 0.4593, has shown a marked improvement in system performance. Important findings include the superior concept performance of the sources from MEDLINE and the finer accuracy achieved by the disease and gene characteristics.

3.5. Deep Learning Models for Medical Information Retrieval

While Bayesian and NLP-based models have demonstrated significant improvements in medical IR, deep learning approaches such as neural networks and transformer-based retrieval models have also gained traction. Techniques like BERT-based retrieval models (e.g., BioBERT and ClinicalBERT) leverage contextual embeddings to enhance query-document relevance. Unlike traditional models, these deep learning methods can capture complex linguistic patterns, improving retrieval performance, particularly in cases of vocabulary mismatches and semantic ambiguities.

Recent studies have shown that transformer-based models outperform classical IR approaches in tasks such as document ranking and question answering in medical domains. For instance, the integration of pre-trained medical language models has resulted in state-of-the-art performance on benchmark datasets like TREC CDS and MIMIC-III. However, these models often require large-scale annotated datasets and significant computational resources, making them less accessible for real-time clinical applications.

By incorporating deep learning into medical IR, future research can further enhance retrieval accuracy and adaptability, bridging the gap between traditional probabilistic models and modern AI-driven retrieval techniques.

3.6. RQ1: Comparison of Different IR Techniques in Medical Information Retrieval

The report described several new IR techniques for the healthcare industry. It provides solutions to problems like vocabularies mismatches, semantics complexity, and the need to integrate multiple data sources in medicine. The studies analyzed used different techniques, including the Cluster-Based External Expansion Model (CBEEM), Part-Of-Speech (POS) term weighting schemes, data-driven sublanguage pattern mining, and Bayesian Precision Medicine (BPM).

CBEEM got really good results in research by actually acquiring model data from databases and using them to make the model more relevant after absorbing the vocabulary' mismatch. Like that, the POS term weighting scheme IR system gains accuracy because the linguistic context is taken into consideration, which is particularly important in the medical setting where the terminologies can differ. The data-driven sublanguage pattern mining technique efficiently showed a way of learning the nature and language leadership of new entities and relationships. Within the BPM framework, the Bayesian methods were used to integrate genomic and textual data, and hence this made retrieval of the information relevant for clinical decision-making easier.

Of the two, the direct comparison of these techniques suggests that the more successful modified IR methods often perform better than traditional tactics, which shows the real need for special applications in medicine.

3.7. RQ2: Benchmarking Against Standard Datasets and Known IR Approaches

The systematic review carried out in this research aimed to overcome the problems encountered in Medical Information Retrieval (IR) by evaluating various methods using the PRISMA framework. The findings of the specific experiments mainly provide the reader with an understanding of the effectiveness of transformed IR techniques in this field. This section is concerned with disclosing the study's findings related to the three main research questions (RQs) that were earlier asked.

The studies that were examined in this lit review used different benchmarking methods to gauge the performance of their IR systems. One example of this is when CBEEM's relative performance was compared to conventional methods such as Query Likelihood (QL) and the initial External Expansion Model (EEM), in which the results showed that there was a significant increase in MAP and NDCG by means of both TREC CDS and OHSUMED sets, respectively. Similarly, the POS term weighting models were also compared to established baselines such as tf-idf and BM25, which exhibited noticeable improvements in MAP.

BPM Framework also used the TREC 2017 Precision Medicine dataset to measure its performance, concluding that the new system is very

strong compared to existing systems. These benchmarking activities not only prove the efficacy of the hypothesized methods but also guarantee that the advances made are based on a reliable evaluation measure. The consistent enhancements throughout these various works are poignant evidence for the thesis that adapting IR methods is the only way to address the peculiar problem of medical information search.

3.8. RQ3: Improvement from Modifying and/or Combining IR Methods

The review indicates that modifying and combining existing IR methods resulted in very good performance. The inclusion of external collections in CBEEM, the linguistics in the POS term weighting scheme, and a novel data-driven approach in sublanguage pattern mining are all examples of how modifying established methods can improve outcomes in the medical field.

Furthermore, the probabilistic method described in the BPM framework for query expansion demonstrates that combining different types of knowledge bases can be accomplished by effectively dealing with vocabulary mismatch and semantic heterogeneity. The feedback from these studies suggests that mixed methods of hybrid models can deliver the accuracy and relevance of the extracted medical data and can be one way to break the boundaries of IR methodologies.

Summing up, the results of this systematic review are the turning point for IR technology to evolve into a highly specialized medical field. The development featured in this study not only increases the functionality of IR systems but also perks up more efficient decision-making by healthcare providers; therefore, end-users and stakeholders are satisfied.

4. Conclusion

The literature review was based on studies conducted using modified information retrieval (IR) techniques in medical data analysis. The PRISMA method was applied, and four scholarly articles were selected according to the given criteria. The study's findings point out that the revised IR models were practically always found to bring about marked advantages in terms of performance compared to standard methods. They also had capabilities in data dependency management, especially in handling vocabulary discrepancies and semantic complications.

Furthermore, the introduction of deep learning techniques, such as transformer-based models, has expanded the capabilities of medical IR. These models have demonstrated superior retrieval performance, particularly in understanding complex biomedical text. However, challenges such as computational costs and data availability remain areas for future exploration.

Future research should focus on developing lightweight deep-learning models that balance computational efficiency with retrieval accuracy. Additionally, incorporating more diverse medical datasets and addressing biases in current models will be essential in improving real-world applicability. Enhancing interpretability and transparency in AI-driven medical IR systems can further facilitate trust and adoption in clinical settings.

As news published suggests, the changes to IR methods can be looked at from both the side of model effectiveness and the improved accuracy of information retrieval; thus, the medical field will also have better clinical decision-making protocols.

References

[1]. J. McGowan and M. Sampson, Systematic reviews need systematic searchers, *Journal of the Medical Library Association,* Vol. 93, No. 1, 2005, pp. 74–80.

[2]. W. Raghupathi and V. Raghupathi, Big data analytics in healthcare: promise and potential, *Health Information Science and Systems,* Vol. 2, 2014, 3.

[3]. P. Zweigenbaum, D. Demner-Fushman, H. Yu, and K. B. Cohen, Frontiers of biomedical text mining: current progress and future directions, *Yearbook of Medical Informatics,* Vol. 16, No. 1, 2007, pp. 73–86.

[4]. D. McGraw and K. D. Mandl, Privacy protections to promote interoperability of digital health data, *New England Journal of Medicine,* Vol. 384, 2021, pp. 2301–2303.

[5]. Y. Wang, N. Afzal, S. Fu, et al., MedSTS: a resource for clinical semantic text similarity, *arXiv:1808.09397,* 2018.

[6]. H. Oh and H. Jung, Cluster-based query expansion using external collections in medical information retrieval, *Journal of Biomedical Informatics*, Vol. 58, 2015, pp. 70–79.

[7]. Y. Wang, S. Wu, D. Li, S. Mehrabi, and H. Liu, A part-of-speech term weighting scheme for biomedical information retrieval, *Journal of Biomedical Informatics,* Vol. 63, 2016, pp. 379–389.

[8]. Y. Zhao, N. J. Fesharaki, H. Liu, et al., Using data-driven sublanguage pattern mining to induce knowledge models: application in medical image reports knowledge representation, *BMC Medical Informatics and Decision Making,* Vol. 18, 2018, 61.

[9]. F. Balaneshinkordan and A. Kotov, Bayesian approach to incorporating different types of biomedical knowledge bases into information retrieval systems for clinical decision support in precision medicine, *Journal of Biomedical Informatics,* Vol. 98, 2019, 103238.

[10]. F. S. Collins and H. Varmus, A new initiative on precision medicine, *New England Journal of Medicine,* Vol. 372, No. 9, 2015, pp. 793–795.

ISBN 978-84-09-71190-1