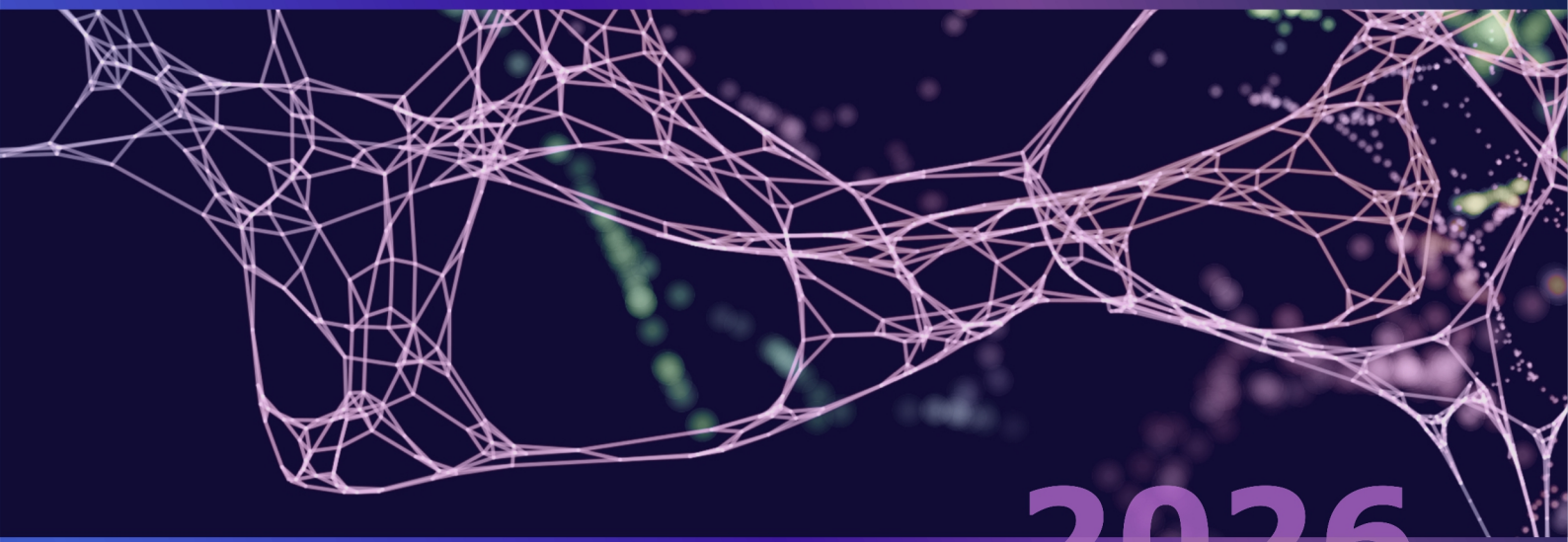
AiMH Mallorca 2026

Artificial Intelligence in Medicine and Healthcare

Proceedings

2nd International Conference on AI in Medicine and Healthcare (AiMH' 2026)



2026

24-26 June 2026, Palma de Mallorca (Balearic Islands), Spain

IFSA

Artificial Intelligence in Medicine and Healthcare

Proceedings of the 2nd International Conference on AI in Medicine and Healthcare (AiMH' 2026)

24-26 June 2026
Palma de Mallorca (Balearic Islands), Spain

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

IFSA Publishing

Sergey Y. Yurish, *Editor*
Artificial Intelligence in Medicine and Healthcare
AiMH' 2026 Conference Proceedings

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

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

ISBN: 978-84-09-86146-0
BN-20260619-XX
BIC: UYQ

Contents

Foreword 5

Diagnostic Performance of an AI System for Mammography Risk Assessment: Low-Prevalence Retrospective Study 6
D. Kvak, K. Kvaková and M. Biroš

Classification of Capillaroscopic Changes in Systemic Sclerosis Using Vision Transformers: Model Refinement, Explainability and Fairness Analysis 11
D. Haag, L. Bogensperger, S. Difrancesco, C. Mihai and M. Krauthammer

Generating a New Objective Index (SURG-STRI) to Evaluate the Surgical Stress from ECG Sensor Data 15
D. Caballero, M. J. Pérez-Salazar, J. A. Sánchez-Margallo and F. M. Sánchez-Margallo

Stereo-Vision Localization of Millimeter-Scale Capsule Targets in a Clinically Inspired Maze: A Repeatability Study 19
I. Frajtag, L. Masjosthusmann, S. Misra and F. Šuligoj

A Pilot Study on Facial Landmark Detection on CT and MR Head Projections for Initial Multimodal Registration 25
Filip Šuligoj, Marko Švaco, Bojan Šekoranja and Bojan Jerbić

Interpretable AI Framework for Raman Spectroscopy Based Diagnostics 30
J. Tomeš, D. Janstová, M. Garnol, and J. Mareš

Transformer-Based Classification of Raman Spectra for Cancer Detection 33
D. Janstová, J. Tomeš, M. Garnol, and J. Mareš

A Machine Learning Approach to identify nutrition-related Diseases based on Fingernail Structure 36
Jan-Torsten Milde, N. Sonntag, L. Müller, K. Plandl, D. Dewald, R. Blum, H. Hollenbach, A. Maxones, T. Kühn, M. Birringer

Trust by Design: Building Reliable Clinical AI to Advance Quality in Patient-Centered Care 39
Maria L. Reyna-Cruz, Martine Ceberio, Christoph Lauter, Cecilia A. Marquez Barraza and Jesus M. López Valles

Artificial Intelligence in Nursing Care: Opportunities, Challenges, and Future Directions 43
K. Wolf-Ostermann

SAVE & SAFE: An AI-Supported Assistive System for Fall Prevention and Nursing Workload Reduction in Acute Geriatric Care 49
E. Mena, T. Schultz and K. Wolf-Ostermann

AI-Driven Services for Care Facilities: Results from a Longitudinal Field Study 52
R. E. Paul and K. Wolf-Ostermann and T. Schultz

Bibliometric Analysis of Nursing Research Related to Artificial Intelligence in Nursing Care 56
Şengül Akdeniz and K. Wolf-Ostermann

Dependence on Fragile AI Systems: Rethinking the Collingridge Dilemma in Clinical AI 63
A. Gerdes

Policy-Focused Evaluation of Individualized Vasopressor Effects with Robustness to Irrelevant Covariates in MIMIC-III ICU Data 67
A. S. Khan, E. Schaffernicht and J. A. Stork

Evaluating Artificial Intelligence in Generating Biochemistry Knowledge Assessments for Medical Education 80
Veljkovic Andrej, Aleksandar Mitic, Ognjen Radovic, Monika Simjanoska Misheva and Stevo Lukic

AI-Based Bacterial Detection Using Multiple Biosensing Technologies under Data Scarcity 83
Felipe Yamada, António Cardoso, Flávia Barbosa, and Luís Guimarães

The Ethical Implications of Artificial Intelligence in Orthopedic Surgery: A Systematic Review 86
Tobi Kamoru, Rebbeca Alemu, Nuhame Mulugeta, Muna Jalani, John Cyrus and Lauren A. Barber

AI-Concordant Care: A Novel Approach toward Precision Depression Treatment 91
Yijun Shao, Yan Cheng, Hank Wen-Chih Wu, Tracey H. Taveira, John E. McGeary, Kevin McConeghy, Ying Yin, Ali Ahmed, Qing Zeng-Treitler

Hierarchical Contrastive Alignment with a Triplet–Focal Objective for Cervical Cytology Classification 94
Bekhzod Olimov and Sung Wook Ahn

CAPF: A Clinical Agent Permission Framework for HIPAA-Aligned Least-Privilege Authorization in Multi-Agent Healthcare AI Systems 100
R. Khanna and J. Nandal

ANESTHOS: A Human-in-the-Loop Perioperative Workflow Architecture for Clinical Decision Support, Safety and Continuous Operational Intelligence 107
A. Binagui Buitureira

Foreword

Artificial intelligence is rapidly transforming modern medicine and healthcare. What only recently appeared to be a promising technological direction has now become an active field of research, development and clinical discussion. AI-based methods are increasingly applied in medical imaging, diagnostics, decision support, robotics, nursing care, hospital workflows, medical education, biosensing, rehabilitation, and personalized treatment. At the same time, the implementation of AI in medicine requires much more than algorithmic performance alone. It demands reliability, transparency, fairness, clinical validation, ethical responsibility, data protection, and a clear understanding of the role of healthcare professionals in the final decision-making process.

The *2nd International Conference on Artificial Intelligence in Medicine and Healthcare* (AiMH' 2026) has been organized as an international forum where researchers, engineers, clinicians, data scientists and healthcare specialists can exchange ideas, present recent achievements, and discuss both the opportunities and the limitations of artificial intelligence in medical and healthcare environments. The papers included in these Proceedings reflect the interdisciplinary nature of this field. They address a wide range of topics, including AI-assisted mammography, medical image analysis, capillaroscopy, Raman spectroscopy-based diagnostics, surgical stress assessment, robotic and capsule systems, AI in nursing care, clinical decision support, mental health applications, bacterial detection, medical education, ethical implications, and governance frameworks for healthcare AI systems.

A particularly important feature of the present volume is its attention not only to technological innovation, but also to clinical relevance. Many contributions emphasize explainability, robustness, human-in-the-loop approaches, fairness analysis, real-world validation and the responsible integration of AI into healthcare practice. This is essential, because in medicine, artificial intelligence should not be seen as a replacement for professional expertise, but as a powerful supporting instrument that may help improve accuracy, efficiency, safety and accessibility of care when properly designed, validated and monitored.

The development of trustworthy AI for medicine is a shared responsibility. Researchers must create reliable and interpretable models; clinicians must define meaningful medical needs and evaluate practical usefulness; engineers must ensure secure and robust implementation; and policy makers and institutions must provide an appropriate ethical and regulatory framework. Conferences such as AiMH' 2026 are important because they bring these communities together and create a space for constructive dialogue between technical innovation and clinical reality.

On behalf of the Organizing Committee, I would like to express my sincere gratitude to all authors for their valuable contributions and for sharing their latest research results with the international community. I also extend my appreciation to the reviewers, whose careful evaluations helped to maintain the scientific quality of the conference, and to all members of the committees and supporting organizations for their dedication and professional efforts in preparing this event and its Proceedings.

I hope that this volume will serve not only as a record of the scientific work presented at AiMH' 2026, but also as a source of inspiration for future research, collaboration and responsible innovation. May the ideas discussed here contribute to the development of artificial intelligence technologies that are not only advanced, but also safe, transparent, clinically meaningful and beneficial for patients and healthcare professionals alike.

Prof., Dr. Sergey Y. Yurish,
AiMH' 2026 Conference Chairman

(009)

Diagnostic Performance of an AI System for Mammography Risk Assessment: Low-Prevalence Retrospective Study

D. Kvak [1,2], K. Kvaková [2] and M. Biroš [2]
[1] Faculty of Medicine, Masaryk University, Kamenice 5, 625 00, Brno, Czech Republic
[2] Carebot s.r.o., Holečkova 3150/25B, 150 00, Prague, Czech Republic
Tel.: + 420 739 174 316
E-mail: daniel.kvak@carebot.com

Summary: AI tools for mammography are increasingly used to support lesion detection and risk stratification, but clinical utility depends on performance in routine practice where disease prevalence is low and negative examinations predominate. We performed a retrospective diagnostic-accuracy study of consecutive screening mammography examinations acquired in January 2024 at AGEL Hospital Nový Jičín (Czech Republic) on a GE Senographe Pristina system. The reference standard was established by three senior breast radiologists using BI-RADS assessment. Of 338 examinations reviewed, 29 did not achieve consensus and were excluded, leaving 309 examinations for analysis. Consensus labels were grouped as Normal (BI-RADS 1), Benign/probably benign (BI-RADS 2–3), and Suspicious (BI-RADS 4–5; imaging-based suspicion, not pathology-confirmed cancer). The index test was the study-level output of an AI mammography system (Carebot AI MMG; Carebot s.r.o., Czech Republic). Two prespecified endpoints were evaluated: (1) any-lesion detection (BI-RADS 2–5 vs BI-RADS 1), with Medium/High Risk considered AI-positive; and (2) suspicious examination identification (BI-RADS 4–5 vs BI-RADS 1–3), with High Risk considered AI-positive. The analysed cohort included 233 Normal, 68 Benign/probably benign, and 8 Suspicious examinations (prevalence of BI-RADS 4–5 suspicion: 2.6 %). For endpoint 1, sensitivity was 0.895 (95 % CI 0.806–0.946) and specificity 0.940 (0.902–0.964). For endpoint 2, sensitivity was 0.875 (0.529–0.978) and specificity 0.857 (0.813–0.892). Predictive values reflected real-world prevalence: for endpoint 1, PPV 0.829 and NPV 0.965; for endpoint 2, PPV 0.140 and NPV 0.996. These findings highlight the importance of reporting predictive values in prevalence-representative cohorts and motivate further multi-centre validation and linkage to outcome-confirmed diagnoses where feasible.

Keywords: Mammography, Artificial intelligence, Diagnostic accuracy, BI-RADS, Sensitivity, Specificity, Predictive value, Real-world evidence.

1. Introduction

Over the last decade, advances in computer vision and deep learning have enabled a new generation of AI tools for mammography. Large retrospective evaluations and reader studies suggest that modern AI systems can approach radiologist-level performance for detecting clinically relevant findings and may add value when used as decision support (e.g., as a second reader, triage aid, or quality-safety layer) rather than as a replacement for clinical judgment [1, 2]. The clinical rationale is particularly strong in screening settings: examinations are high-volume, most are negative, subtle abnormalities can be missed, and radiology services must balance sensitivity for important findings against the downstream burden of false positives and unnecessary recalls. In this context, an AI system that reliably highlights examinations more likely to contain actionable findings could support prioritization, reduce oversight errors, and improve workflow.

Professional societies emphasize that currently deployed AI tools should be treated as adjuncts that may be accepted or overruled by the reporting radiologist, and that robust governance and post-deployment evaluation are essential to ensure safe and effective use in clinical practice [3]. A key methodological issue is that predictive values depend strongly on prevalence. While sensitivity and specificity describe performance relative to a reference standard, positive and negative predictive values answer the operational clinical question: given a positive (or negative) AI output, what is the probability that the examination truly meets the condition of interest [4]. Therefore, performance measured on cancer-enriched datasets may not translate to routine screening populations dominated by negative examinations. In this work, we report a retrospective evaluation of a mammography AI system in a prevalence-representative clinical cohort.

2. Methodology

2.1. Study Design

This retrospective study evaluates the diagnostic accuracy of a study-level AI output against a radiologist consensus reference standard. The manuscript is structured according to the Standards for Reporting Diagnostic Accuracy Studies (STARD 2015) guidance, adapted to the data elements [5].

2.2. Data Acquisition

Mammography examinations were collected retrospectively at AGEL Hospital Nový Jičín (Czech Republic) and acquired in January 2024. All examinations in the evaluated cohort were acquired on a GE Senographe Pristina mammography system. Data were transferred for analysis under the contract "Agreement on cooperation in the development of Carebot software" dated 31 March 2025.

2.3. Ground Truth

Ground truth (reference standard) was established by three senior breast radiologists (>20 years of experience). Each radiologist independently reviewed each mammography study and recorded structured annotations at the study and projection level. For this evaluation, annotations were converted into study-level consensus categories. The category Suspicious denotes BI-RADS 4–5-aligned imaging suspicion rather than pathology-confirmed cancer and was derived from suspicious mass or architectural distortion annotations. The category Benign/probably benign denotes BI-RADS 2–3-aligned findings and was derived from benign mass annotations. Normal denotes absence of suspicious mass, architectural distortion, or benign mass annotation. Microcalcifications and risk-area markings were not used in endpoint definitions. Consensus ground truth was created at study-level with prespecified rules: a study was labeled Suspicious if at least 2 of 3 radiologists assigned a suspicious mass or architectural distortion label; Benign/probably benign if at least 2 of 3 radiologists assigned a benign mass label and fewer than 2 assigned a Suspicious label; and Normal if at least 2 of 3 radiologists assigned Normal and no radiologist assigned Suspicious. Studies not meeting these criteria were considered discordant and excluded from performance evaluation.

A total of 338 consecutive mammography examinations were reviewed. Consensus reference standard was available for 309 examinations; 29 examinations were discordant and excluded. Within the analysed cohort, radiologist consensus classified 233 examinations (75.4 %) as Normal, 68 (22.0 %) as Benign/probably benign, and 8 (2.6 %) as Suspicious. This distribution reflects a prevalence-representative screening cohort dominated by negative examinations.

Table 1. Cohort demographics by BI-RADS category.

Class	BI-RADS	n	Age (mean, SD)	Age (range)
Normal	1	233	60.6 ± 11.0	40-84
Benign	2–3	68	59.5 ± 10.4	40-84
Suspicious	4–5	8	52.2 ± 5.9	45-64

2.4. Software

The AI system (Carebot AI MMG software version 2.9; Carebot s.r.o., Czech Republic; Fig. 1) is a software medical device designed to support radiologists in detection of breast lesions on standard digital mammography. The system processes mammography studies provided in DICOM format and generates a study-level categorical output with three ordered levels: Low Risk, Medium Risk, and High Risk. The algorithm was developed using 105,496 unique radiologist annotations collected between 21 February 2023 and 16 June 2025. No examinations from the evaluation cohort were used for training, validation, tuning, or model development.

Data leakage was prevented by separating the evaluation cohort at the examination level. All StudyInstanceUIDs from the January 2024 evaluation cohort were excluded from the development dataset, and the radiologist consensus labels used in this study were generated only for the locked evaluation cohort.

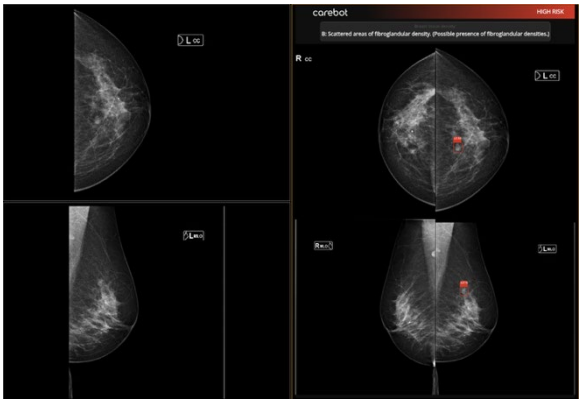


Fig. 1. Example AI system output.

2.5. Endpoints and Statistical Analysis

Two prespecified endpoints were evaluated to reflect distinct clinical questions: (1) whether the AI system flags examinations that contain any lesion-level finding requiring attention, and (2) whether the AI system identifies examinations assessed as Suspicious by the radiologist consensus reference standard (i.e., imaging-based suspicion, not pathology-confirmed malignancy). Endpoints and AI positivity thresholds were fixed before analysis.

Endpoint 1: Any-lesion detection. Reference-standard positive examinations were those labeled Benign/probably benign (BI-RADS 2–3) or Suspicious (BI-RADS 4–5); reference-standard negative examinations were Normal (BI-RADS 1). AI output was considered test-positive if the study-level category was Medium Risk or High Risk, and test-negative if Low Risk.

Endpoint 2: Suspicious examination identification. Reference-standard positive examinations were those labeled Suspicious (BI-RADS 4–5); reference-standard negative examinations were Normal (BI-RADS 1) or Benign/probably benign (BI-RADS 2–3). AI output was considered test-positive if the study-level category was High Risk, and test-negative if Low Risk or Medium Risk.

Performance was summarized using 2×2 confusion matrices and the following metrics: sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and accuracy. Balanced accuracy was also reported and defined as the mean of sensitivity and specificity to account for class imbalance. Two-sided 95 % confidence intervals were computed using Wilson score intervals for binomial proportions [6]. Uncertainty for balanced accuracy was estimated using non-parametric bootstrap percentile intervals. All analyses were performed in Python and are reported at the study-level.

3. Results

3.1. Endpoint 1: Lesion Detection

At this operating point, the AI system correctly identified 68 lesion-positive (TP) and missed 8 (FN) examinations, corresponding to a sensitivity of 0.895 (two-sided 95 % CI 0.806–0.946). All 8/8 Suspicious examinations were test-positive (7 High Risk, 1 Medium Risk); therefore, all false negatives at this threshold occurred in the Benign/probably benign category. Among Normal examinations, the AI produced 14/233 test-positive results (FP = 14; 11 High Risk and 3 Medium Risk), yielding a specificity of 0.940 (0.902–0.964). Overall classification accuracy for endpoint 1 was 0.929 (0.895–0.953). Balanced accuracy was 0.917 (0.878–0.954, bootstrap). Predictive values reflected the prevalence-representative cohort: PPV was 0.829 (0.734–0.895) and NPV was 0.965 (0.932–0.982).

Table 2. Diagnostic accuracy for endpoint 1.

Metric (Endpoint 1)	Estimate (95 % CI)
Sensitivity	0.895 (0.806–0.946)
Specificity	0.940 (0.902–0.964)
Accuracy	0.929 (0.895–0.953)
Balanced accuracy	0.917 (0.878–0.952)
PPV	0.829 (0.734–0.895)
NPV	0.965 (0.932–0.982)

3.2. Endpoint 2: Suspicious Lesion Identification

Using High Risk as the decision threshold, the AI correctly flagged 7 Suspicious examinations as test-positive (TP) and missed 1 (FN), yielding sensitivity 0.875 (0.529–0.978). Because this endpoint included only eight Suspicious examinations, this sensitivity estimate has wide uncertainty and should be interpreted cautiously. Among non-suspicious examinations, 43 were classified as High Risk (FP), corresponding to specificity 0.857 (0.813–0.892). The false positives were concentrated among Benign/probably benign examinations: 32/43 (74.4 %) false positives were Benign/probably benign and 11/43 (25.6 %) were Normal. Accuracy for endpoint 2 was 0.858 (0.814–0.892), and balanced accuracy was 0.866 (0.719–0.943, bootstrap). Because BI-RADS 4–5 Suspicious examinations were rare (prevalence 2.6 %), PPV for the High-Risk rule was 0.140 (0.070–0.262), while NPV remained very robust at 0.996 (0.978–0.999). Thus, in this screening-prevalence cohort, a High-Risk output enriched for Suspicious examinations but should not be interpreted as confirmed malignancy.

Table 3. Diagnostic accuracy for endpoint 2.

Metric (Endpoint 2)	Estimate (95 % CI)
Sensitivity	0.875 (0.529–0.978)
Specificity	0.857 (0.813–0.892)
Accuracy	0.858 (0.814–0.892)
Balanced accuracy	0.866 (0.721–0.944)
PPV	0.140 (0.070–0.262)
NPV	0.996 (0.978–0.999)

4. Discussion

This retrospective diagnostic-accuracy study evaluated a mammography AI system against a three-radiologist consensus reference standard in a prevalence-representative cohort from routine clinical practice. The cohort intentionally reflects a real screening-like setting where most examinations are negative and only a small fraction is assessed as Suspicious. In such settings, predictive values are highly prevalence-dependent, and PPV/NPV are essential for interpreting operational utility beyond sensitivity and specificity [4].

For endpoint 1 (any-lesion detection), using Medium Risk or High Risk as AI-positive yielded high sensitivity (0.895) and specificity (0.940). This operating point aligned well with expert assessment of whether an examination contains a mass-like or distortion-like finding that warrants attention. For endpoint 2 (Suspicious examination identification), using High Risk as AI-positive resulted in sensitivity 0.875 and specificity 0.857, whereas PPV was 0.140. This combination is expected when the target condition is rare (here, 8/309 examinations, 2.6 %). Even a moderate false-positive rate can dominate the set of AI-positive examinations and drive PPV down, while NPV becomes very high (0.996). Because endpoint 2 was based on only eight Suspicious examinations, the point estimate should be regarded as exploratory and not overinterpreted. These results emphasize that, in prevalence-representative cohorts, High Risk outputs should be interpreted as prioritization signals rather than direct proxies for confirmed malignancy. In practice, this reinforces professional guidance that AI outputs should be treated as decision support subject to radiologist confirmation and monitored after deployment [3].

4.1. Limitations

This study has several limitations. First, it is retrospective and single-center, and all examinations were acquired on a single mammography platform, which may limit generalizability. Second, uncertainty was substantial for endpoint 2 sensitivity due to the small number of Suspicious examinations, and estimates should be interpreted cautiously. Third, 29 discordant examinations were excluded because no consensus reference standard could be assigned. This yields a cleaner reference standard, but may also inflate apparent performance by removing difficult or ambiguous cases. Fourth, the reference standard does not represent pathology-confirmed cancer. The Suspicious category reflects imaging-based suspicion aligned with BI-RADS principles [7], and the study therefore measures agreement with expert interpretation rather than true cancer status. Outcome linkage to histopathology and follow-up would be required to quantify performance for confirmed malignancy and to interpret certain apparent false positives and false negatives. We report uncertainty using Wilson score confidence intervals for binomial metrics and follow diagnostic-accuracy reporting principles [5, 6].

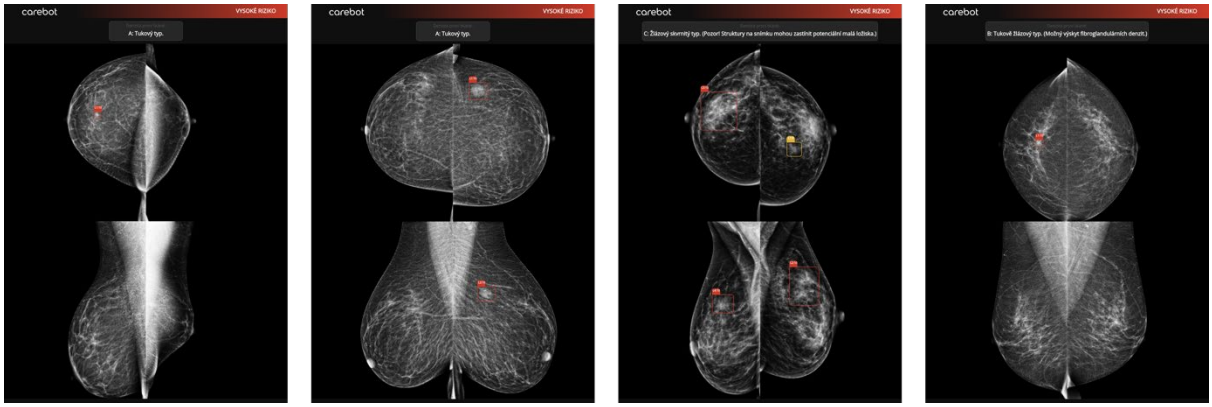


Fig. 2. True positive (TP) examples from the Suspicious category, correctly flagged by the AI system.

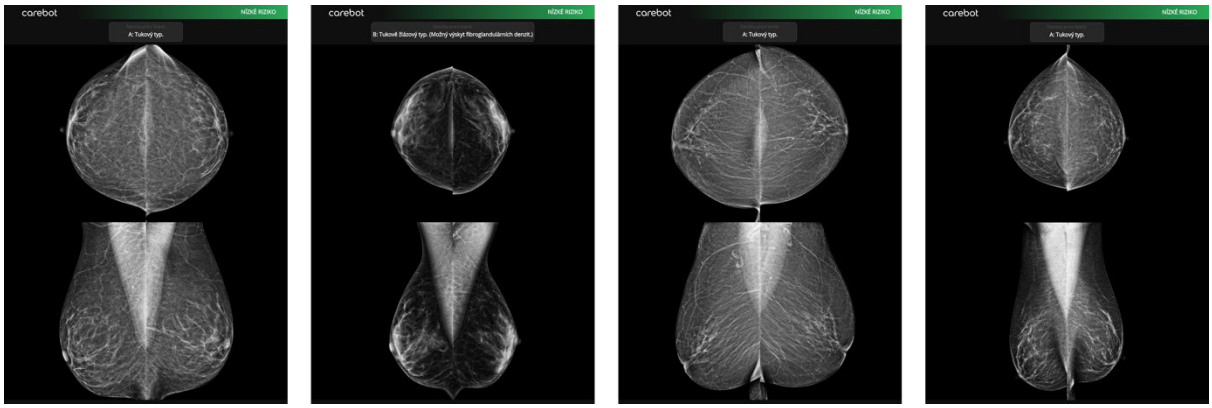


Fig. 3. True negative (TN) examples from the Normal category, correctly flagged by the AI system.

5. Conclusions

In this retrospective, real-world prevalence-representative screening cohort evaluated against a radiologist consensus reference standard, the AI system demonstrated robust agreement with expert assessment for identifying examinations with lesion-level findings. Performance for identifying radiologist-assessed Suspicious examinations was strongly influenced by the low prevalence and small number of such findings, underscoring the need to interpret outputs in a workflow context rather than as confirmed malignancy. Overall, the results support the use of AI as decision support that can help prioritize examinations for review.

Abbreviations

AI: Artificial intelligence;
BI-RADS: Breast Imaging Reporting and Data System;
CI: Confidence interval;
NPV: Negative predictive value;
PPV: Positive predictive value;
TP: True Positive;
TN: True Negative;
FP: False Positive;
FN: False Negative.

References

[1]. S. M. McKinney, M. Sieniek, V. Godbole, et al., International evaluation of an AI system for breast cancer screening, *Nature*, Vol. 577, Issue 7788, 2020, pp. 89-94.

[2]. A. Rodriguez-Ruiz, K. Lång, A. Gubern-Merida, et al., Stand-alone artificial intelligence for breast cancer detection in mammography: Comparison with 101 radiologists, *Journal of the National Cancer Institute*, Vol. 111, Issue 9, 2019, pp. 916-922.

[3]. S. Schiaffino, D. Bernardi, N. Healy, et al., ESR essentials: Artificial intelligence in breast imaging – practice recommendations by the European Society of Breast Imaging, *European Radiology*, Vol. 36, 2026, pp. 1909-1918.

[4]. D. G. Altman, J. M. Bland, Diagnostic tests 2: Predictive values, *BMJ*, Vol. 309, Issue 6947, 1994, 102.

[5]. P. M. Bossuyt, J. B. Reitsma, D. E. Bruns, et al., STARD 2015: An updated list of essential items for reporting diagnostic accuracy studies, *BMJ*, Vol. 351, 2015, h5527.

[6]. L. D. Brown, T. T. Cai, A. DasGupta, Interval estimation for a binomial proportion, *Statistical Science*, Vol. 16, Issue 2, 2001, pp. 101-133.

[7]. M. S. Newell, S. V. Destounis, J. W. T. Leung, W. B. DeMartini, et al., ACR BI-RADS v2025 Manual, *American College of Radiology*, Reston, 2025.

(014)

Classification of Capillaroscopic Changes in Systemic Sclerosis Using Vision Transformers: Model Refinement, Explainability and Fairness Analysis

D. Haag [1], L. Bogensperger [2], S. Difrancesco [2], C. Mihai [3] and M. Krauthammer [2]
[1] ETH Zurich, Rämistrasse 101, 8092 Zurich, Switzerland
[2] University of Zurich, Krauthammer Lab, Schmelzbergstrasse 26, 8091 Zurich, Switzerland
[3] Department of Rheumatology, University Hospital Zurich, University of Zurich, Rämistrasse 100, 8091 Zurich, Switzerland
E-mail: delia.haag@gmx.ch

Summary: Nailfold capillaroscopy (NFC) is an important imaging technique for detecting microvascular changes associated with systemic sclerosis (SSc). In this work, we evaluate a Vision Transformer (ViT)–based model for automated classification of NFC images. We explore different image preprocessing techniques, multiple training configurations, and model fairness across different demographic subgroups. Gradient-weighted class activation mapping (Grad-CAM) was implemented to enhance model explainability. Data from two cohorts, the European Scleroderma Trials and Research Group (EUSTAR) and Very Early Diagnosis of Systemic Sclerosis (VEDOSS), were used to train three model variants for multi-label classification of five capillary abnormalities. The best-performing model, trained with a weighted loss function, achieved F1 scores between 0.48 and 0.66. Furthermore, there was no statistically significant evidence of performance disparity across sex and age subgroups. Overall, the ViT may have potential as a supplementary tool in clinical practice, but the final clinical assessment should remain under clinician supervision.

Keywords: Nailfold capillaroscopy, Systemic sclerosis, Multi-label image classification, Vision transformers, Medical imaging, Explainable AI, Fairness.

1. Introduction

Systemic sclerosis (SSc) is a rare autoimmune disease characterized by vasculopathy and progressive fibrosis of the skin and internal organs and is associated with high morbidity [1, 2]. Early diagnosis is therefore essential to improve patient outcomes. Nailfold capillaroscopy (NFC) is a non-invasive imaging technique used to visualize microvascular alterations and represents an important diagnostic tool for SSc [1]. Microangiopathic changes, such as enlarged or giant capillaries, capillary loss, and microhemorrhages, which together form the scleroderma pattern, are important indicators to support diagnosis and monitoring of the disease [3]. However, NFC image interpretation is predominantly based on qualitative or semi-quantitative assessments, making it time-consuming and dependent on expert assessors [4].

Recent work has explored automated analysis using machine learning, including a vision transformer–based model proposed by Garaiman et al. [1], which provides the basis for this study. Our work is based on the same underlying data, although the parameters of our data selection differ slightly. Additionally, the dataset was partitioned into new training, validation, and test splits. Human annotations from the original study were reused for the overlapping subset of images, enabling comparison with clinician performance. The aim of this study is to refine and extend this model for potential clinical use.

Due to the complexity of deep learning models, methods to increase their explainability gain importance [5]. Furthermore, enabling more insight into the decision-making process of a model is essential to build trust and to facilitate its potential integration in clinical practice [5]. A widely used technique is Gradient-weighted Class Activation Mapping (Grad-CAM) which computes the gradient of the output with respect to the model's embeddings [6]. This allows for the spatial localization of image regions that most strongly influence the prediction.

Furthermore, the clinical presentation of systemic sclerosis necessitates a critical examination of algorithmic fairness. Specifically, the underrepresentation of male patients introduces a risk of demographic bias that may compromise the generalizability of automated diagnostic tools.

While the results by Garaiman et al. [1] show great potential of automated image evaluation in this domain, explainability of the model and potential biases are not examined in detail and could fill an important gap, particularly with regard to a clinical application. Furthermore, class imbalance is not addressed, even though it plays an important role due to the underrepresentation of certain labels.

This work aims to address these technical and ethical considerations by evaluating explainability and fairness and by accounting for inherent class imbalance of the data. To summarize, our contributions include: (i) the refinement of image preprocessing; (ii) the exploration of training

configurations, such as loss function adjustments, to address data imbalance; (iii) the assessment of model explainability; and (iv) a fairness analysis across demographic subgroups.

2. Methods

2.1. Dataset and Preprocessing

This study included patients from the Rheumatology Department of the University Hospital Zurich enrolled either in the European Scleroderma Trials and Research Group (EUSTAR) or in the Very Early Diagnosis of Systemic Sclerosis (VEDOSS) cohort. The dataset comprised 16,192 nailfold capillaroscopy images (16 images per patient and visit).

Capillary abnormalities, such as the previously mentioned enlarged capillaries, giant capillaries, capillary loss, microhemorrhages and the scleroderma pattern, were encoded as binary labels, indicating presence or absence. This results in a multi-label classification problem, where the clinical features are non-exclusive and may occur simultaneously. Images were preprocessed using standard transformations, such as horizontal flipping and random cropping. After evaluating multiple scale ranges on the validation set, a random crop within a scale range of 0.7 to 1 resulted in the highest generalization performance. In addition, Contrast Limited Adaptive Histogram Equalization (CLAHE) was applied to enhance contrast while preserving local characteristics of the image [7].

2.2. Model Architecture and Training

A Vision Transformer model following Dosovitskiy et al. [8] was implemented. The resulting embeddings were passed through MLP heads to learn the final probability distribution for each label [1]. A 5-fold cross-validation scheme was used to ensure model robustness. Splitting was performed at the patient level, such that all images from a given patient were assigned to the same fold, thereby preventing data leakage. Given the class imbalance, with the representation of the positive class ranging from 19 % to 37 %, two model extensions were implemented to improve predictive performance on minority labels. First, a weighted loss function was introduced using focal loss with class-specific weights derived from label frequencies [9]. Second, synthetic images were generated using a flow matching approach to increase representation of minority classes [10]. The image generation was based on full label combinations rather than on individual minority classes.

2.3. Explainability

To assess explainability of the model, Grad-CAM maps were generated for each label independently, visualizing the image regions that drove the model's predictions [6]. For more stable results, the activations from the last three transformer encoder blocks of the model were averaged.

2.4. Fairness

Model fairness was evaluated with respect to sex and age, with age stratified into three groups: patients within one standard deviation of the mean age (majority group), patients below this range, and patients above this range. Performance differences were assessed post hoc using a bootstrapping approach described by Liang et al. [11], in which sampled matching the size of the minority group were repeatedly drawn from the majority group. The resulting distribution of performance metrics was compared with the model performance on the minority group using a z-test. Fairness analysis with respect to race was also considered. However, statistically meaningful evaluation was not feasible due to limited sample sizes in several subgroups.

3. Results

3.1. Image Preprocessing

Regarding image preprocessing, classification in this domain relies on quantifying specific features within localized regions of the image. Therefore, it was essential to preserve informative regions while also introducing variability into the training set. This was achieved by identifying an optimal scale range for random cropping. Another important aspect concerns the image quality. Applying CLAHE improved performance across all labels, with the largest improvement observed for the label capillary loss where the F1 score increased from 0.58 to 0.62. The preprocessing step enhanced the visibility of capillaries, particularly faint background structures, which may be relevant for detecting capillary loss, as illustrated in Fig. 1.

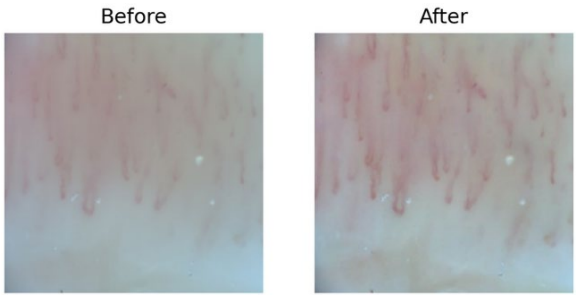


Fig. 1. Synthetic image before and after the application of CLAHE.

3.2. Model Evaluation

To address class imbalance and focus on the positive class, model performance was evaluated using the F1-score, recall, and precision. As shown in

Table 1, the weighted loss variant improved recall across all labels, while precision decreased. This indicates that the weighted loss function shifts the model toward improved detection of positive cases, at the expense of a higher false positive rate. Within the context of the evaluation of NFC images, where not missing positive cases is critical, this trade-off favors the weighted model variant. The highest overall improvement reflected in the F1 score was seen in the label enlarged capillaries which also has one of the most pronounced class imbalances of all labels.

Table 1. Performance metrics of all variants. Abbrs.: enlarged capillaries (EC), giant capillaries (GC), capillary loss (CL), microhemorrhages (MH), scleroderma pattern (SP)

Metric	F1 score			Recall			Precision		
Variant	Baseline	Weighted	Augmented	Baseline	Weighted	Augmented	Baseline	Weighted	Augmented
EC	0.32	**0.48**	0.31	0.23	**0.65**	0.23	**0.49**	0.38	**0.49**
GC	0.65	**0.66**	0.65	0.58	**0.65**	0.57	**0.75**	0.69	0.74
CL	0.62	**0.65**	0.61	0.57	**0.75**	0.56	0.68	0.59	**0.7**
MH	0.53	**0.54**	0.53	0.44	**0.55**	0.44	**0.68**	0.55	0.59
SP	0.63	**0.64**	0.63	0.55	**0.64**	0.55	**0.74**	0.67	0.73

In contrast, synthetic data augmentation did not improve performance. We hypothesize that this might be due to biases inherent in the original dataset which may have been transferred to the generated data [12]. In the context of NFC, these biases are primarily driven by technical variability, such as discrepancies in image quality which the model may replicate as structural noise. This issue is further amplified by label instability, since the ground truth is determined solely by the treating physician where the inherent difficulty of interpreting low-quality images may result in higher inter-observer variability. Furthermore, disease progression itself often introduces a bias, as advanced disease stages can lead to diminished visibility of the capillaries. Furthermore, since the generation process was conditioned on different label combinations rather than on individual labels independently, the low frequency of certain label combinations may have been insufficient for the model to reliably learn the underlying patterns.

Additionally, model performance was compared with annotations from four rheumatologists on a subset of images. Overall, human annotators achieved higher performance for most labels. However, the model performed comparably to, or slightly better than, one annotator for capillary loss and the scleroderma pattern. In this context, substantial inter-annotator variability was observed for some labels, which may indicate inherent label noise. However, it should be noted that the relatively small sample size of this evaluation may introduce additional variability. Furthermore, this comparison should be interpreted with caution, as it does not fully reflect a fair setting. Human annotators are able to incorporate prior clinical knowledge and contextual information across multiple images, whereas the model processes each image as an independent instance. A more equitable comparison would require a model capable of jointly analyzing and comparing the full capillaroscopy of a patient.

Due to the multi-label nature of the task, individual labels exhibit heterogeneous responses to changes in the model configuration. This variability may be attributed to differing levels of inherent label noise, as well as varying sensitivity to image quality across all labels. Consequently, modifications in preprocessing or training configurations can yield label-dependent effects, with some labels benefiting more than others. In particular, the label enlarged capillaries exhibiting high imbalance benefits from the weighted loss function, whereas the less imbalanced label capillary loss shows substantial improvement when preprocessing steps targeting image quality enhancement are applied. The latter observation is likely explained by the fact that assessment of capillary loss typically relies on quantifying the number of capillaries within a defined image region, making accurate detection highly dependent on the visibility of capillary structures.

3.3. Explainability

Fig. 2 shows a Grad-CAM visualization of a synthetic image labeled with the presence of giant capillaries. The importance of each region in contributing to the model's prediction is indicated by the color, with red areas indicating highly important areas. The visualization indicates that the model focuses on capillary-like structures. To our best knowledge, this represents the first application of Grad-CAM to NFC images. Such methods may offer a valuable addition in clinical application by directing physicians' attention to specific image regions, validating the physiological plausibility of the model's decision-making process and thereby increasing trust in its predictions [5]. However, these local explanations are limited to the level of individual images. Expert evaluation is required to assess the validity and clinical relevance of the highlighted regions [13].

3.4. Fairness

The z-test of the fairness analysis showed no evidence that the model is not fair with respect to sex

and age. Quantitative analysis of potential racial bias could not be reliably performed because several demographic subgroups were underrepresented in the dataset. The limited number of positive labels within certain subgroups hindered the computation of reliable fairness metrics. Consequently, it remains unclear whether these imbalances reflect random variation or a systematic bias, such as potential disparities in model sensitivity across racial subgroups. The analysis regarding race remains to be further investigated and might be an important direction for future work.

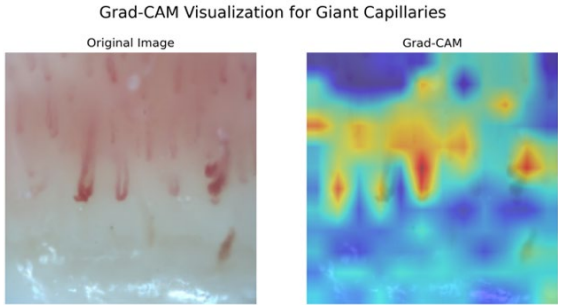


Fig. 2. Grad-CAM visualizations (synthetic image).

4. Conclusions

This study extends a Vision Transformer–based model for automated analysis of nailfold capillaroscopy images, with a focus on image preprocessing techniques, training configurations, explainability and fairness. In terms of preprocessing, the application of CLAHE notably improved the prediction results. Given the multi-label nature of the task, determining the optimal training configuration was challenging. However, evaluation of the model's performance suggests that the model variant using a weighted loss function may be most effective in identifying positive cases. Grad-CAM provides a potential tool for model interpretability in clinical contexts. However, further work is needed to assess the validity of Grad-CAM visualizations. Fairness analysis showed no evidence of performance differences across sex or age.

Although these findings suggest the potential of AI-assisted nailfold capillaroscopy analysis to support clinical assessment, such tools may be considered adjunctive to clinical expertise. Future research could assess the extent to which they can be effectively integrated into clinical practice.

Declaration of the Use of AI-based Tools

ChatGPT 5.0: Rephrasing of text.

References

[1]. A. Garaiman, F. Nooralahzadeh, C. Mihai, N. P. Gonzalez, et al., Vision transformer assisting rheumatologists in screening for capillaroscopy changes in systemic sclerosis: An artificial intelligence model, *Rheumatology*, Vol. 62, 2023, pp. 2492-2500.

[2]. C. P. Denton, D. Khanna, Systemic sclerosis, *The Lancet*, Vol. 390, Issue 10103, 2017, pp. 1685-1699.

[3]. V. Smith, E. Hysa, M. Snow, T. Frech, et al., Nailfold capillaroscopy, *Best Practice & Research Clinical Rheumatology*, Vol. 37, Issue 1, 2023, 101849.

[4]. A. L. Herrick, M. Berks, C. J. Taylor, Quantitative nailfold capillaroscopy – update and possible next steps, *Rheumatology*, Vol. 60, Issue 5, 2021, pp. 2054-2065.

[5]. M. Champendal, H. Müller, J. O. Prior, C. S. dos Reis, A scoping review of interpretability and explainability concerning artificial intelligence methods in medical imaging, *European Journal of Radiology*, Vol. 169, 2023, 111159.

[6]. R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, et al., Grad-CAM: Visual explanations from deep networks via gradient-based localization, *International Journal of Computer Vision*, Vol. 128, Issue 2, 2020, pp. 336-359.

[7]. S. M. Pizer, R. E. Johnston, J. P. Ericksen, B. C. Yankaskas, et al., Contrast-limited adaptive histogram equalization: Speed and effectiveness, in *Proceedings of the 1st Conference on Visualization in Biomedical Computing*, 1990, pp. 337-345.

[8]. A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, et al., An image is worth 16×16 words: Transformers for image recognition at scale, in *Proceedings of the 9th International Conference on Learning Representations (ICLR'21)*, 2021.

[9]. T.-Y. Lin, P. Goyal, R. Girshick, K. He, et al., Focal loss for dense object detection, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 42, Issue 2, 2020, pp. 318-327.

[10]. Y. Lipman, R. T. Q. Chen, H. Ben-Hamu, M. Nickel, et al., Flow matching for generative modeling, in *Proceedings of the 11th International Conference on Learning Representations (ICLR)*, 2023.

[11]. Y. Liang, H. Chao, J. Zhang, G. Wang, et al., Unbiasing fairness evaluation of radiology AI model, *Meta-Radiology*, Vol. 2, 2024, 100084.

[12]. B. Khosravi, S. Purkayastha, B. J. Erickson, H. M. Trivedi, et al., Exploring the potential of generative artificial intelligence in medical image synthesis: Opportunities, challenges, and future directions, *The Lancet Digital Health*, Vol. 7, 2025, 100890.

[13]. B. H. M. van der Velden, H. J. Kuijf, K. G. A. Gilhuijs, M. A. Viergever, Explainable artificial intelligence (XAI) in deep learning-based medical image analysis, *Medical Image Analysis*, Vol. 79, 2022, 102470.

(016)

Generating a New Objective Index (SURG-STRI) to Evaluate the Surgical Stress from ECG Sensor Data

D. Caballero [1], M. J. Pérez-Salazar [1], J. A. Sánchez-Margallo [1] and F. M. Sánchez-Margallo [2]
[1] Robotic, Image-guided surgery and Bioprinting Unit, Jesús Usón Minimally Invasive Surgery Center, N-521 Road Km. 41.8, ES-10071 Cáceres, Spain
[2] Scientific Direction, Jesús Usón Minimally Invasive Surgery Center, N-521 Road Km. 41.8, ES-10071 Cáceres, Spain
Tel.: +34 927 181032
E-mail: jasanchez@ccmijesususon.com

Summary: Traditional methods to evaluate surgeons' stress during their surgical procedures have been established subjectively. In this study, a new objective index is proposed to measure stress during Minimally Invasive Surgery performance, called Surgery Stress Index (SURG-STRI). This new index will be based on the ECG sensor characteristics measured on the surgeons during surgical activities. To this end, data were gathered during 44 surgical sessions completed by 18 surgeons. Once the data were collected, two preprocessing techniques were applied, scaled, and normalized. The dataset was splitted into 80 % for training and cross-validation and 20 % for testing. The results showed that the model using Multiple Linear Regression achieved the best performance. This model was successfully validated, demonstrating the possibility for evaluating stress objectively. These results enable an innovative method for measuring stress objectively, which could help prevent health risks for surgeons during surgical procedures, thereby improving both, healthcare quality and surgical performance.

Keywords: Minimally invasive surgery, Wearable device, Stress levels, Artificial intelligence, Stress index, Predictive analysis.

1. Introduction

Minimally invasive surgery (MIS) procedures have grown its application in recent years. The main advantages of MIS are described widely in the scientific literature [1]. However, there are some limitations for the surgeons, such as, the high stress levels managing and ergonomic deficiencies are still addressed. These unsolved deficiencies that can be detrimental to the surgeon's health and healthcare quality, and consequently, for the surgical performance.

The preventive identification of high stress levels may potentially reduce the health risks. This is especially critical in clinical environments, since it has implications for the surgical quality and performance [2]. In this context, some advances have been developed for monitoring stress levels in the surgical research field [3].

To determine stress levels during MIS performance, the control of ergonomic and physiological parameters is essential. Some wearable devices, including electrocardiogram sensors (ECG) that were analyzed for recording and evaluating stress level. These systems allow us to quantify, among others, parameters related to ergonomic, kinematics or physiology [4]. In this way, an objective and robust solution can be provided the stress analysis in an objective way as a function of the ergonomic, physiological and kinematic parameters of the surgeon during MIS procedure.

The application of Artificial Intelligence (AI) has widely grown in its use, implementation, and development. Among the different AI techniques, there are several algorithms with predictive purposes [1]. Specifically, some systems are specialized on the identification of primary factors associated with high stress levels during MIS procedures. Three AI algorithms were tested with predictive purposes in the current study, with a linear approach, a non-linear approach and a machine learning approach [5].

The absence of stress objective index during the development of MIS activities generates a need to monitor the stress objectively. For that, a new objective stress index is proposed in order to measure the surgeon's stress during the development of surgical activities: the Surgery Stress Index (SURG-STRI). This index is based on ergonomic, kinematic and physiological characteristics of the surgeon, measured with the ECG sensor. The proposed objective stress index is the main novelty of this study. This approach is different to previous studies that developed an objective index based on other sensors such as EDA [6] and the test of several AI techniques on ECG sensor data to define the most effective AI technique for predicting surgical stress levels [7].

Therefore, the objective of this study is to propose a new objective index (SURG-STRI) that allows quantifying the stress level from physiological, ergonomic and kinematic parameters from ECG sensor of the surgeons during the development of surgical procedures. This study also aims to validate the proposed stress index, comparing these values to the SURG-TLX stress index values after the end of the surgical activities.

2. Material and Methods

The data for this study were gathered during 44 MIS sessions, 18 with conventional laparoscopy surgery and 26 with Robot-Assisted Surgery (RAS). A total of 18 surgeons (4 women and 14 men, between 25 and 50 years old) participated in this study. Ten surgeons have an experienced level in conventional laparoscopy surgery (> 100 laparoscopic surgical procedures) and the remaining have an intermediate-novices level (< 100 procedures). All participants have an intermediate-novice on RAS (< 100 procedures). Fig. 1 shows the experimental design.

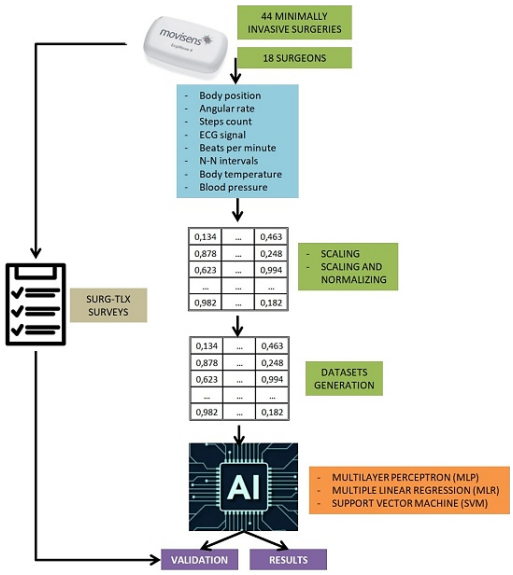


Fig. 1. Experimental design.

2.1. Wearable Device

The ergonomic, kinematic, and physiological data were recorded with a single wearable device (ECGMove 4 sensor, Movisens®, Karlsruhe, Germany). The ergonomic, kinematic and physiological data extracted from the device were processed by using the Unisens Viewer software (Movisens®, Karlsruhe, Germany). This device allows measurements on body position, angular rate, steps count, usual electrocardiogram, beats per minute, N-N intervals, body temperature and blood pressure. The device was attached to the left side of the participants' torso under the pectoral muscle by means of two pectoral electrodes.

2.2. Conventional Laparoscopy

For conventional laparoscopy, Olympus® VISERA ELITE (Olympus®, Tokyo, Japan) imaging system and Karl Storz® laparoscopic instruments (Karl Storz SE and Co., Tuttlingen, Germany) were used. 18 sessions were performed with conventional laparoscopy.

2.3. Robotic Platforms

For RAS were used two different robotic platforms, the robotic platform Versius® (CMR surgical®, Cambridge, United Kingdom) for laparoscopy approach and Symani® (MMI®, Pisa, Italy) for microsurgery. Twenty-six sessions were performed with RAS, 14 of them were performed with Symani® and the remaining were developed with Versius®.

2.4. Datasets

The original dataset was transformed by applying scaled and normalized preprocessing techniques. Once the datasets were preprocessed, these ones were divided into four datasets: 80 % of data for training and cross-validation and 20 % of data for test. For the cross-validation, this splitting process was carried out taking into account the same surgeon and surgical session.

Scaled preprocessing technique allows each parameter to be described on a scale between 0 and 1. For normalized preprocessing technique transform the dataset into a more integrated and robust one with fewer redundancies.

The R^2 coefficient was used to evaluate the goodness the fit of the predictive models and for validation, according to the rules given by Colton [8], where R^2 of 0 to 0.25 is considered as a poor to no relationship; 0.25 to 0.50 indicates a weak degree of relationship; 0.50 to 0.75 designates a moderate to good relationship; and 0.75 to 1 shows a very good to excellent relationship. The root mean square error (RMSE) was also used to validate the prediction results [9]. RMSE measures the difference between actual and predicted values. RMSE values of less than 0.05 are considered adequate [9].

2.5. Subjective Stress Index

After the end of each of the 44 surgical sessions, surgeons were asked to complete a SURG-TLX survey about the level of stress [6] that they had felt during different moments of the surgical performance.

2.6. New Objective Stress Index

For the development of the new objective stress index (SURG-STRI), three AI-based predictive techniques were tested to identify the most suitable approach. The first used linear predictive models (Multiple Linear Regression –MLR–). MLR is a predictive technique that consists in representing the linear relationship between a dependent variable and several independent variables. The configuration used was M5 method for the feature selection and a ridge value of 1×10^{-4} [10]. Another predictive approach used non-linear predictive models (Support Vector Machine –SVM–). SVM is based on the

transformation of the vector space in other higher dimensional space, in which the model can be found to solve the problems [10]. A radial basis function was used as the kernel with its default parameter values. The last predictive model was based on with artificial neural networks (ANN), specifically a Multilayer Perceptron –MLP–. MLP is a type of ANN that tries to simulate the human brain connections. In this way, the model is based on the values and signals that completed the ANN. Therefore, MLP was configured with 500 epochs, a learning rate of 0.01, and a momentum rate of 0.05. The PyTorch library from Python 3.13.7 programming language was used for generating the predictive models.

3. Results and Discussion

3.1. Training Results

Table 1 shows the results of the training dataset for the predictive models. The best results were achieved for MLR as AI technique, and Scaled as preprocessing technique (R^2 = 0.9125, RMSE = 0.0503) followed by MLP (R^2 = 0.7801, RMSE = 0.1601) and with SVM (R^2 = 0.3884, RMSE = 0.2887) [8]. The results are in accordance with previous studies, obtaining the most favourable results with the MLR model [6].

Table 1. R^2 and RMSE for the training dataset.

	Stress – R^2 (RMSE)	
	Scaled	**Scaled – Norm**
MLP	0.7801 (0.1601)	0.7451 (0.1855)
MLR	**0.9125 (0.0503)**	0.8775 (0.0757)
SVM	0.3884 (0.2887)	0.3534 (0.3141)

3.2. Cross-Validation Results

Table 2 shows the results of the cross-validation for the predictive models. The best results were achieved for MLR as AI technique, and scaled as preprocessing technique (R^2 = 0.8924, RMSE = 0.0591) followed by MLP (R^2 = 0.7785, RMSE = 0.2093) and with SVM (R^2 = 0.3782, RMSE = 0.2903) [8]. The results are in line with previous studies, obtaining the best results using the MLR model [6].

Table 2. R^2 and RMSE for the cross-validation.

	Stress – R^2 (RMSE)	
	Scaled	**Scaled - Norm**
MLP	0.7785 (0.2093)	0.7435 (0.2347)
MLR	**0.8924 (0.0591)**	0.8574 (0.0845)
SVM	0.3782 (0.2903)	0.3432 (0.3157)

3.3. Test Results

Table 3 shows the results of the test dataset for the predictive models. The best results were achieved for MLR as AI technique, and scaled as preprocessing technique (R^2 = 0.8723, RMSE = 0.0677) followed by MLP (R^2 = 0.7769, RMSE = 0.2096) and with SVM (R^2 = 0.3684, RMSE = 0.2918) [8]. The results are in agreement with previous studies, obtaining the best results by using the MLR model [6].

Table 3. R^2 and RMSE for the test dataset.

	Stress – R^2 (RMSE)	
	Scaled	**Scaled - Norm**
MLP	0.7769 (0.2096)	0.7435 (0.2349)
MLR	**0.8723 (0.0677)**	0.8373 (0.0931)
SVM	0.3684 (0.2918)	0.3334 (0.3172)

3.4. Preprocessing Results

Table 4 shows the results of the preprocessing techniques, by applying MLR as AI technique and the training dataset, the cross-validation results and the test dataset.

Table 4. Comparative of preprocessing techniques.

	Stress – R^2 (RMSE)	
	Scaled	**Scaled – Norm**
Training dataset	**0.9125 (0.0503)**	0.8775 (0.0757)
Cross-validation results	0.8924 (0.0591)	0.8574 (0.0845)
Test dataset	0.8723 (0.0677)	0.8373 (0.0931)

The best results in all cases were reached by scaled as preprocessing technique (R^2 = 0.8723, RMSE = 0.0677), followed by scaled and Normalized (R^2 = 0.8373, RMSE = 0.0931). The results are in accordance with previous studies with similar results, in which scale obtained the best results as preprocessing technique [6].

3.5. Model Validation

Taking into account the obtained result in the current study, the authors have been decided to develop the predictive model by using MLR as AI technique and scaled as a preprocessing technique, for the new SURG-STRI objective stress index based on ECG parameters.

Equation (1) presents the predictive model of the objective index for stress characterization SURG-STRI. In this model, the feature with the greatest weight in stress is beats per minute and body temperature. On the other hand, the characteristics with the least weight are the angular rate on X axis and the body position on Z axis.

In future studies, considering the promising results obtained and the fact that this is a preliminary study, a

comparison could be made with a model using only the most representative parameters.

$$\begin{aligned} Stress = & -0.082 \\ & * Body\ position\ on\ X\ axis - 0.421 \\ & * Body\ position\ on\ Y\ axis - 0.064 \\ & * Body\ position\ on\ Z\ axis + 0.008 \\ & * Angular\ rate\ on\ X\ axis + 0.144 \\ & * Angular\ rate\ on\ Y\ axis - 0.327 \\ & * Angular\ rate\ on\ Z\ axis - 1.574 \\ & * Beats\ per\ minute + 0.467 * ECG\ signal - \\ & 0.047 * Steps\ count - 1.396 \\ & * NN\_intervals - 0.466 \\ & * Blood\ pressure + 1.255 \\ & * Body\ Temperature + 1.025. \end{aligned} \quad (1)$$

Fig. 2 displays the results of the comparison between the values obtained by the SURG-TLX (Real Values – X Axis) and the values obtained by the SURG-STRI (Predicted Values – Y Axis). The R2 = 1 line is also shown to evaluate the deviation of the SURG-STRI values respect to the SURG-TLX values, as well as the 95 % confidence intervals lines. This demonstrates the high correlation between the values obtained by our objective index (SURG-STRI), based on our linear model, and the subjective values provided by surgeons in the SURG-TLX surveys after each surgical activity. As can be seen, most stress levels are below 6, although high stress levels with values above 8 were also successfully predicted. The green boxes indicate the measures evaluated as residuals, which should be removed to improve the model in future studies.

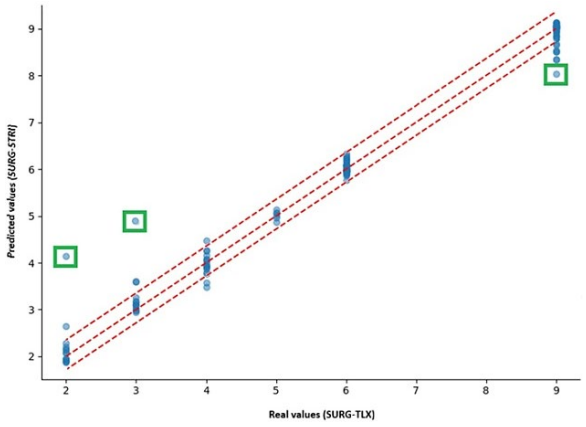


Fig. 2. Scatter plot of SURG-TLX values (real values) versus SURG-STRI values (predicted values) to validate our predictive model with the confidence intervals and with outliers marked in green boxes.

4. Conclusions

In this study, a new objective stress index (SURG-STRI) has been successfully developed, based on predicting stress from parameters related to ECG sensor during MIS procedures. The best results were reached by using MLR as AI technique and scaled as preprocessing technique. These results validate a new method to evaluate stress during MIS practice in an innovative and novel way.

Acknowledgments

This work has been financed by the Ministry of Science and Innovation with funds from the European Union Next Generation EU, from the Recovery, Transformation and Resilence Plan (PRTR-C17.I1) and the European Regional Development Fund (ERDF) of Extremadura Operational Program 2021-2027.

References

[1]. D. Caballero, J. A. Sánchez-Margallo, M. J. Pérez-Salazar, F. M. Sánchez-Margallo, Applications of artificial intelligence in minimally invasive surgery training: A scoping review, *Surgeries*, Vol. 6, Issue 1, 2025, 7.

[2]. A. M. Hurley, P. J. Kennedy, L. O'Connor, T. G. Dinan, et al., SOS save our surgeons: Stress levels reduced by robotic surgery, *Gynecological Surgery*, Vol. 12, Issue 3, 2015, pp. 197-206.

[3]. D. Caballero, M. J. Pérez-Salazar, J. A. Sánchez-Margallo, F. M. Sánchez-Margallo, 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 10, 2024, pp. 1953-1963.

[4]. F. Saoughi, A. Behmanesh, N. Sayfouri, Internet of things in medicine: A systematic mapping study, *Journal of Biomedical Informatics*, Vol. 103, 2020, 103383.

[5]. J. F. Ávila-Tomás, M. A. Mayer-Pujadas, V. J. Quesada-Varela, La inteligencia artificial y sus aplicaciones en medicina I: Introducción, antecedentes a la IA y sus aplicaciones en robótica, *Atención Primaria*, Vol. 52, Issue 10, 2020, pp. 778-784 (in Spanish).

[6]. D. Caballero, M. J. Pérez-Salazar, J. A. Sánchez-Margallo, F. M. Sánchez-Margallo, Creación de un nuevo índice objetivo (SURG-STRI) para la evaluación del estrés quirúrgico a partir de datos del sensor EDA, in *Proceedings of the 43rd Annual Conference of the Spanish Society of Biomedical Engineering (CASEIB)*, 2025 (in Spanish).

[7]. D. Caballero, M. J. Pérez-Salazar, J. A. Sánchez-Margallo, I. Díaz-Romero, et al., Application of machine learning and deep learning methods on ECG sensor data to predict stress levels in minimally invasive surgery, *Lecture Notes in Computer Science*, Vol. 16148, 2026, pp. 185-199.

[8]. T. Colton, Statistics in Medicine, *Little Brown and Co*., 1974.

[9]. R. Hyndman, A. B. Koehler, Another look at measures of forecast accuracy, *International Journal of Forecasting*, Vol. 22, Issue 4, 2006, pp. 679-688.

[10]. U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, From data mining to knowledge discovery in databases, *AI Magazine*, Vol. 17, Issue 3, 1996, pp. 37-54.

(017)

Stereo-Vision Localization of Millimeter-Scale Capsule Targets in a Clinically Inspired Maze: A Repeatability Study

I. Frajtag [1], L. Masjosthusmann [2], S. Misra [2] and F. Šuligoj [1]
[1] University of Zagreb, Faculty of Mechanical Engineering and Naval Architecture,
Ivana Lučića 5, 10000 Zagreb, Croatia
[2] University of Twente, Faculty of Engineering Technology, Department of Biomechanical Engineering,
Surgical Robotics Laboratory, Drienerlolaan 5, 7522 NB Enschede, Netherlands
E-mail: ines.frajtag@fsb.unizg.hr

Summary: Reliable localization of magnetically actuated microrobots and capsule robotic systems is essential for safe navigation in complex anatomical pathways, yet remains challenging in tortuous and occlusion-prone environments. This study presents a stereo vision framework that combines YOLO-based detection in two camera views with triangulation to reconstruct metric 3D target center positions. When YOLO-OBB is used, the same framework also provides an orientation estimate that can be translated into a headin cue. The system is evaluated in a maze phantom designed to mimic bowel- and vessel-like pathways, using five targets of different sizes and geometries. On a test set, YOLO-OBB improves detection reliability over standard YOLO (precision/recall 95.6/94.5 % vs. 86.1/85.9 %), maintains comparable center error (4.31 vs. 4.22 px mean), and reducing CPU inference time by half. Repeatability analysis across ten maze points showed lower total 3D error for cylindrical than symmetric targets (E_{total} = 0.91/2.33 vs. 2.40/11.1), indicating higher robustness for navigation-relevant localization in medically relevant settings.

Keywords: Vision system, Microrobots, Detection, Error, Repeatability.

1. Introduction

Over the past decade, robotics has transitioned from specialized systems to a growing presence in routine clinical workflows. Robotic assistants are now used across many surgical disciplines [1-3] to support demanding procedures and improve precision. In parallel, the drive toward minimally invasive diagnosis and therapy has accelerated the development of microrobots and capsule robotic systems that can operate inside hard-to-reach anatomical spaces, including pathways such as the gastrointestinal (GI) tract and vasculature [4]. Among these technologies, magnetically actuated devices are particularly attractive because external magnetic fields can penetrate biological tissue and interact with magnetic components inside the device to generate force and torque locally. This allows the actuation source to remain outside the body, enabling compact and fully internalized device designs. As a result, magnetically actuated microrobots and capsule robots are increasingly explored for applications such as targeted drug delivery, localized therapy, and *in situ* diagnostics [5]. In this context, magnetically actuated microrobots and capsule robotic systems are especially relevant, but their practical deployment strongly depends on reliable localization and orientation tracking to support safe navigation in complex, occlusion prone anatomy. These requirements become especially strict in confined anatomies where visibility is intermittent and trajectories are highly curved. Recent surveys and experimental studies [6, 7] describe multiple localization modalities [8], such as magnetic sensing and magnetic localization, vision-based approaches such as visual odometry/SLAM or learning-based pose estimation from image sequences, as well as hybrid sensor-fusion strategies. Each modality entails practical constraints related to system complexity, susceptibility to artifacts, and achievable spatial and temporal resolution. Hybrid and vision-based solutions are appealing because they can provide high spatial detail using relatively low-cost imaging hardware. Yet robust performance remains difficult under clinically realistic conditions. Tortuous geometry, occlusions, reflections and low-texture scenes can all degrade tracking - especially for millimeter-scale targets.

In this work, we develop and evaluate a stereo-vision localization framework designed for a clinically inspired scenario: a maze phantom that emulates the vision challenges of gastrointestinal tract and vasculature anatomy. A central component of the framework are YOLO detectors trained on a dedicated dataset of capsule targets spanning multiple sizes and geometries. The detectors are used to detect the target in each camera view and to extract its 2D center; with YOLO-OBB (Oriented Bounding Box), it also provides an in-plane orientation, which we use to infere a heading cue. Given the synchronized left/right detections (cameras on Fig. 1a**)**, we reconstruct the 3D target center via stereo triangulation and, when applicable, track position and heading. Our work focuses on repeatability and interpretable total error of capsule robotic systems of different sizes and shapes because absolute accuracy is limited by manual target placement and the absence of an external high-precision ground-truth tracking system. We

decompose the error into variability induced by manual repositioning across repeated trials, within-trial measurement noise while the target remains stationary, and a detection/geometry related error. This separation enables identification of the dominant error source and provides an interpretable total error for navigation tasks, where stable relative localization is often more critical than absolute accuracy.

2. Related Research

Reliable pose estimation of the device is widely recognized as key enabler for magnetically actuated microrobots and capsule robotics systems, as they directly determine whether closed-loop control and re-navigation can be performed safely in confined, deformable anatomy. Recent surveys emphasize that no single sensing modality is universally reliable. Magnetic and vision-based approaches all face practical limitations when moving from controlled scenarios to clinically realistic conditions, including calibration burden, sensitivity to disturbances, and reduced robustness under occlusions and specular reflections.

A large body of work therefore uses camera-based perception as a first step, because it can provide fine spatial detail with relatively lightweight hardware. However, maintaining stable detection and tracking is challenging when the target is small and the scene appearance varies strongly with viewpoint. Beyond standard camera views, several studies focus on pose/orientation estimation under microscope or otherwise constrained optical setups. Choudhary *et al*. [9] propose a deep learning-based method for three-dimensional microrobot orientation estimation and tracking, enabling richer pose feedback than only object center localization. Related work also addresses 3D pose estimation for visually challenging targets, like transparent or low-contrast microrobots, by inferring depth and orientation from microscope images, where classical handcrafted visual cues are often unreliable [10]. While these approaches advance pose observability, they are validated in highly controlled imaging settings tied to a specific optical system, and experiments are not always structured to separate repeated repositioning, multi-location effects, and within-trial measurement noise. In addition to optical imaging, clinically relevant modalities such as ultrasound have also been investigated. Liu *et al*. [11] address capsule robot pose and mechanism-state detection in an anatomically representative environment using ultrasound imaging and deep learning, demonstrating that clinically relevant scenarios introduce failure modes that must be explicitly handled. Finally, datasets such as USMicroMagSet [12] highlight the importance of standardized evaluation for small device tracking under difficult imaging conditions, reinforcing that performance comparisons should be tied to well-defined metrics and protocols.

Against this background, machine learning has increasingly been adopted to enhance robustness in the visual perception stage. Deep learning object detectors such as YOLO [13] have become common in medical vision tasks due to their favorable balance between accuracy and speed, as well as the ease of deployment. In the microrobotics context, Li *et al*. [14] demonstrate real-time microrobot detection and tracking using deep-learning models, at the same time illustrating that YOLO detectors can remain effective even when the target is small, and the scene is visually challenging. Many vision-only demonstrations primarily report detection/tracking success, while less frequently quantifying how detection variability propagates into navigation-relevant 3D stability across repeated trials and multiple spatial locations. Hybrid and sensor-fusion pipelines are often advocated to mitigate modality-specific weaknesses and retain metric consistency under realistic disturbances [15].

In contrast to prior works, our study combines a calibrated stereo setup with YOLO-based detectors to introduce an evaluation protocol that explicitly quantifies repeatability by decomposing localization error into detection error, manual repositioning variability and within-trial measurement noise. This design directly supports navigation interpretation: not only whether the detector works, but how stable and predictable the resulting 3D localization is across repeated trials and multiple maze locations.

3. Materials and Methods

The goal of this work is to assess the detection of robustness and tracking repeatability in a clinically inspired experimental scenario. We are using a designed setup, shown in Fig. 1a, composed of a maze phantom and a calibrated stereo vision system with the dimensions shown in Fig. 1b.

The maze phantom is fabricated from plexiglass and rigidly mounted on a magnetized base plate, providing a stable and reproducible substrate during repeated trials. To reduce the impact of unwanted optical effects, the maze is lined with a white background layer, which improves contras and helps reduce the influence of reflections. The stereo rig remains unchanged throughout calibration and data collection sessions. Camera intrinsic, stereo extrinsic, and the world reference frame are obtained using an OpenCV ChArUco board pipeline, where ChArUco detections are used for per-camera intrinsic calibration, stereo calibration between C0 and C1, and a fixed board pose to estimate the rigid transform from C0 to the world coordinate system W, Fig. 1a. To support learning-based perception in this clinically inspired maze, we made a stereo dataset tailored to the geometry and materials of the experimental setup. The dataset is collected using both camera views and is designed to cover full experimental matrix: five capsule robots represented by magnets of different sizes and geometry (Fig. 2d).

All the recorded images are annotated in Roboflow using polygon-based labels to capture the true target extent for elongated and rotated objects. At each location, mentioned as P1-P10 in Fig. 1c, the magnet was manually repositioned, and the procedure was repeated 10 times to capture across-trial variability; within each repetition, 100 frames were recorded to characterize within-trial measurement fluctuations while the target remained stationary. Two deep learning detectors were evaluated in this work: YOLOv12n and YOLOv12n-OBB. We use standard YOLO model as a strong, widely adopted baseline for reliable and efficient 2D center detection and we use OBB to test whether explicitly modeling target orientation (via oriented bounding box) improves robustness and provides an additional heading cue- thereby quantifying the benefit of orientation-aware detection for navigation in curved, occlusion-prone scenes. For cylindrical targets, the 3D heading vector is obtained by triangulating the endpoints of the major axis estimated by OBB in the two camera views, and heading error is computed as the angular difference between the estimated and reference 3D direction. The standard YOLO and the OBB model were trained using the same configuration, with identical training parameters and a confidence threshold of 0.5 for inference. All experiments are conducted on a research workstation running Ubuntu 22.04, equipped with a 64-bit Intel Core i5 CPU operating at 2.80 GHz. In our system, the OBB orientation is used together with the object's 2D center to provide an orientation cue from both camera views. Quantitative 3D heading estimation is not the primary focus of the present study.

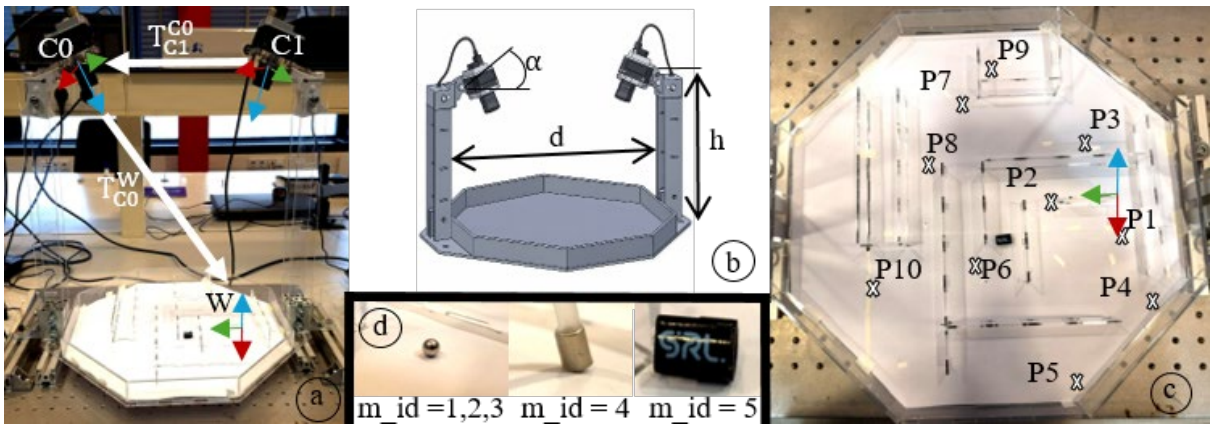


Fig. 1. a) Clinically inspired maze phantom setup with a calibrated stereo vision rig. Two fixed cameras define the camera coordinate frames C0 and C1. The stereo calibration provides rigid transformation between cameras, $\boldsymbol{T_{C1}^{C0}}$, while an extrinsic calibration defines the transformation from camera 0 to the world coordinate system, $\boldsymbol{T_{C0}^{W}}$. The world coordinate system – W is used as the reference for reporting metric 3D positions and repeatability results. The orientation of coordinate axes and the direction of transformations used in the reconstruction process is shown with arrow, b) Monochrome Basler cameras, model avA1000-100gm, are used. Each camera is operating at resolution 1024x1024 with 6 mm focal length lenses. Cameras are mounted at a fixed height of h = 420 mm and tilted by 28° to satisfy the required field of view and working distance. The maze area observed by the stereo pair corresponds to an approximately circular field of view with a diameter d = 350 mm, c) Top view of the maze phantom with the 10 predefine spatial locations (P1 – P10) used for repeatability testing. The marked points span straight segments and curved regions, capturing visibility changes, occlusions, and reflections induced by the maze geometry. These positions define systematic evaluation grid for comparing target-dependent and location-dependent effects, d) Representative targets used in the study. Capsule /microrobots surrogate were realized using five magnets of different size and shapes. Circular magnets are m_id = 1, 2, 3, and their dimensions are: m_id = 1 has diameter 2 mm, m_id = 2 has diameter 3 mm, and m_id = 3 has 5 mm. Two magnets are cylindrical sizes, m_id = 4 (Ø3.5 x 6 mm) and m_id = 5 (Ø9 x 15 mm). This set allows evaluation of detection and localization repeatability across varying target appearance, scale and aspect ratio.

4. Results

We assessed detection performance and positioning error in a bowel-/vessel-like maze phantom that mimics clinically relevant visual constraints, using multiple target locations that would be difficult to probe in vivo. Our goal was to determine which factor has the greatest impact on positioning accuracy and which error sources dominate in repeatability.

Detection performance was evaluated on a test image set using a standard YOLOv12n detector and a YOLOv12n-OBB detector. The annotated dataset comprises 10 948 frames and was split 80/10/10 into train/validation/test, yielding a test set of 1094 images used for the quantitative results in Table 1. We report IoU (Intersection-over-Union) based overlap statistics to quantify spatial agreement between predicted regions and the reference annotations. IoU is defined as the ratio between the intersection area and the union area of the prediction and the ground truth. To summarize the IoU distribution, we report IoU_{mean}, IoU_{p95} and the proportion of detections with IoU≥0.75, a commonly used criterion for tight localization. On the test set, IoU statistics indicate that the standard

YOLO achieves slightly tighter spatial alignment with the ground-truth regions than OBB, as reflected by higher average IoU, a higher p95 values of IoU and a larger fraction of detections exceeding the IoU≥0.75 threshold.

In contrast, OBB shows a modest drop in IoU overlap, which is expected because the overlap is more sensitive to small orientation deviations. Even minor angle errors can reduce IoU while detection remains visually correct. Because our downstream objective is stereo triangulation, we prioritize detection properties that directly affect 3D reconstruction: image-level detection rate, stability of the estimated 2D center, and for OBB orientation error, since it provides an angle estimate. Overall, OBB yields more reliable image-level detections and a more consistent precision-recall behavior, which is reflected in the mAP metrics reported in Table 1**.** It also triggers fewer false positives in difficult scenes with partial occlusions and specular reflections, compared with the standard YOLO model. For stereo triangulation, two properties matter most: consistent detection in both camera views and stable 2D center localization. In our experiment, both detection models achieve comparable center accuracy in the image plane. To better characterize center stability beyond mean/p95, we report the full distribution of center errors using an empirical cumulative distribution function (CDF), Fig. 2b, which reveals how frequently small pixel-level errors occur.

Table 1. Detection and 2D localization performance of the standard YOLO model and the YOLO-OBB evaluated on the test set.

Metric	YOLOv12n	YOLO-OBBn
IoU_{mean}	0.72	0.70
IoU_{p95}	0.96	0.93
IoU≥0.75	0.51	0.49
Test images	1094	
Detection rate	98.61 %	99.11 %
Precision	86.11 %	95.6 %
Recall	85.93 %	94.5 %
mAP50	85.74 %	95.25 %
mAP50-95	48.61 %	66.53 %
Center error mean/p95 (px)	4.22/11.68	4.31/11.99
Angle error mean/p95 (°)	-	4.63/11.1
Inference time mean (ms)	427.33 (~3 FPS)	203.90 (~5 FPS)

Since IoU is often used as a localization quality proxy, we further analyze how IoU relates to the 2D center error for OBB detections, Fig. 2c. It shows that higher overlap generally corresponds to smaller center errors, but with increased variability at lower IoU. This suggests that the primary performance gap is driven by detection robustness (whether the target is detected at all), rather than by systematic differences in the predicted center once a detection is present. A key benefit of OBB is the additional orientation signal. This enables estimation of target heading and extends localization beyond 3D position toward direction estimation, which is directly relevant for navigation in narrow and tortuous paths where heading informs steering and collision avoidance. It is important to note that orientation is not equally meaningful for all target geometries. For circular or symmetric targets, the heading can be ambiguous because multiple orientations look identical in the image. In those cases, even a correct detection can yield an unstable OBB angle, since small appearance changes can rotate the fitted box without reflecting a true physical rotation. This is why orientation estimates are inherently more reliable for elongated targets, which provide stronger directional visual cues. In our navigation setting, this implies that pose estimation should be interpreted as a robust 3D position estimate for all targets, and a reliable heading estimate primarily for non-symmetric targets.

Runtime was measured on a CPU configuration. Even under these constraints, OBB runs approximately two times faster than the standard model in our setup. Although the current system is not yet real-time, these measurements provide a clear CPU baseline and motivate GPU deployment in future work.

For navigating a microrobot or a capsule the stability of the 3D pose estimate across repetitions is critical. The system must return to a consistent location and orientation to support motion planning, heading correction, and revisiting regions of interest. to quantify this, we designed an experiment with 5 magnets (representing different capsule/microrobot geometries) and 10 spatially defined points within the maze. Each point was repeated 10 times per magnet. Within each repetition, we recorded 100 frames while the object remains stationary. This design lets us separately quantify variability across repetitions and variability within a repetition under stationary conditions. Table 2 reports the 3D localization errors grouped by magnet type. As expressed in Equation (1), the total error is decomposed into components associated with repositioning variability, within-trial noise, and repeatability, allowing a more interpretable analysis of localization performance.

$$E_{total} = \sqrt{E_{place}^2 + E_{noise}^2 + E_{det}^2} \quad (1)$$

All components are reported as mean/p95. The mean summarizes average behavior, and the p95 value indicates a conservative bound below which most measurements fall, and it is useful for safety-aware navigation. The results show that total error is primarily driven by manual repositioning. Within-trial noise is smaller and reflects the intrinsic stability of the measurement process when the target does not move. This indicates that the largest improvements in repeatability will come from standardizing the placement procedure and reducing manual variability, for example through fixtures, guided placement, or automated positioning. At the same time, the measurement process itself behaves relatively

consistently during stationary measurements, which is encouraging for close loop use once placement variability is controlled. Finally, the spatial dependence of the worst-case 3D localization error is visualized as a heatmap of E_{total} at p95 across all points and magnet types, Fig. 2a, highlighting specific maze locations that act as error hotspots.

We also observe a clear dependence on target geometry, Fig. 2a. Circular/symmetric capsules are more sensitive to imaging conditions such as bends, partial occlusions, and specular reflections. This typically appears as higher across repetition variability, because small changes in viewpoint and visibility produce larger changes in the detected outline and thus in triangulated depth. Cylindrical targets provide a stronger, more stable visual signature across viewpoints. The largest cylindrical magnet (m_id = 5) achieves the best repeatability and the most stable total error, consistent with its higher visual salience and more reliable detections in both camera views. The smaller cylindrical target (m_id = 4) preserves the same geometric advantage but is more sensitive to detection instability and center jitter under occlusions and reflections, because its projected image area is reduced and the influence of background clutter becomes relatively larger.

Table 2. Decomposition of 3D localization error in the maze phantom for each magnet type, reported as mean/p95 (mm) across all evaluated points and repetitions using the final selected model, YOLOv12n-OBB. The p95 statistic provides a conservative upper-tail summary that is particularly relevant to navigation robustness. We decompose errors into interpretable components. E_{place} captures variability induced by manual repositioning between repetitions. E_{noise} captures within-trial measurement noise while the target is stationary. E_{repeat} summarizes repeatability at the same point (how constantly the system returns to the same 3D location over repeated placements). E_{total} represents the total error.

m_id	E_{place} [mm]	E_{noise} [mm]	E_{repeat} [mm]	E_{total} [mm]
1	1.21/8.34	0.28/1.37	1.36/8.38	1.64/8.52
2	1.80/8.34	0.28/0.99	1.85/8.38	2.06/8.52
3	1.33/6.80	0.49/2.61	1.70/8.45	2.40/11.1
4	1.54/7.44	0.18/0.91	1.58/7.59	1.77/7.78
5	0.45/1.11	0.48/2.02	0.72/2.28	0.91/2.33

5. Conclusions

In this paper, we present an evaluate a stereo-vision-based experimental setup for capsule robotic systems in a bowel-/vessel-inspired maze phantom, intended to support the testing and validation of localization and future control strategies, with emphasis on repeatability rather than absolute accuracy. The proposed pipeline combines YOLO-based 2D detections in both camera views with stereo triangulation to reconstruct the 3D target center, while YOLO-OBB additionally provides an estimate of target heading. On a test set, YOLO-OBB demonstrates superior detection reliability compared with a standard YOLO model, achieving higher precision/recall and overall detection quality, while maintaining comparable 2D center localization accuracy. These properties translate into a higher likelihood of obtaining valid, synchronized detections in both views, which is critical for robust stereo triangulation and consistent 3D reconstruction.

Beyond detector performance, a key contribution of this work is an interpretable repeatability analysis that decomposes 3D localization variability into variability induced by manual repositioning and within-trial measurement noise while the target remains stationary. This decomposition provides a practical total error that aligns with navigation requirements, where stable relative localization is often more important than global absolute accuracy. These results indicate that within-trial noise is low across most maze regions, whereas the upper-tail error is primarily driven by manual repositioning variability and localized failure modes. Heatmap and worst-case analyses identify a few spatial hotspots that dominate the error tail under visually challenging conditions. Finally, the experiment highlights a clear geometry dependence: cylindrical targets exhibit more robust behavior, while symmetric provide limited directional cues, making orientation estimates inherently less informative and more sensitive under adverse imaging conditions. Overall, the framework supports stable relative localization (and heading where meaningful) across most maze segments, establishing a solid basis for closed-loop navigation of magnetically actuated devices in tortuous anatomy.

Future work will focus on reducing dominant error sources and increasing real-time capability. We will introduce automated repositioning to isolate sensing error from setup variability. Second, we will improve robustness in the identified hotspot regions by collecting additional data. Hard-negative mining and failure-case reweighting will be explored to further suppress false positives in reflective scenes. The full process will be running on GPU and optimize end-to-end latency, with the goal of achieving update rates suitable for closed-loop control. Finally, we will extend evaluation to more diverse lighting, materials, and phantom configuration, and investigate multimodal fusion (magnetic sensing an/or inertial cues) to increase robustness under clinically realistic variability.

Acknowledgements

This research was funded by the project INSPIRATION non-INvaSive PatIent RegistrATIon for rObotic Neurosurgery (grant no. NPOO.C3.2.R2-I1.06.0153), financed by the European Union through the National Recovery and Resilience Plan (NPOO).

This work was supported by the European Commission under the Horizon Europe program under Grant no. 101070066 (RĚGO).

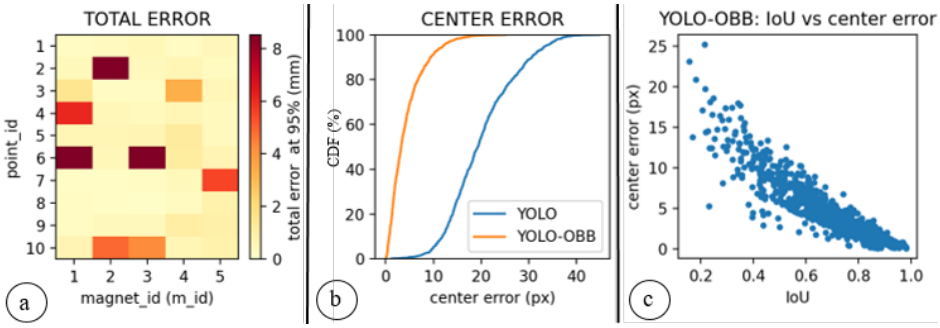


Fig. 2. a) Heatmap of the 95th percentile total 3D localization error (E_{total}), evaluated for each combination of maze position P1-P10 (y-axis) and magnet type m_id = 1-5 (x-axis). The most critical spots are concentrated at point 2 for magnet 2 and point 6 for magnets 1 and 3, where the total 3D error at p95 increases substantially relative to the remaining locations. This pattern indicates that worst-case performance is driven by a specific combination of maze geometry and magnet type and is further amplified by local visibility constraints and material-induced optical artifacts, b) Empirical cumulative distribution function (CDF) of the 2D center localization error in the image plane for both detection models, where the x-axis denotes center error of the target and the y-axis denotes the cumulative percentage of detections. Curves that rise faster and lie further left correspond to more frequent small pixel errors and thus higher center stability. This plot complements mean/p95 summaries by revealing the full distribution and the presence of long-error tails that can affect stereo triangulation, c) Scatter plot relating IoU (x-axis) to 2D center error (y-axis) for YOLO-OBB detection model. The negative trend indicates that higher overlap with the reference annotation generally coincides with smaller center errors, while lower IoU values are associated with larger variability in center estimates. This relationship helps interpret IoU as a proxy for localization quality: high IoU detections tend to provide stable centers, whereas low IoU detections include a wider range of center errors and are more likely to degrade 3D reconstruction.

References

[1]. D. Dlaka, M. Švaco, et al., Frameless stereotactic brain biopsy: A prospective study on robot-assisted brain biopsies performed on 32 patients by using the RONNA G4 system, *The International Journal of Medical Robotics and Computer Assisted Surgery*, Vol. 17, Issue 3, 2021, e2244.

[2]. A. Handa, et al., Role of robotic-assisted surgery in public health: Its advantages and challenges, *Cureus*, Vol. 16, Issue 6, 2024, e62331.

[3]. T. Haidegger, et al., Robot-assisted minimally invasive surgery-surgical robotics in the data age, *Proceedings of the IEEE*, Vol. 110, Issue 7, 2022, pp. 835-846.

[4]. Q. Cao, et al., Robotic wireless capsule endoscopy: Recent advances and upcoming technologies, *Nature Communications*, Vol. 15, 2024, 9355.

[5]. D. Zhang, et al., Advanced medical micro-robotics for early diagnosis and therapeutic interventions, *Frontiers in Robotics and AI*, Vol. 9, 2023, 1056508.

[6]. B. J. Nelson, et al., Magnetically actuated medical robots: An in vivo perspective, *Proceedings of the IEEE*, Vol. 110, Issue 7, 2022, pp. 1028-1037.

[7]. A. Schmidt, et al., Tracking and mapping in medical computer vision: A review, *Medical Image Analysis*, Vol. 94, 2024, 103134.

[8]. M. A. Ali, et al., Recent advancements in localization technologies for wireless capsule endoscopy: A technical review, *Sensors*, Vol. 25, Issue 1, 2025, 23.

[9]. S. Choudhary, et al., Three-dimensional optical microrobot orientation estimation and tracking using deep learning, *Robotica*, Vol. 43, Issue 2, 2025, pp. 616-637.

[10]. D. Zhang, et al., Micro-object pose estimation with sim-to-real transfer learning using small data, *Communications Physics*, Vol. 5, 2022, 80.

[11]. X. Liu, et al., Capsule robot pose and mechanism state detection in ultrasound using attention-based hierarchical deep learning, *Scientific Reports*, Vol. 12, 2022, 18671.

[12]. K. Botross, USMicroMagSet: Using deep learning analysis to benchmark the performance of microrobots in ultrasound images, *IEEE Robotics and Automation Letters*, Vol. 8, Issue 6, 2023, pp. 3254-3261.

[13]. A. Chandrashekhar, et al., An efficient YOLOv12-based framework for detecting extremely small-scale objects, *Scientific Reports*, Vol. 16, 2026, 2062.

[14]. H. Li, et al., Magnetic-controlled microrobot: Real-time detection and tracking through deep learning approaches, *Micromachines*, Vol. 15, Issue 2, 2024, 255.

[15]. X. Tang, Vision-based automated control of magnetic microrobots, *Micromachines*, Vol. 13, Issue 2, 2022, 269.

(018)

A Pilot Study on Facial Landmark Detection on CT and MR Head Projections for Initial Multimodal Registration

Filip Šuligoj, Marko Švaco, Bojan Šekoranja and Bojan Jerbić
Faculty of Mechanical Engineering and Naval Architecture, University of Zagreb Ivana Lučića 5,
10000 Zagreb, Croatia
Tel.: +385 1 6168 222
E-mail: filip.suligoj@fsb.unizg.hr

Summary: This pilot study investigates facial landmark detection on synthetic CT and MR head projections as an automatic coarse-initialization step for MR-to-CT and image-to-patient registration. Head surfaces reconstructed from volumetric data were rendered from multiple viewpoints, analyzed using RetinaFace, and back-projected to the 3D surface to obtain sparse anatomical correspondences. A landmark-guided view frontalization procedure was introduced to combine coarse viewpoint search from unknown initial pose with landmark-based correction toward an approximately frontal canonical view. On one CT and one MR scan, single-view detection was strongly viewpoint dependent, succeeding in 32/169 CT views and 52/169 MR views for complete heads, and in 20/169 CT views and 14/169 MR views after cropping below the nose. In contrast, the guided search converged to an acceptable first and final frontalyzed detections satisfying predefined geometric acceptance criteria in 100 % of all trials in this pilot evaluation for both modalities. Eye landmarks showed the highest repeatability across repeated trials. This preliminary feasibility demonstration supports facial landmark detection as an effective correspondence mechanism for registration initialization.

Keywords: Multimodal registration, Image-to-patient registration, CT, MR, Facial landmarks.

1. Introduction

Multimodal registration is a prerequisite for MR-to-CT fusion, image-to-patient alignment, and mixed-reality guidance in craniofacial diagnostic, therapeutic, and surgical workflows. In rigid point or surface-based registration, the objective is to estimate a transformation $g = (\boldsymbol{R}, \boldsymbol{t}) \in \mathrm{SE}(3)$ that aligns a source representation of anatomy with a target representation. In practice, however, image-to-patient and MR-to-CT registration rarely start near the correct solution. Surface based methods such as iterative closest point (ICP) and related geometric matching approaches are effective local refinements, but they require a sufficiently accurate initial pose because otherwise they can converge to an incorrect local minimum or fail to establish useful correspondences [1–3]. This initialization problem is particularly difficult when the relative head orientation is unknown and when cross-modal appearance prevents straightforward point matching.

Facial landmarks offer a compromise because they provide anatomically meaningful but sparse 3D cues that can be estimated before dense registration. In this paper, these landmarks are not used as the final registration result. Instead, they provide an automatic coarse alignment that can initialize downstream ICP, surface matching, or other established registration algorithms [1–3]. Accordingly, the presented method should be interpreted as an initialization stage rather than a complete multimodal registration pipeline. The idea is motivated by progress in optical face analysis, where deep networks have become robust to substantial appearance changes, pose variation, and partial occlusion [4, 5]. RetinaFace, introduced by Deng et al., demonstrated accurate joint face and sparse landmark localization in unconstrained optical images [6]. In medical imaging, however, the problem has mostly been addressed with modality-specific pipelines. De Boer et al. reported automatic facial landmark detection for neurosurgical mixed-reality applications in 262 CT and MR scans, achieving a mean localization error of 4.02±2.65 mm with a dedicated deep-learning model trained directly on medical data [7]. Frajtag et al. evaluated MediaPipe-based optical facial landmark localization in a surgical setting, demonstrating the sensitivity of landmark localization to camera pose and illumination changes under controlled operating-room conditions [8]. Other studies have focused on de-identification of MR data through facial-feature detection and removal [9] or on modality specific landmark estimation on 3D facial scans using deep-learning approaches [10]. These contributions confirm that medical facial landmarks are informative, but they do not directly address reuse of an optical detector on synthetic CT/MR renderings together with automatic orientation recovery from arbitrary volumetric pose.

In this pilot study, we therefore investigate exploratory but practically relevant question: can an optical face detector be reused on synthetic CT and MR-derived projections to provide an automatic initialization for multimodal registration? The proposed contribution is threefold. First, we reuse RetinaFace on synthetic projections generated from reconstructed head surfaces, thereby avoiding training a dedicated modality-specific detector. Second, we introduce guided reorientation that searches for a detectable view from arbitrary pose and then refines the orientation toward a frontal rendering before final

landmark extraction. Third, we explicitly test viewpoint sensitivity on a fixed grid of sampled poses and evaluate landmark-guided view frontalization from randomized initial poses. The output is a consistent sparse 3D landmark set that can initialize subsequent registration.

2. Materials and Methods

2.1. Volumetric Data Visualization

Volumetric data were reconstructed into a three-dimensional head surface using VTK. DICOM series were loaded with vtkDICOMImageReader. For CT, the outer head surface was extracted by isosurface rendering at -400 HU. For MR, a modality-adaptive threshold $\tau = I_{\min} + 0.09(I_{\max} - I_{\min})$ was applied, where $I_{\min}$ and $I_{\max}$ denote the non-background intensity extrema. Gaussian smoothing ($\sigma = 0.4$) was applied to MR volumes before surface extraction. The resulting meshes were smoothed with vtkSmoothPolyDataFilter (20 iterations for CT and 8 for MR), restricted to the largest connected component using vtkPolyDataConnectivityFilter, and, for MR, repaired with vtkFillHolesFilter. The surface centroid was estimated from the non-empty voxel bounding box and used to translate each model to the origin.

Synthetic projections were generated by rendering the reconstructed surface with a perspective camera (640×640, $60°$) on a viewing sphere centered at the surface centroid. A non-positional directional light (intensity 0.6, diffuse (1,1,1), specular (0.3,0.3,0.3)) was aligned with the camera up vector and aimed at the centroid so that illumination remained rigidly coupled to the viewing direction. Rendered images were captured with vtkWindowToImageFilter and transferred to OpenCV for face analysis. This rendering stage intentionally converts modality-specific volumetric information into synthetic optical-like views that remain geometrically linked to the original 3D surface.

2.2. Face Detection

Face detection and landmark localization were performed on the synthetic projections with the RetinaFace ResNet-50 model [6] through the OpenCV DNN module. The network was evaluated at 640×640 resolution with a confidence threshold of 0.5 and a non-maximum suppression threshold of 0.4. For every successful detection, RetinaFace returned five 2D landmarks: left eye, right eye, nose tip, left mouth corner, and right mouth corner. Each landmark was subsequently back-projected to the 3D surface with vtkCellPicker, producing a sparse set of 3D points that can initialize a later rigid registration stage.

It is important to emphasize that the detector operates on synthetic projections rather than directly on CT or MR slices. Consequently, detection quality remains strongly viewpoint dependent, degrades under partial anatomy, and may fail under severe self-occlusion or lower-face removal. These limitations motivated the guided reorientation procedure described below.

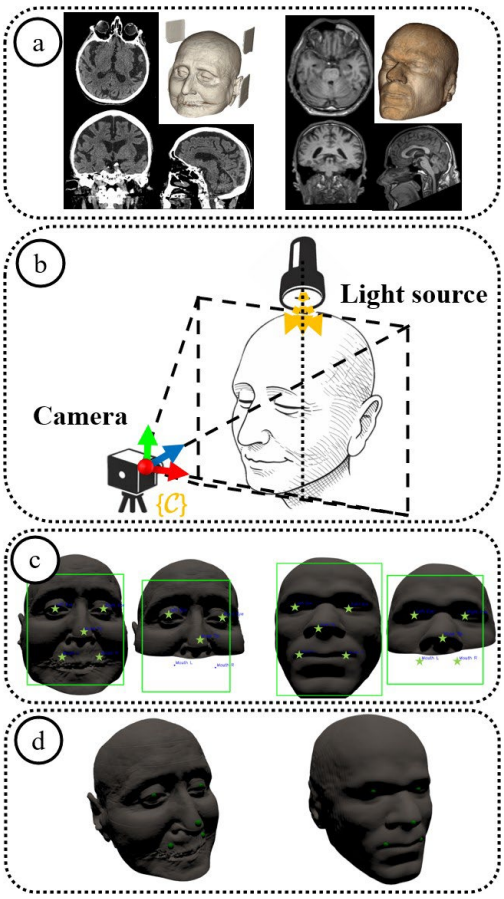


Fig. 1. Automatic initialization workflow, (a) CT and MR volumes, (b) Surface reconstruction and synthetic projection generation, (c) Face detection in rendered views, (d) Back-projection of landmarks to 3D for initialization.

2.3. Landmark-Guided View Frontalization

To obtain repeatable three-dimensional landmark locations from initially unknown surface orientations, we introduced a landmark-guided view frontalization procedure. The reconstructed CT or MR surface remained fixed in the world coordinate system, while reorientation was represented by changes in the virtual camera pose. A coarse-to-fine search was first performed over candidate view poses. For each rendered view, RetinaFace was applied to detect the face bounding box and five 2D landmarks: both eyes, nose tip, and both mouth corners. Detected 2D landmarks were back-projected onto the 3D surface using vtkCellPicker. Let $\mathbf{l}_{LE}$, $\mathbf{l}_{RE}$, $\mathbf{l}_{N}$, $\mathbf{l}_{LM}$, and $\mathbf{l}_{RM}$ denote the detected 2D locations of the left eye, right eye, nose tip, left mouth corner, and right mouth corner, respectively. The eye midpoint and mouth midpoint were computed as:

$$\boldsymbol{e} = \frac{\mathbf{l}_{LE}+\mathbf{l}_{RE}}{2}, \boldsymbol{m} = \frac{\mathbf{l}_{LM}+\mathbf{l}_{RM}}{2} \quad (1)$$

The detected landmark configuration was used to compute a bounded correction of the current view pose. The in-plane roll correction was obtained from the interocular angle:

$$\Delta\phi = -\operatorname{atan2}\left(l_{RE,y} - l_{LE,y},\ l_{RE,x} - l_{LE,x}\right) \quad (2)$$

Yaw and pitch were computed as heuristic view corrections rather than absolute anatomical pose estimates. The horizontal nose displacement from the eye midpoint was normalized by the inter-ocular distance:

$$a_x = \frac{l_{N,x} - e_x}{\|\mathbf{l}_{RE} - \mathbf{l}_{LE}\|}, \quad (3)$$

and the vertical nose position was normalized by the eye–mouth distance:

$$r_y = \frac{l_{N,y} - e_y}{m_y - e_y} \quad (4)$$

The yaw and pitch corrections were then defined as

$$\Delta\theta = -k_\theta a_x, \Delta\psi = k_\psi\left(r_y - r_y^0\right), \quad (5)$$

where $\Delta\theta$, $\Delta\psi$, and $\Delta\phi$ denote yaw, pitch, and roll corrections, respectively. In the present implementation, values of $k_\theta = 35°$, $k_\psi = 25°$, and $r_y^0 = 0.45$ were empirical selected during preliminary tuning to provide stable convergence. The corrected view was rendered again and RetinaFace was reapplied. The final detection was accepted only if the face bounding box and landmarks satisfied predefined geometric consistency criteria, including valid eye–mouth configuration, near-central nose position, a sufficient number of valid 3D landmark projections, and approximate alignment between the estimated facial plane and the image plane.

Algorithm summary:

1. Render candidate views using coarse-to-fine camera-pose sampling;
2. Apply RetinaFace and back-project detected 2D landmarks to the 3D surface;
3. Stop at the first geometrically acceptable face detection;
4. Estimate roll, yaw, and pitch corrections from eye, nose, and mouth landmark geometry;
5. Apply the correction in the local camera coordinate system;
6. Render the corrected frontal candidate and repeat landmark detection;

Accept the final view if 2D and 3D landmark consistency criteria are satisfied.

3. Results

The proposed method was evaluated on one head CT dataset from the public CQ500 dataset and one head MR dataset acquired at Dubrava University Hospital (Croatia) under approved clinical protocols with informed consent. The evaluation consisted of two complementary experiments. First, a viewpoint sensitivity analysis quantified how strongly single-view RetinaFace detection depends on head orientation and partial facial visibility on synthetic CT and MR renderings. Second, landmark-guided view frontalization was evaluated from randomized initial poses to determine whether a coarse-to-fine search strategy can recover a detectable face and refine it toward a frontal view suitable for consistent 3D landmark extraction.

Viewpoint sensitivity analysis. During the single-view experiment, the camera orientation was sampled on a fixed-radius sphere around the frontal nominal pose, as illustrated in Fig. 2. One rotational axis approximately aligned with the facial normal was fixed, while the remaining two rotational degrees of freedom were swept within $\pm 90°$at $\Delta = 15°$ increments, yielding 169 evaluated viewpoints per modality. The experiment evaluated whether isolated renderings provide sufficiently stable initialization cues without additional orientation recovery. Face detection succeeded in 32/169 CT views and 52/169 MR views for complete head models. After inferior cropping below the nose, thereby removing the mouth and lower face while preserving the upper facial region, detection succeeded in 20/169 CT views and 14/169 MR views. These results confirm strong viewpoint dependence, while also demonstrating that synthetic CT and MR renderings can still produce facial configurations compatible with an optical-image detector for sparse anatomical correspondence recovery.

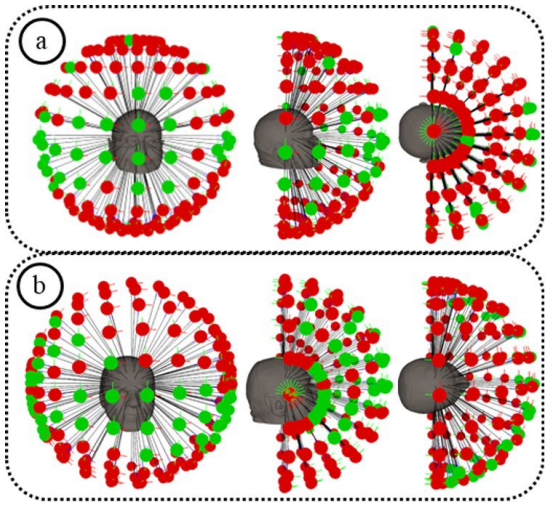


Fig. 2. Single-view detection outcomes over sampled camera poses. Green indicates successful detection and red indicates failure.

Landmark-guided view frontalization from random initial pose. Because the fixed-grid experiment demonstrated limited single-view

robustness, the proposed landmark-guided frontalization procedure was evaluated from randomized initial orientations. In contrast to the isolated single-view experiment, arbitrary initial orientations were not treated as final detections. Instead, the system iteratively rendered candidate views until a geometrically acceptable first face detection was obtained, after which the detected landmark configuration was used to estimate roll, yaw, and pitch corrections toward a frontal rendering. The procedure converged to an acceptable first face detection in all 50 independent runs for both CT and MR. After landmark-guided frontalization, final detections satisfying the predefined 2D/3D geometric consistency criteria were obtained in every case.

3D landmark repeatability and localization error. To quantify the consistency of the sparse 3D correspondences produced by the initialization stage, repeated guided-reorientation trials were analyzed using Euclidean 3D landmark localization error. For landmark i in trial k, the Euclidean localization error was computed as

$$e_{i,k} = \left\| \boldsymbol{p}_{i,k} - \bar{\boldsymbol{p}}_i \right\|_2, \bar{\boldsymbol{p}}_i = \frac{1}{N}\sum_{k=1}^{N} \boldsymbol{p}_{i,k}, \quad (6)$$

where $\boldsymbol{p}_{i,k}$ denotes the detected 3D landmark position in trial k, $\bar{\boldsymbol{p}}_i$ denotes the mean 3D landmark position over all repeated detections, and $N = 50$ for each modality. These quantities characterize the repeatability of the sparse initialization landmarks and should not be interpreted as final target registration error for intracranial anatomy.

From the Euclidean distances, mean ± SD, RMS error, and maximum error were computed for every landmark and are summarized in Table 1.

Table 1. Euclidean 3D landmark localization error over 50 trials per modality.

Mod.	Landmark	Mean ± SD (mm)	RMS (mm)	Max (mm)	Success
MR	L eye	2.67 ± 0.94	2.82	5.19	100 %
MR	R eye	2.47 ± 1.21	2.75	5.35	100 %
MR	L mouth	2.86 ± 1.68	3.31	7.66	100 %
MR	R mouth	3.92 ± 2.41	4.59	10.44	100 %
MR	Nose	3.50 ± 2.06	4.05	10.33	100 %
CT	L eye	1.66 ± 1.02	1.94	5.25	100 %
CT	R eye	1.58 ± 0.85	1.79	4.13	100 %
CT	L mouth	1.83 ± 1.57	2.4	8.03	100 %
CT	R mouth	5.03 ± 2.74	5.72	12.42	100 %
CT	Nose	2.06 ± 1.58	2.58	7.16	100 %

The repeatability analysis showed that the eye landmarks were the most stable across repeated guided reorientation trials, while the nose tip exhibited intermediate variability and the mouth landmarks showed the largest spatial dispersion. Overall, the results demonstrate that although isolated single-view detections remain strongly viewpoint dependent, the proposed landmark-guided search and frontalization procedure recovers a repeatable frontal configuration and stable sparse 3D correspondences suitable for multi-modal registration initialization.

4. Conclusion

This pilot study demonstrates that facial landmark detection on synthetic CT and MR projections can provide a feasible coarse-initialization mechanism for MR-to-CT and image-to-patient registration. Although single-view detection remained strongly viewpoint dependent, the proposed guided reorientation procedure converged to acceptable detections in all randomized CT and MR trials and produced repeatable 3D landmark estimates after frontalization. The eye landmarks were the most stable, while the mouth landmarks showed the largest variability. The main limitation of this study is the limited dataset, since the method was evaluated on only one CT and one MR case. Nevertheless, this study confirms that MR-to-CT-to-physical-space coarse registration is an effective option and that face detection can bridge preoperative and intraoperative automatic patient registration frameworks. Future work will therefore focus on validation on a larger and more diverse dataset.

Acknowledgements

This research was funded by the project INSPIRATION non-INvaSive PatIent RegistrATIon for rObotic Neurosurgery (grant no. NPOO.C3.2.R2-I1.06.0153), financed by the European Union through the National Recovery and Resilience Plan (NPOO).

References

[1]. P. J. Besl, N. D. McKay, A method for registration of 3-D shapes, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 14, Issue 2, 1992, pp. 239-256.

[2]. Y. Chen, G. Medioni, Object modelling by registration of multiple range images, *Image and Vision Computing*, Vol. 10, Issue 3, 1992, pp. 145-155.

[3]. A. Segal, D. Haehnel, S. Thrun, Generalized-ICP, in *Proceedings of the Robotics: Science and Systems (RSS)*, 2009, 21.

[4]. M. Wang, W. Deng, Deep face recognition: A survey, *Neurocomputing*, Vol. 429, 2021, pp. 215-244.

[5]. B. Johnston, P. de Chazal, A review of image-based automatic facial landmark identification techniques, *EURASIP Journal on Image and Video Processing*, Vol. 2018, 2018, 86.

[6]. J. Deng, J. Guo, E. Ververas, I. Kotsia, et al., RetinaFace: Single-shot multi-level face localisation in the wild, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 5203-5212.

[7]. M. de Boer, et al., Automatic facial landmark detection for neurosurgical mixed reality applications in MRI and CT scans using deep learning, in *Proceedings of*

the Medical Imaging with Deep Learning (MIDL), 2024, nIIXQGPZJc.

[8]. I. Frajtag, M. Švaco, F. Šuligoj, Evaluation of facial landmark localization performance in a surgical setting, in *Proceedings of the 33rd International Conference on Robotics in Alpe-Adria Danube Region (RAAD)*, 2025, pp. 278-287.

[9]. Y. U. Jeong, S. Yoo, Y.-H. Kim, W. H. Shim, De-identification of facial features in magnetic resonance images: Software development using deep learning technology, *Journal of Medical Internet Research*, Vol. 22, Issue 12, 2020, e22739.

[10]. Y. Chong, et al., Automated anatomical landmark detection on 3D facial images using U-Net-based deep learning algorithm, *Quantitative Imaging in Medicine and Surgery*, Vol. 14, Issue 3, 2024, pp. 2466-2474.

(023)

Interpretable AI Framework for Raman Spectroscopy Based Diagnostics

J. Tomeš [1], D. Janstová [1], M. Garnol [2,3] and J. Mareš [1]
[1] Department of Mathematics, Informatics and Cybernetics, University of Chemistry and Technology, Prague, Prague, Czech Republic
[2] Department of Analytical Chemistry, Faculty of Chemical Engineering, University of Chemistry and Technology, Prague, Czech Republic
[3] Department of Physical Chemistry, Faculty of Chemical Engineering, University of Chemistry and Technology, Prague, Czech Republic
E-mail: Daniela.Janstova@vscht.cz

Summary: Interpretability remains a key challenge in applying artificial intelligence to medical diagnostics. While machine learning models can achieve high performance in Raman spectral classification, their outputs are often difficult to relate to underlying biochemical processes. This work proposes an interpretable framework for Raman spectroscopy-based diagnostics that integrates data-driven models with domain knowledge. A literature-based database of Raman vibrational bands is used to associate spectral regions with molecular components such as proteins, lipids, and nucleic acids. The framework further incorporates attention maps from transformer-based models, enabling direct mapping between model decisions and biochemical interpretation. Implemented as an interactive diagnostic tool, the system allows exploration of spectral features, identification of relevant bands, and contextual interpretation of results. The proposed approach aims to enhance transparency and potentially support clinical decision-making, and represents a step toward explainable and user-oriented spectroscopic diagnostics.

Keywords: Raman spectroscopy, Explainable AI, Clinical decision support, Spectral interpretation, Attention mechanism, Biomedical diagnostics.

1. Introduction

Raman spectroscopy enables label-free molecular characterization of biological tissues by capturing vibrational signatures of proteins, lipids, nucleic acids, and extracellular matrix components [1-4]. In pathological conditions, such as cancer, biochemical alterations can lead to measurable changes in spectral features, including band intensities and band distributions [2, 5, 6].

Although machine learning has enhanced Raman spectral classification, model outputs are often not straightforward to interpret in terms of underlying biochemical mechanisms, highlighting the need for tools that support transparent analysis.

Interpretation of Raman spectra traditionally relies on expert knowledge linking spectral bands to molecular vibrations. However, this process can be complex and time-consuming. Therefore, there is a need for approaches that connect data-driven models with domain knowledge and support more transparent interpretation of spectroscopic data.

While frameworks such as SpecReX [7] or SHAP-based Raman XAI [8] utilize general explainability methods, and tools like RAMBO [9] offer automated preprocessing and band assignment, our approach provides a unique integration. Unlike these existing tools, our framework explicitly links transformer-based attention maps in real-time with a curated biochemical database. This creates a direct, user-facing diagnostic link between the model's internal decision-making and established molecular vibrations, providing context that goes beyond simple band detection or post-hoc explanation.

2. Methods

The framework integrates a transformer classifier with a curated Raman band database [12].

2.1. Data Acquisition

Spectra were acquired from healthy and neoplastic thymus tissue of adult Wistar rats (n = 3; one with spontaneously developed thymic tumour, two healthy controls), serving as a model for automated oncological screening. The pathological status was validated by specialists from Charles University during dissection. A total of 788 spectra were collected across 10 measurement days (healthy: 489; tumour: 299) using a 785 nm laser calibrated against the silicon reference band. For model training and evaluation, a per-day split strategy was employed to prevent data leakage; classification results are reported in the companion paper [12].

2.2. Data Processing and Technology

Numerical processing and signal manipulation were implemented in Python (v. 3.13.2) using NumPy [10] and SciPy [11], while spectral assignment data were managed with SQLite.

Spectra were preprocessed with StandardScaler normalization only; no baseline correction or smoothing was applied, matching the preprocessing strategy of the companion classification pipeline [3]. The classification model uses a 3-layer transformer encoder (d_model = 64, 8 attention heads) trained with binary cross-entropy loss and weighted random sampling to handle class imbalance, achieving AUROC = 0.996 ± 0.004 and specificity = 1.000 [12]. Precomputed lookup tables enable real-time mapping of spectral positions to molecular components through a curated Raman band database, linking model attention regions to established biochemical assignments.

2.3. Database and Biochemical Mapping

A curated database of Raman vibrational bands links spectral regions to molecular components such as proteins, lipids, nucleic acids, and structural motifs. The assignments were based on available literature [1-5], ensuring that each spectral position could be interpreted in the context of previously published Raman studies. Relevant bands for each spectral position are identified using inclusive range and nearest-center matching, supporting accurate assignment and interpretability.

2.4. Visualization and Interaction

Interactive visualization allows spectra and attention overlays to be explored in real time. Users can inspect which spectral regions contributed to classification decisions, providing transparency and insight into the model's behavior. The modular design supports scalability and future extensions, including multi-spectrum comparison, data export, or integration with additional predictive models.

3. Results

The framework successfully integrated with a 3-layer transformer-based classification model, which achieved accuracy = 0.991 ± 0.003, AUROC = 0.996 ± 0.004, F1 = 0.988 ± 0.000, and specificity = 1.000 [12]. The interactive tool effectively aligns high-confidence model predictions with relevant biochemical bands (e.g., proteins and lipids), providing immediate diagnostic context.

A curated reference database of Raman vibrational bands enabled real-time assignment of spectral positions to biochemical components. Users could interactively explore model-relevant regions and compare them with literature-reported assignments, supporting the interpretation of complex spectra and enhancing transparency of the classification model.

Optimized data handling and efficient querying of the database allowed smooth performance and rapid inspection of multiple spectra, without additional computational overhead. The modular system design supports integration of new spectra, attention maps, or predictive models, demonstrating flexibility and scalability for future extensions.

4. Conclusion

This work presents an interactive diagnostic framework that bridges the gap between transformer-based Raman classification and biochemical interpretability. By coupling attention-based model insights with a curated database of Raman vibrational bands, the tool enables real-time exploration of predictions, linking machine-learned spectral features to specific molecular components. The framework demonstrates strong potential for enhancing transparency in spectroscopic diagnostics, and its modular design supports future integration with additional classifiers, spectral modalities, or clinical datasets. While current validation is limited to ex vivo rat thymus tissue, the approach is readily extensible and represents a step toward clinically deployable, explainable spectroscopic tools.

Acknowledgements

This work was supported by the Ministry of Education, Youth, and Sports of the Czech Republic – grant of Specific University Research (No. A2-FCHI-2026-007).

Data Availability Statement

The Raman spectral dataset used in this study is available in the Zenodo repository (https://www.doi.org/10.5281/zenodo.18989676).

References

[1]. A. Synytsya, et al., Ex vivo vibration spectroscopic analysis of colorectal polyps for the early diagnosis of colorectal carcinoma, *Diagnostics*, Vol. 11, Issue 11, 2021, 2048.

[2]. B. Brozek-Pluska, J. Musial, R. A. Kordek, Analysis of human colon by Raman spectroscopy and imaging-elucidation of biochemical changes in carcinogenesis, *International Journal of Molecular Sciences*, Vol. 20, Issue 14, 2019, 3398.

[3]. D. Janstová, et al., Machine learning pipeline with custom grid search for colorectal Raman spectroscopy data, *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, Vol. 345, 2026, 126749.

[4]. A. C. S. Talari, Z. Movasaghi, S. Rehman, I. U. Rehman, Raman spectroscopy of biological tissues, *Applied Spectroscopy Reviews*, Vol. 50, Issue 1, 2015, pp. 46-111.

[5]. A. Synytsya, et al., Evaluation of IR and Raman spectroscopic markers of human collagens: Insights for indicating colorectal carcinogenesis, *Spectrochimica*

Acta Part A: Molecular and Biomolecular Spectroscopy, Vol. 296, 2023, 122664.

[6]. M. Karnachoriti, et al., Biochemical differentiation between cancerous and normal human colorectal tissues by micro-Raman spectroscopy, *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, Vol. 299, 2023, 122852.

[7]. N. Blake, et al., SpecReX: Explainable AI for Raman spectroscopy, *arXiv*, 2025, arXiv:2503.14567.

[8]. L. Bellantuono, et al., An explainable artificial intelligence analysis of Raman spectra for thyroid cancer diagnosis, *Scientific Reports*, Vol. 13, Issue 1, 2023, 16590.

[9]. A. Guzzetti, et al., RAMBO: An open-source web application for Raman spectral analysis and tissue characterization, *Proceedings of SPIE*, Vol. 13846, 2026, 138460D.

[10]. C. R. Harris, et al., Array programming with NumPy, *Nature*, Vol. 585, Issue 7825, 2020, pp. 357-362.

[11]. P. Virtanen, et al., SciPy 1.0: Fundamental algorithms for scientific computing in Python, Nature Methods, Vol. 17, Issue 3, 2020, pp. 261-272.

[12]. D. Janstová, et al., Transformer-based classification of Raman spectra for cancer detection, in *Proceedings of the 2nd International Conference on AI in Medicine and Healthcare (AiMH)*, 2026, pp. 33-35.

(024)

Transformer-Based Classification of Raman Spectra for Cancer Detection

D. Janstová [1], J. Tomeš [1], M. Garnol [2,3] and J. Mareš [1]
[1] Department of Mathematics, Informatics and Cybernetics,
University of Chemistry and Technology, Prague, Czech Republic
[2] Department of Analytical Chemistry, Faculty of Chemical Engineering,
University of Chemistry and Technology, Prague, Czech Republic
[3] Department of Physical Chemistry, Faculty of Chemical Engineering,
University of Chemistry and Technology, Prague, Czech Republic
E-mail: Daniela.Janstova@vscht.cz

Summary: Raman spectroscopy provides detailed molecular information for non-invasive tissue diagnostics and has strong potential in histopathological assessment and intraoperative decision-making. However, the complexity of spectral data presents challenges for conventional machine learning approaches. This study proposes a transformer-based model for classification of Raman spectra into healthy and cancerous categories. The model leverages self-attention to capture global dependencies across spectral features and operates with minimal preprocessing. Across 5 per-slice splits, the transformer achieves AUROC = 0.996 ± 0.004 and specificity = 1.000, matching logistic regression (AUROC = 1.000) while providing attention-based interpretability. Ablation confirms that positional encoding is critical (AUROC drops from 0.998 to 0.946).

Keywords: Raman spectroscopy, Transformer, Cancer detection, Spectral analysis, Deep learning, Biomedical diagnostics.

1. Introduction

Raman spectroscopy enables label-free biochemical characterization of biological tissues based on molecular vibrational signatures [1-5]. Unlike conventional histopathology, it does not require extensive sample preparation and allows direct chemical analysis rather than purely morphological evaluation [1, 2, 6]. Its compatibility with fiber-optic probes further enables integration into endoscopic systems or intraoperative workflows, supporting rapid diagnostic decision-making [7-10].

Despite these advantages, Raman spectra are high-dimensional, noisy, and contain complex biochemical information. Therefore, they require evaluation using machine learning methods [2, 7]. Simpler machine learning approaches typically need preprocessing steps, such as noise removal or baseline correction [2, 7, 9]. Deep learning methods, on the other hand, often do not require such preprocessing and can analyze spectra in a more flexible way. Certain spectral bands are interdependent because molecular vibrations appear across multiple regions of the spectrum. Transformers are well-suited for capturing these long-range dependencies, but they are still rarely applied in this domain. In this work, we aim to demonstrate that transformers can handle Raman spectral data effectively and can simplify the workflow by eliminating the need for complex preprocessing, as they are capable of automatically identifying the most relevant spectral features [4, 11].

In contrast to our prior classical ML pipeline for colorectal tissue [1, 2], this work introduces a transformer-based architecture on a distinct thymus dataset with a spontaneous tumour model.

2. Methods

2.1. Data Acquisition

Raman spectra were acquired from thymus tissue samples of adult Wistar rats (n = 3; one with spontaneously developed thymic tumour, two healthy controls). The presence of neoplastic tissue was identified and validated by specialists from the Department of Physiology (Faculty of Science, Charles University) during dissection. Tissues were snap-frozen, cryosectioned (60 µm), and measured ex vivo without fixation or staining.

A total of 788 spectra were collected across 10 measurement days (healthy: 489; tumour: 299) using a 785 nm laser calibrated against the silicon reference band. To prevent data leakage, spectra were split on a per-slice basis: all spectra from the same measurement day were assigned entirely to training (~620), validation (~80), or test (~90). Five random healthy-day assignments produced five independent splits; metrics are reported as mean ± standard deviation.

2.2. Classification Model

Spectra were classified into healthy and cancerous classes using a transformer model with 2–3 encoder layers (d_model = 64–128, 8 attention heads). Only StandardScaler normalisation was applied (no baseline correction, smoothing, or cosmic-ray removal), validating that the model handles raw Raman data effectively.

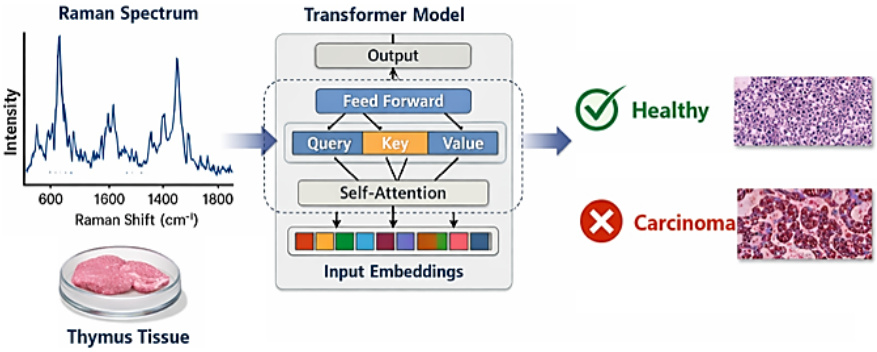


Fig. 1. Flowchart of the proposed transformer-based approach for tissue diagnostics.

3. Results

The transformer was evaluated against SVM, random forest, and logistic regression baselines across 5 per-slice splits (train/validation/test). Table 1 summarises test-set performance.

Table 1. Test-set classification performance (mean ± std over 5 independent per-slice splits).

Model	Accuracy	F1	AUROC
Transformer (L = 3, d = 64)	**0.991 ± 0.003**	**0.988 ± 0.000**	0.996 ± 0.004
Transformer (L = 2, d = 128)	0.990 ± 0.001	0.983 ± 0.011	0.995 ± 0.004
Logistic Regression	0.989 ± 0.007	0.986 ± 0.005	**1.000 ± 0.000**
Transformer (L = 2, d = 64)	0.983 ± 0.009	0.968 ± 0.033	0.992 ± 0.005
Random Forest	0.976 ± 0.011	0.969 ± 0.006	0.984 ± 0.003
1D-CNN	0.956 ± 0.031	0.947 ± 0.028	0.979 ± 0.008
SVM	0.928 ± 0.043	0.913 ± 0.033	0.982 ± 0.014

All transformer variants achieved perfect specificity (Sp = 1.000), with no false positives on healthy tissue. In an ablation study on a representative split, removing positional encoding reduced AUROC from 0.998 to 0.946. Attention maps revealed prominent peaks in the Amide III (1250–1350 cm^{-1}), lipid CH_2 deformation (1448 cm^{-1}), and Amide I (1655 cm^{-1}) regions, consistent with known protein conformational and lipid changes in tumour tissue. On the same split, raw spectra without baseline correction yielded comparable performance to baseline-corrected spectra (AUROC = 0.997 vs. 0.994), confirming the minimal-preprocessing approach.

4. Conclusion

This study demonstrates that transformer-based models are well-suited for Raman spectral analysis, offering both efficient classification and interpretative capabilities. The model's perfect specificity across all splits indicates strong potential for clinical screening applications. The current work serves as a proof-of-concept for future validation on clinical datasets.

Acknowledgements

This work was supported by the Ministry of Education, Youth, and Sports of the Czech Republic – grant of Specific University Research (No. A2-FCHI-2026-007).

References

[1]. J. Tomeš, D. Janstová, S. Mohsen, A. Sinica, et al., Rapid colorectal tissue classification using data-driven Raman techniques, *IEEE Access*, Vol. 13, 2025, pp. 29601-29612.

[2]. D. Janstová, J. Tomeš, J. Vališ, A. Synytsya, et al., Machine learning pipeline with custom grid search for colorectal Raman spectroscopy data, *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, Vol. 345, 2026, 126749.

[3]. M. N. Jensen, E. M. Guerreiro, A. Enciso-Martinez, S. G. Kruglik, et al., Identification of extracellular vesicles from their Raman spectra via self-supervised learning, *Scientific Reports*, Vol. 14, Issue 1, 2024, 6791.

[4]. M. Chang, C. He, Y. Du, Y. Qiu, et al., RaT: Raman transformer for highly accurate melanoma detection with critical features visualization, *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, Vol. 305, 2024, 123475.

[5]. A. Synytsya, D. Janstová, M. Šmidová, A. Synytsya, et al., Evaluation of IR and Raman spectroscopic markers of human collagens: Insides for indicating colorectal carcinogenesis, *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, Vol. 296, 2023, 122664.

[6]. A. Synytsya, A. Vaňková, M. Miškovičová, J. Petrtýl, et al., Ex vivo vibration spectroscopic analysis of colorectal polyps for the early diagnosis of colorectal carcinoma, *Diagnostics*, Vol. 11, Issue 11, 2021, 2048.

[7]. J. Vališ, M. Fousková, D. Janstová, L. Habartová, et al., Automated classification pipeline for real-time in vivo examination of colorectal tissue using Raman spectroscopy, *Spectrochimica Acta Part A: Molecular*

and Biomolecular Spectroscopy, Vol. 313, 2024, 124152.

[8]. H. Ding, A. W. Dupont, S. Singhal, L. D. Scott, et al., Effect of physiological factors on the biochemical properties of colon tissue – an in vivo Raman spectroscopy study, *Journal of Raman Spectroscopy*, Vol. 48, Issue 7, 2017, pp. 902-909.

[9]. M. Fousková, J. Vališ, A. Synytsya, L. Habartová, et al., In vivo Raman spectroscopy in the diagnostics of colon cancer, *Analyst*, Vol. 148, Issue 11, 2023, pp. 2518-2526.

[10]. M. S. Bergholt, et al., Simultaneous fingerprint and high-wavenumber fiber-optic Raman spectroscopy enhances real-time in vivo diagnosis of adenomatous polyps during colonoscopy, *Journal of Biophotonics*, Vol. 9, Issue 4, 2016, pp. 333-342.

[11]. S. Weng, C. Wang, R. Zhu, Y. Wu, et al., Identification of surface-enhanced Raman spectroscopy using hybrid transformer network, Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, Vol. 316, 2024, 124295.

(025)

A Machine Learning Approach to identify nutrition-related Diseases based on Fingernail Structure

Jan-Torsten Milde, N. Sonntag, L. Müller, K. Plandl, D. Dewald, R. Blum, H. Hollenbach, A. Maxones, T. Kühn and M. Birringer
Fulda University of Applied Sciences, 36037 Fulda, Germany
E-mail: milde@hs-fulda.de

Summary: Nutritional deficiencies, particularly regarding micronutrients like iodine, iron, and calcium, remain a significant public health challenge despite high food availability. Current diagnostic methods often rely on invasive and expensive blood tests that are not part of routine screenings. This paper presents the KISMA project, which develops a non-invasive, AI-supported Point-of-Care (PoC) system. By utilizing high-resolution macro-imaging of fingernails and multimodal Convolutional Neural Networks (CNNs), the system aims to identify risks for micronutrient deficiencies and nutrition-related diseases. Preliminary studies indicate that mineral profiles in nails correlate significantly with lifestyle factors and specific health conditions.

Keywords: Predictive image-based analysis, Nutritional deficiencies, Fingernail structure, Non-invasive diagnostics, Micronutrients, Mineral analysis.

1. Introduction

The integration of Artificial Intelligence (AI) into medical diagnostics has demonstrated significant success, particularly in dermatology where machine learning algorithms assist in the early detection of skin cancer.

By training on large datasets of skin lesions - independently validated by experts - these systems can distinguish between benign moles and malignant lesions with high validity.

The KISMA project adopts a similar methodology but shifts the focus to *nutritional science and preventive healthcare*. It is well-established in clinical literature that various parameters, such as age, dietary patterns, and chronic diseases, influence the *mineral composition and morphology of fingernails*. KISMA seeks to leverage these correlations by developing a non-invasive, AI-driven tool for health assessment.

2. Project Objective

The primary goal of KISMA is to develop an AI system capable of *inferring a user's mineral supply and identifying potential health conditions* based solely on digital images of fingernail patterns. By correlating visual data with chemical ground truth and health histories, the project aims to bypass the need for expensive and time-consuming laboratory analyses of nail clippings. This approach to *causality analysis* is currently underrepresented in scientific literature and is intended to culminate in a proprietary diagnostic application.

2.1. Fingernail Images as a Non-Invasive Diagnostic Window

Current research results document a link between fingernails and general health status, though the topic of nutrition is highlighted somewhat indirectly within the context of deficiency symptoms and diet-related diseases [1]. It is evident that the fingernail represents a valuable tool for the non-invasive monitoring of general health [2, 3].

Morphological changes in the nail - particularly regarding color, texture, thickening, and distortion - are not merely superficial cosmetic phenomena. Rather, these visual patterns can function as vital markers indicating underlying health issues. Specifically, such changes are associated with nutritional deficiencies. Furthermore, nail morphology serves as an indicator of systemic disorders, as it can also characterize more serious conditions such as anemia and diabetes. The analysis of these features thus offers significant potential for early and patient-friendly diagnostics.

2.2. CNNs as a Method for Medical Image Analysis

CNNs (Convolutional Neural Networks) possess powerful feature learning and representation capabilities, demonstrating strong performance in image classification, segmentation, and object detection [4]. The advancement of deep learning reduces the need for the manual development and optimization of new feature extractors, thereby enabling automatic feature extraction and self-learning [5].

3. Interdisciplinary Methodology

The project is structured into two collaborative sub-projects involving artificial intelligence and nutrional sciences and healthcare.

3.1. Artificial Intelligence and Image Processing

The AI sub-project focuses on the development of specialized tools and procedures for the *high-quality capture of fingernail images*. These images serve as the foundation for the training and testing phases of the recognition system. The methodology utilizes advanced machine learning techniques specifically optimized for *image processing and pattern recognition*, building upon previous research in autonomous systems and gesture recognition.

3.2. Nutritional Science and Analytical Ground Truth

The nutritional sub-project provides the essential metadata required for the AI to "learn" the biological context of the visual patterns. This includes:

- *Anonymized Questionnaires*: An online survey collects data on demographics (age, gender, BMI), dietary patterns (e.g., vegan, vegetarian, omnivore), chronic illnesses (e.g., diabetes, autoimmune diseases), and medication use.
- *Quantitative Mineral Analysis*: To provide a verifiable scientific basis, nail samples are analyzed using Inductively Coupled Plasma Mass Spectrometry (ICP-MS) at the Core Facility's instrumental analysis laboratory. This state-of-the-art method allows for the quantitative measurement of up to 40 elements, including essential minerals, trace elements, and toxic heavy metals.

3.3. Preliminary Work and Pilot Study

The foundations of KISMA were established through a pilot study (NutriNails) conducted in 2024. While results of the nutrional sciences and healthcare part of this pilot study have been published in [5], the results of the technical part have not been presented to the scientific community.

The study was registered with the German Register of Clinical Trials (DRKS).

- *Participants*: 200 subjects provided data via a 30-item online survey.
- *Data Collection*: Fingernail surfaces were photographed and physically collected for chemical analysis.
- *Analysis*: Over 20 bulk and trace elements were quantified via ICP-MS.
- *Integration*: All data packets (photos, survey results, and chemical analysis) were linked via a barcode system to ensure complete anonymity while maintaining data integrity for initial descriptive statistics.

In the technical part of this pilot study a dataset of 5,000 fingernail images was collected, primarily sourced from a population of healthy young students (aged 20–30) of diverse ethnicities. In parallel, a mobile application for the independent recording of fingernail images was developed iteratively. To ensure that only high-quality, usable images were included in the dataset, an integrated AI-based quality control system was implemented using MediaPipe for hand detection and YOLO for fingernail image classification.

The central result of this preliminary work is the successful development of a CNN-based autoencoder that enables the differentiation between standard and non-standard fingernail images based on medium-resolution captures. This autoencoder allows for the efficient preprocessing of large volumes of image data and the rapid identification of interesting special cases.

However, the pilot study also revealed a limitation of this approach: medium-resolution images are insufficient for fine-grained analysis and correlation with biochemical analytical markers. To achieve the necessary level of detail, macro photography under precisely defined and reproducible conditions is required. Currently, no commercially available device exists for the standardized capture of such macro images, necessitating the development of a dedicated hardware solution for the project proposed here.

4. Current Work

We are currently working on the development of a multimodal AI system. The Convolutional Neural Networks (CNNs) tested in the pilot study have already proven their functionality with a small number of images. To date, however, these AI systems have exclusively processed image information. To successfully correlate visual features with biochemical measurement data (from the analysis of fingernail clippings), a multimodal expansion of the network is mandatory. The prerequisite for this is the methodologically sound construction of a high-quality data corpus that integrates both the high-resolution macro images from the new device and the corresponding clinical metrics, thereby enabling the modeling of the complex relationships between nail patterns and nutritional deficiencies.

In order to create high-quality image data, we are building an imaging device for creating high-resolution macro images of fingernails, which serve as the necessary foundation for the precise assessment of nutritional parameters. Since the device is intended for use outside the laboratory—directly on-site with users or customers—a special focus is placed on high usability and user experience (UX) to ensure ease of

use by non-scientists. At the same time, high hygiene standards must be met. The device's connectivity is essential to allow for the direct transfer of high-resolution images to the AI-based analysis system for the generation of a personalized nutritional assessment.

4. Conclusions

The KISMA project represents an advancement in AI-supported diagnostics. By combining nutritional science with computer vision, it offers a scalable solution for early detection of malnutrition, potentially reducing long-term healthcare costs and improving quality of life for an aging population.

References

[1]. J. R. Navarro-Cabrera, M. A. Valles-Coral, M. E. Farro-Roque, N. Reátegui-Lozano, et al., Machine vision model using nail images for non-invasive detection of iron deficiency anemia in university students, *Frontiers in Big Data*, Vol. 8, 2025, Article 1557600.

[2]. A. Munjal, S. Sharma, Onychomycosis: A portable nail infection detection system using advanced image processing, in *Proceedings of the 2025 9th International Conference on Inventive Systems and Control (ICISC)*, IEEE, 2025, pp. 1–7.

[3]. P. M. Gote, P. Kumar, T. Dhale, G. V. Mishra, Development of a nail-based application for the detection of disease using image processing, in *Proceedings of the 2024 2nd DMIHER International Conference on Artificial Intelligence in Healthcare, Education and Industry (IDI-CAIEI)*, IEEE, 2024, pp. 1–6.

[4]. Z. Can, Ş. Işık, Diagnosing diseases from fingernail images, *Eskişehir Osmangazi Üniversitesi Mühendislik ve Mimarlık Fakültesi Dergisi*, Vol. 30, Issue 3, 2022, pp. 464–470.

[5]. P. Kaushik, P. Sharma, Optimized nail disease detection using DenseNet121: Advancing dermatological diagnostics with deep learning, in *Proceedings of the 2024 International Conference on Decision Aid Sciences and Applications (DASA)*, IEEE, 2024, pp. 1–6.

[6]. N. Sonntag, L. Müller, K. Plendl, D. Dewald, et al., Fingernail mineral profiling as a non-invasive tool to assess dietary and lifestyle factors: Results from the cross-sectional Fulda NutriNAIL study, *BioFactors*, Vol. 51, Issue 6, 2025, Article e70056.

(026)

Trust by Design: Building Reliable Clinical AI to Advance Quality in Patient-Centered Care

Maria L. Reyna-Cruz [1], Martine Ceberio [1], Christoph Lauter [1], Cecilia A. Marquez Barraza [2] and Jesus M. López Valles [3]

[1] University of Texas at El Paso, CR2G Laboratory,
500 W. University Avenue., 79968 El Paso Tx., United States of America
[2] Centro Medico de Especialidades, Av. de Las Americas 201, Margaritas,
32300 Ciudad Juárez, Chihuahua, México
[3] Instituto Mexicano del Seguro Social, 8560 Av. Vicente Guerrero, Las Quintas,
31240 Ciudad Juarez, Chihuahua, México
Tel.: + 1 9152669878
E-mail: mlreynacruz@miners.utep.com

Summary: Artificial intelligence (AI) is increasingly integrated into clinical workflows, yet real-world adoption remains limited by concerns around trust, transparency, and usability. We present a practical framework for developing trustworthy clinical AI, based on two ongoing research projects: (1) automated diagnosis of develop- mental dysplasia of the hip (DDH) using ultrasound and uncertainty-aware machine learning, and (2) personalized treatment prediction for relapsed non-Hodgkin lymphoma (NHL) using interpretable deep learning models. Although these domains differ in modality, risk, and workflow, they share common principles: domain expert-informed design, explainability, demographic bias assessment, clinical validation, and uncertainty quantification. Drawing from these cases, we propose five pillars for trustworthy AI in healthcare: rigorous validation, transparency, bias mitigation, clinician engagement, and ethical oversight. Our approach emphasizes that AI success in medicine depends not only on predictive performance but on how well it aligns with the needs of those it aims to serve: clinicians, institutions, and patients.

Keywords: Clinical AI, Trustworthy AI, Explainable AI, Uncertainty quantification, Bias mitigation, Medical decision-making.

1. Introduction

Artificial intelligence (AI) systems are becoming increasingly embedded in clinical decision-making, with the potential to improve diagnostic accuracy, support treatment planning, and enable more individualized care. Yet, much of this progress has focused on optimizing performance metrics, such as accuracy or area under the curve, rather than on the lived experience of the patients and clinicians who interact with AI in practice. As a result, we are now recognizing that trust is not missing due to lack of intent, but because AI development has been prioritizing task completion over human alignment. As we move from proof-of-concept models to real-world adoption, the healthcare community must re-center AI development around reliability, transparency, and user trust. Clinical adoption requires systems that are not only technically accurate but also transparent, reliable, and aligned with medical reasoning [1, 2].

In recent years, increasing attention has been directed toward explainable AI, fairness in machine learning, and uncertainty quantification in medical models. However, there remains a need for practical approaches that integrate these concepts throughout the development lifecycle of clinical AI systems [3, 4].

This paper explores how trustworthy AI principles can be operationalized through real research projects. This paper presents a practical approach to building trustworthy clinical AI, drawing from two of our ongoing research projects: (1) a machine learning system for classifying Developmental Dysplasia of the Hip (DDH) using ultrasound and Graf's method, and (2) a deep learning-based predictive model for treatment in relapsed Non-Hodgkin Lymphoma (NHL). While these projects differ in imaging modality, clinical workflow, and data types, they intersect in their shared goal: making AI clinically usable by integrating domain expertise, explainability tools, and quality-focused validation techniques, including quantification of uncertainty.

This work makes three contributions. First, we describe two applied clinical AI projects with different data types addressing distinct medical tasks: diagnostic imaging and treatment prediction. Second, we analyze methodological challenges encountered when attempting to translate machine learning systems into clinically usable tools. Third, drawing from these experiences, we propose a framework consisting of five design pillars that support the development of trustworthy clinical AI systems. The projects presented in this work serve as case studies illustrating how trustworthy AI principles can be implemented in real-world clinical research settings.

This paper is not intended as a comprehensive technical evaluation of the underlying machine learning models. Instead, the focus is on identifying and synthesizing design principles that emerged during the development of clinically-oriented AI systems across two distinct healthcare domains. Detailed

technical validation and quantitative model evaluation are the subject of ongoing and future work.

2. Background: Trustworthy AI in Healthcare

AI-driven decision support systems are progressively explored across medical domains, including medical imaging, oncology, and critical care. In diagnostic imaging, deep learning models have demonstrated performance comparable to human experts in several tasks [5]. Similarly, predictive models using electronic health record data have been proposed to assist in prognosis estimation, treatment planning, and hospital resource management [6].

Despite these advances, clinical implementation remains challenging. Concerns regarding model interpretability, algorithmic bias, and generalization across patient populations have limited widespread adoption. Clinicians must understand how predictions are generated and when model outputs should be trusted or questioned. Without transparency and clear communication of model limitations, even highly accurate systems may fail to gain acceptance.

Explainable AI (XAI) methods have emerged as one approach to address these challenges. Techniques such as SHAP and LIME provide mechanisms for interpreting model predictions by identifying influential features. At the same time, uncertainty-aware learning approaches aim to quantify the confidence associated with predictions, allowing clinicians to recognize ambiguous cases that may require additional evaluation.

Together, these developments highlight the importance of designing AI systems that prioritize not only performance but also interpretability, fairness, and alignment with clinical practice.

3. Project 1: Automating DDH Diagnosis

Developmental dysplasia of the hip (DDH) is a pediatric condition characterized by abnormal development of the hip joint. Early diagnosis is critical, as timely intervention can prevent long-term complications such as joint instability, gait abnormalities, and early-onset osteoarthritis. Ultrasound imaging is commonly used for screening infants, particularly during the first months of life.

Graf's method remains the most widely used ultrasound-based classification approach for assessing hip maturity [7]. The method relies on precise measurement of anatomical angles from ultrasound images. While standardized, the procedure is highly operator-dependent and requires significant training. Variability in image acquisition and interpretation can lead to inconsistent diagnoses, particularly in settings with limited specialist availability.

Our DDH research focuses on automating Graf's classification of hip maturity from ultrasound images using convolutional neural networks while having a pediatric orthopedist in the loop since the design of the study (Dr. Marquez). Graf's method, while standardized, is highly operator-dependent, which can lead to variability in diagnosis and missed early interventions. To address this, our approach incorporates a curated dataset labeled by pediatric orthopedists following standardized clinical evaluation procedures and aims to integrate angle measurement predictions that correlate with clinical judgment, while using uncertainty quantification to communicate model confidence in ambiguous or borderline cases [8]. This supports trust in the system and fosters potential clinical adoption, especially in resource-limited environments where access to trained personnel is constrained.

This approach not only improves consistency in measurements but also provides a structured way to communicate model uncertainty in clinically meaningful terms.

4. Project 2: Treatment n Relapsed NHL

Non-Hodgkin lymphoma (NHL) represents a diverse group of hematologic malignancies with varying clinical courses and treatment responses. While many patients initially respond to therapy, a subset experience relapse and require additional treatment decisions. Selecting an appropriate therapy in the relapsed setting is complex, as clinicians must consider prior treatments, patient health status, laboratory results, and emerging therapeutic options.

Our second research project investigates the use of machine learning to support treatment decision-making in relapsed NHL. The system utilizes structured clinical data, which includes treatment history, laboratory values, and long-term outcomes, to predict personalized treatment strategies [9]. The primary objective of the system is treatment-response prediction and treatment decision support in relapsed NHL rather than survival analysis alone.

Because treatment decisions directly affect patient survival and quality of life, interpretability is particularly important in this context, as model outputs must be understandable not only to oncologists but also to patients facing life-altering treatment choices. To address this need, the system aims to integrate SHAP value visualizations and LIME-based explanations. These methods provide both global insights into model behavior and local explanations for individual predictions.

Collaboration with a hematologist has been central to the design of this system. Clinical expertise is used to ensure that model outputs align with established treatment guidelines and medical reasoning. Model evaluation therefore considers not only predictive performance but also agreement with historical expert treatment decisions and patient survival outcomes over 5–10 years.

5. Intersecting Principles and Distinctions

While the DDH and NHL projects differ in domain and complexity (addressing different clinical tasks and

using different data modalities), they reveal several shared principles relevant to trustworthy AI development. Shared guiding principles include:

• Domain-expert-informed annotation and design. First, both projects emphasize close collaboration with medical experts. Domain knowledge is essential for defining appropriate prediction targets, interpreting model outputs, and validating results. Without clinical guidance, even technically sophisticated models risk producing outputs that are difficult to integrate into real workflows.

• Explainability at both the global and local level. Second, interpretability plays a central role in both applications. In the DDH project, interpretability is achieved by predicting anatomical measurements that correspond directly to clinical practice. In contrast, the NHL system relies on feature-attribution methods to explain predictions derived from structured patient data. These differences illustrate how explainability techniques must be adapted to the characteristics of the data and the clinical context.

• Uncertainty quantification to support risk-aware interpretation. Uncertainty awareness emerges as a key component of trustworthy AI. In imaging-based diagnosis, uncertainty can highlight ambiguous cases that require expert review. In treatment prediction, uncertainty may help clinicians assess the reliability of model recommendations before integrating them into clinical decisions.

• Bias assessment across demographic groups. Both projects recognize the importance of evaluating model performance across demographic and clinical subgroups to identify potential disparities in predictions and ensure fairer deployment in heterogeneous patient populations.

• Validation based on real-world clinical decision-making. Model evaluation considers not only predictive performance but also consistency with clinician reasoning, treatment guidelines.

These principles highlight the versatility and challenge of implementing trustworthy AI across different care contexts with contrasting demands. DDH emphasizes early detection in pediatric imaging, with goals of reducing missed cases and increasing diagnostic consistency. NHL, on the other hand, deals with longitudinal decision-making and individualized treatment planning in a complex, high-risk setting. These contrasts require different explainability approaches (visual vs. tabular), different model evaluation criteria (precision vs. survival alignment), and different user interactions (pediatric orthopedist versus oncologist-patient discussions). These observations suggest that trust in clinical AI is not domain-specific, but emerges from consistent design principles that adapt to different clinical contexts.

6. Framework for Trustworthy Clinical AI

The five pillars proposed in this framework emerged through comparative analysis of recurring challenges encountered across the two clinical AI projects, including model interpretability, clinician interaction, uncertainty communication, data limitations, and workflow integration. Rather than representing isolated technical requirements, the pillars reflect practical considerations repeatedly identified during the development of clinically-oriented AI systems. Building on these experiences, the proposed five essential pillars for the development of trustworthy clinical AI systems:

Rigorous Validation. Clinical AI systems must be evaluated across diverse patient populations and realistic clinical scenarios. Validation should extend beyond retrospective performance metrics to include testing on independent datasets and, when possible, prospective clinical evaluation. Such practices help ensure that models remain reliable when deployed outside their original development environment.

Explainability and Transparency. Clinicians must be able to understand how model predictions are generated. Explainable AI techniques provide mechanisms for visualizing feature importance, highlighting influential variables, and identifying potential errors in model reasoning. Transparent systems encourage meaningful collaboration between human experts and AI tools [1, 2, 10].

Bias Detection and Mitigation. Machine learning models trained on historical healthcare data may inherit existing biases present in clinical datasets. Systematic evaluation across demographic groups is necessary to identify disparities in model performance. Incorporating fairness assessments and uncertainty-aware approaches can help mitigate these risks [3, 4].

Clinician Engagement. Successful clinical AI systems are rarely developed in isolation. Continuous engagement with clinicians throughout the development lifecycle, from model design to evaluation, ensures that models address real clinical needs and integrate effectively into decision-making workflows.

Ethical Governance. The development and deployment of clinical AI must adhere to ethical and regulatory standards. Institutional review boards, data governance policies, and transparency in data usage are essential components of responsible AI research in healthcare. [11] [12]

These principles are not discipline-specific but are adaptable to various clinical AI use cases, from early diagnostics to treatment planning. By grounding AI development in this framework, we aim to align model development with quality improvement goals such as safety, equity, operational efficiency, clinician experience and high-quality patient care.

7. Discussion

While the projects described in this work illustrate promising directions for trustworthy clinical AI, several challenges remain. Access to high-quality clinical datasets can be limited due to privacy

regulations and institutional constraints. In addition, models trained on retrospective data may not fully capture the complexity of real-time clinical decision-making. The current status of these projects includes dataset curation and ongoing development of uncertainty-aware approaches for diagnostic and treatment-support tasks.

Future work will involve further validation of these systems with ongoing collaboration with clinical partners and exploration of uncertainty-aware decision support tools. Integrating AI predictions with clinician feedback mechanisms may also help refine model behavior over time. Ultimately, successful deployment will require not only technical improvements but also careful consideration of workflow integration and clinician training.

8. Conclusion

Artificial intelligence has the potential to transform healthcare by supporting earlier diagnosis, improving treatment planning, and enabling more personalized care. However, the success of these technologies depends not only on predictive performance but also on their reliability, transparency, and alignment with clinical practice.

AI in healthcare must go beyond optimizing for technical performance. By integrating explainability, uncertainty quantification, clinicians' insight, and patient-centered design, we can build systems that are not only accurate, but also understandable to clinicians and patients, and safe for institutional adoption. Our work on DDH and NHL illustrates that while models may differ in form and function, the foundation of clinical AI must be designed with trust as a core outcome.

Designing AI systems with trust as a foundational objective can help ensure that these technologies serve clinicians and patients effectively, ultimately advancing the quality of patient-centered care.

While these findings are based on ongoing research projects and retrospective data, further prospective validation will be essential for real-world clinical deployment. The success of AI in medicine depends not only on what it can predict, but on who it serves and how it is experienced.

References

[1]. A. Bhattacharya, Applied Machine Learning Explainability Techniques: Make ML Models Explainable and Trustworthy for Practical Applications Using LIME, SHAP, and More, *Packt Publishing*, Birmingham, 2022.

[2]. F. Dallanoce, Explainable AI: A comprehensive review of the main methods, Medium, 2022, https://medium.com/mlearning-ai/explainable-ai-a-comprehensive-review-of-the-main-methods-4cb1e2b6b2ba

[3]. J. Gawlikowski, C. R. N. Tassi, M. Ali, J. Lee, et al., A survey of uncertainty in deep neural networks, *Artificial Intelligence Review*, Vol. 56, Issue Supplement 1, 2023, pp. 1513-1589.

[4]. P. Wang, N. C. Bouaynaya, L. Mihaylova, J. Wang, et al., Bayesian neural networks uncertainty quantification with cubature rules, in *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*, 2020, pp. 1-7.

[5]. A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, et al., Dermatologist-level classification of skin cancer with deep neural networks, *Nature*, Vol. 542, 2017, pp. 115-118.

[6]. Z. Obermeyer, E. J. Emanuel, Predicting the future – big data, machine learning, and clinical medicine, *The New England Journal of Medicine*, Vol. 375, Issue 13, 2016, pp. 1216-1219.

[7]. R. Graf, Hip Sonography: Diagnosis and Management of Infant Hip Dysplasia, *Springer*, Berlin, 2006.

[8]. M. L. Reyna-Cruz, R. Tabares, M. Ceberio, V. Kreinovich, et al., Machine learning-based screening for pediatric hip dysplasia: Towards a validated approach, in *Proceedings of the 58th Asilomar Conference on Signals, Systems, and Computers*, 2024, pp. 1891-1894.

[9]. M. L. Reyna-Cruz, M. Ceberio, C. Lauter, J. M. L. Valles, Deep learning explainability on non-Hodgkin lymphoma: Relapse & treatment, in *Proceedings of the 1st International Conference on AI in Medicine and Healthcare (AiMH)*, 2025, pp. 75-79.

[10]. R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, et al., A survey of methods for explaining black box models, *ACM Computing Surveys*, Vol. 51, Issue 5, 2018, 93.

[11]. C. J. Guerrini, A. L. McGuire, An ethics framework for evaluating ownership practices in biomedical citizen science, *Citizen Science: Theory and Practice*, Vol. 7, Issue 1, 2022, 23.

[12]. Ethics and governance of artificial intelligence for health: WHO guidance, *World Health Organization*, Geneva, 2021.

(030)

Artificial Intelligence in Nursing Care: Opportunities, Challenges, and Future Directions

K. Wolf-Ostermann
University of Bremen, Institute for Public Health and Nursing Research (IPP), PF 330440, D-28334 Bremen, Germany
Health Sciences Bremen, University of Bremen, PF 330440, D-28334 Bremen, Germany
E-mail: wolf-ostermann@uni-bremen.de

Summary: Artificial intelligence (AI) has the potential to transform nursing care by improving quality, reducing workloads and enabling personalised, preventive care. However, its successful integration requires overcoming significant technical, organisational, ethical and social barriers, including poor data quality, limited digital infrastructure, insufficient workforce competencies and low nurse involvement in development. Although applications such as fall detection, automated documentation and decision support are emerging, most research remains hospital-centric and lacks real-world validation. To address these issues, the AI Nursing Care Readiness Assessment (AINCRA) has been developed. Consisting of five dimensions – regulatory, processual, technical, social and ethical, and community building – across five maturity levels, AINCRA provides 69 actionable attributes to guide interdisciplinary teams in planning, evaluating, and advancing AI projects throughout their lifecycle. Sustainable AI integration demands participatory design, workforce training, and evidence-based development. Prioritising readiness, equity, and user-centred approaches is essential to realise AI's full potential. The future of AI in nursing is not about replacing nurses, but empowering them. This requires a strategic, human-centred approach that uses technology to enhance, rather than replace, the core values of nursing.

Keywords: Artificial intelligence (AI), Digital health, Healthcare technology, Implementation, Nursing care.

1. Introduction

Digitalization has become an integral part of modern healthcare systems. The World Health Organization [1] has issued a call to policymakers, urging them to prioritise the development and implementation of digital tools and technologies as a fundamental strategy to address the mounting health system and population demands. The integration of artificial intelligence (AI) into nursing practice and nursing education necessitates substantial investment in infrastructure, organisational processes, and workforce capacity building, and gives rise to critical inquiries concerning equity and inclusion. It is imperative that nursing professionals possess the necessary preparation and support to competently utilise digital technologies across all aspects of their work. This encompasses training, clinical practice, research and management.

A recent report on nurses' views regarding the role of AI tools in their future work [2] indicates an increasing utilisation of AI tools by nurses for professional development, patient education, and medical research purposes. A significant proportion of nurses also perceive considerable future potential for AI tools in enhancing patient outcomes and consultation quality. While nurses as a group tend to view AI positively and believe it can support patient care and professional autonomy, they are more cautious than physicians regarding time savings and routine clinical use. The acceptance of AI by nurses is contingent upon three factors: transparency, usability, and safety. Nurses value clear and comprehensible outputs, reliable evidence-based information, and transparent sources with citations.

Whilst there is considerable potential for artificial intelligence (AI) to transform and improve processes, its successful implementation requires an interdisciplinary approach that balances technological, ethical and practice-based perspectives. The purpose of this balanced approach is to foster the development of effective, user-centred AI solutions. In order to achieve the full potential of artificial intelligence (AI) in the field of nursing care, particularly within long-term care settings, it is essential that AI systems are trustworthy, practical, and designed to support – rather than replace – human caregivers.

2. Applications of AI in Nursing Care

The objective of AI is to replicate human cognitive functions such as reasoning, decision-making, and pattern recognition. In the domain of nursing, the integration of AI technologies has the potential to facilitate care delivery, enhance patient autonomy, and partially substitute nursing activities [3].

A recent review [7] targeting diverse healthcare settings demonstrates the multifaceted effects of AI integration in nursing, concluding that AI in nursing significantly improves clinical outcomes by increasing diagnostic accuracy, enabling earlier detection of health changes, and supporting faster interventions that reduce complications and hospital stays. Furthermore, it has been demonstrated that such systems can enhance operational efficiency and staff well-being by

automating routine tasks and reducing burnout. The overarching objective of the initiative is twofold: firstly, to enhance the quality of care, and secondly, to concomitantly reduce the workload of nursing personnel.

2.1. Types of AI Systems

Despite increasing interest, AI in nursing lacks a standardized classification and is commonly structured around methods. AI systems in nursing can be categorized into:

- machine learning systems for predictive analytics;
- rule-based expert systems for decision support;
- hybrid systems combining multiple approaches [4, 5].

2.2. Examples in Practice

The employment of AI in nursing is increasingly becoming a significant aspect of professional practice, with AI already being utilised in numerous fundamental domains of nursing care. However, the majority of research is concentrated on hospital settings, while applications in long-term care and outpatient settings remain restricted [6].

Examples of AI applications include monitoring patient health and detecting falls, improving communication, predicting care needs, and supporting administrative processes such as scheduling [6].

AI systems have been demonstrated to be capable of providing care professionals with support through the automation of routine tasks, the optimisation of administrative workflows, and the provision of decision support in patient care [7-9]. Furthermore, AI-driven tools have been shown to promote self-management, independence and quality of life for individuals with chronic conditions, while simultaneously alleviating the burden on family caregivers [6, 7, 10]. Existing practical applications include AI-powered fall detection systems, automated wound monitoring, voice-based nursing documentation tools, and social or emotional robots used in care and support settings [12-15].

The following examples of applications of AI in nursing care exemplarily illustrate evidence based success:

- An AI-powered early warning system (EWS) for sepsis detection in hospitals [16] which demonstrated that AI-enabled systems significantly improved early detection of clinical deterioration, including sepsis, leading to better patient outcomes. Nurses reported increased confidence in identifying at-risk patients and initiating timely interventions.
- AI-enhanced wearable sensors for real-time fever detection [17] confirmed that AI-enhanced wearables significantly improve early detection of postoperative complications and enable nurses to intervene proactively, reducing postoperative infections and complications.
- An AI-driven speech assistant for nursing documentation [18] confirmed that AI speech assistants are well-received and effective in reducing administrative burden.
- An AI-driven predictive workload classification for nurse staffing [19] validated AI as a reliable tool for workload prediction, supporting efficient resource allocation in clinical settings.
- An AI-powered mobile app for nurse burnout reduction [20] confirmed that AI-driven mental health tools can effectively support nurse well-being and reduce burnout.

These examples show that AI is already making a measurable impact in nursing care, with proven applications across diverse clinical settings. These success stories demonstrate that AI is not replacing nurses but empowering them to deliver safer, more efficient, and more compassionate care. However, often nursing professionals have so far had only limited direct experience with AI systems in their daily work [11].

2.3. Methodological State of Research

A review of the existent literature reveals several limitations. A considerable proportion of studies are of a descriptive nature and primarily focus on technical performance as opposed to clinical outcomes. A paucity of randomised controlled trials and high-quality real-world implementation studies is evident [6, 8, 21]. Moreover, the development of AI applications is predominantly undertaken within data-rich environments, such as hospitals, thereby constraining their transferability to alternative care settings.

3. Challenges in AI Implementation

AI projects in nursing face numerous challenges due to technical and organizational challenges. But its implementation requires also careful attention to social and ethical issues such as privacy, bias, and the need to preserve clinical judgment [7, 9, 22-24].

3.1. Technical Challenges

The primary technical challenges are not only confined to the development of the AI algorithm itself, but also extend to the creation of a robust, secure, and interoperable technological ecosystem that can reliably deliver AI-driven insights within the dynamic and demanding environment of nursing care.

The implementation process is complicated by technical and organisational integration issues, including the necessity of connecting to existing IT

systems in healthcare settings. A significant technical challenge is the design of AI tools that are not only functional but also seamlessly integrate into the existing, often complex, workflows of nurses.

The seamless integration of AI systems with existing, often fragmented, healthcare IT infrastructure is often of crucial importance. The utilisation of AI tools is contingent upon the availability of comprehensive, real-time data from electronic health records (EHRs), medical devices and nursing documentation, amongst other sources. Nevertheless, a considerable number of healthcare organisations function with heterogeneous systems that are deficient in standardised data formats and communication protocols. This complicates the effective aggregation and analysis of data for artificial intelligence applications.

The quality of the data is often compromised by the presence of heterogeneous, incomplete, or non-digitised nursing records. The dearth of data availability beyond the confines of hospital settings, the paucity of digital infrastructure, and the absence of standardised interoperability protocols collectively impede the effective development and deployment of AI solutions in nursing. The quality of AI models is contingent upon the quality of the data on which they are trained. The generation of unstructured data by nursing care is a significant phenomenon. Examples of such data include clinical notes, nursing assessments and verbal reports. The quality of this data is often inconsistent, incomplete or poorly documented. The absence of standardised data collection methodologies across diverse care settings and institutions engenders substantial challenges in the development of reliable and generalisable AI models.

The implementation of AI on a large scale necessitates significant computational resources, particularly for complex models such as deep learning. It is evident that a significant number of healthcare organisations are not equipped with the requisite hardware and robust network infrastructure to support the processing demands of AI applications. The management of ongoing maintenance and updates of these models is complex and resource-intensive.

3.2. Organizational and Human Challenges

The potential of AI to transform nursing education is significant. It has the capacity to enable personalised learning, enhance clinical judgment through immersive tools such as virtual reality (VR) simulations and chatbots, and reduce educator workload through automated feedback and content generation. However, this potential is currently constrained by a critical lack of standardisation in the incorporation of AI literacy within nursing curricula. This has given rise to concerns regarding overreliance on AI, erosion of interpersonal skills, and the risk of plagiarism. These issues underscore the pressing need to integrate fundamental AI concepts, ethical training, and practical skills into nursing programs.

Furthermore, it is imperative to acknowledge the pivotal role of acceptance of AI systems by nurses and care recipients, in addition to effective organisational change management, for the attainment of successful outcomes. The integration of artificial intelligence (AI) within nursing practice is frequently impeded by the limited involvement of nursing staff in the design and development of AI systems. This often results in solutions that do not align with clinical realities. This issue is further compounded by the considerable workload and limited availability of time experienced by healthcare professionals, who often encounter difficulties in engaging with new technologies. Moreover, the dearth of adequate digital competencies among the workforce further curtails the capacity to adopt and utilise AI tools with any degree of effectiveness. This situation serves to emphasise the necessity for targeted training initiatives and inclusive participation in endeavours pertaining to digital transformation.

3.3. Ethical and Social Challenges

The integration of artificial intelligence within the domain of nursing represents a substantial transformation. However, it concomitantly gives rise to significant ethical and social challenges, necessitating the formulation of effective strategies to ensure the judicious implementation of this technology. The primary concerns encompass data privacy breaches, algorithmic bias, and an absence of transparency in AI decision-making – a phenomenon often termed the "black box" problem. This can result in the erosion of patient trust and the exacerbation of health inequalities. These risks are exacerbated by the reluctance of nurses to adopt these systems, often due to concerns regarding job security and a lack of confidence in the transparency and reliability of AI-driven decision-making processes. Moreover, unequal access to AI technologies, particularly in resource-constrained settings, has the potential to exacerbate existing disparities in healthcare delivery. In order to establish trust and ensure the equitable adoption of AI systems, it is essential that they are designed with robust data security, clear accountability, and explainable outputs. This will enable nurses and patients to understand and engage with AI in a meaningful way. It is imperative to recognise that addressing these ethical and social dimensions is not a matter of choice; rather, it is fundamental for the sustainable and trustworthy integration of AI in nursing.

4. AI Readiness and Success Factors

4.1. Success Factors

As demonstrated by the examples in Section 2.2, the implementation of AI offers significant opportunities to enhance patient care, support clinical

decision-making, and alleviate workload pressures. The successful implementation of artificial intelligence (AI) in nursing care necessitates a comprehensive, multi-faceted approach that extends beyond the mere deployment of technology. It is imperative that a strategic alignment of technical, organisational, ethical and human factors is achieved.

In addition to the solutions to the challenges identified in Section 3 being implemented, it is essential that nurses are involved in the design of the system in a timely manner. This will help to build trust and demonstrate the benefits. The development of AI tools must be a collaborative process with nurses to ensure that these tools are intuitive, seamlessly integrate into existing workflows, and provide genuine support to nurses without disrupting their work. It is vital to facilitate proactive communication and to ensure the involvement of nurses in the implementation process in order to overcome fears of job displacement and to build trust in the technology.

Consequently, the successful implementation of AI in nursing is not merely a technical project, but rather a transformational change management initiative. A comprehensive strategy is imperative, one that prioritises the human element, ensures ethical integrity, and fosters a supportive organisational culture. This strategy must also leverage technology to enhance, rather than replace, the irreplaceable human aspects of nursing care. The necessity of organisational readiness across multiple dimensions is apparent. These include technical infrastructure, human resources, and ethical considerations.

4.2. Measuring AI Readiness

Notwithstanding the considerable potential of artificial intelligence in the field of nursing care, initiatives often prove unsuccessful due to a lack of fulfilment of sociotechnical, regulatory and organisational prerequisites. To address this, an evidence-based tool was developed to systematically support the planning, implementation and evaluation of AI in nursing care: the AI Nursing Care Readiness Assessment (AINCRA) tool [25, 26]. This structured, validated framework enables stakeholders to proactively assess and strengthen their project readiness. AINCRA is not merely a checklist but a holistic maturity model designed to reflect the complex interplay of technical, ethical, social, and organisational factors critical for successful AI integration in nursing.

The development of AINCRA followed a sequential exploratory multimethod design, ensuring methodological rigor and stakeholder relevance. The process was structured in five distinct, iterative phases:

1. Preliminary Foundation: This study builds on prior research by the research team, which includes a rapid review of 292 publications on AI in nursing care, an expert workshop (n = 21), expert interviews (n = 14), and an online survey (n = 53). The findings of this study identified five key dimensions: regulatory, processual/translational, technical, social/ethical, and community building.

2. Systematic Literature Review (SLR): A comprehensive SLR was conducted across five major databases (Scopus, PubMed, ACM, AIS, ECONLit) to identify existing AI readiness models in healthcare and nursing. The identification of AI readiness attributes was achieved through inductive grouping, which resulted in the formation of 13 dimensions.

3. Integration and Consensus Building: The amalgamation and harmonisation of all attributes from preliminary work and the SLR was achieved through two rounds of nominal group consensus involving five expert stakeholders. This process entailed the consolidation of semantically equivalent attributes and the elimination of those not directly influenced by AINC project stakeholders. The outcome of this process was a refined set of 67 attributes across five core dimensions.

4. Indicator Development and Pilot Testing: A preliminary version of AINCRA was developed, incorporating feedback from 13 AINC project experts. The team established a standardised 5-level maturity scale (initial, assessing, determined, managed, and optimised) and employed GPT-4 to generate initial level descriptions. These descriptions were then subjected to a rigorous review process, during which they were revised by the entire research team.

5. The process of expert validation and refinement is a key component in the overall methodology. The final phase of the study involved think-aloud interviews (n = 18 experts) and focus group discussions with experts who had direct experience of conducting AINC projects. This methodological approach yielded profound insights into the cognitive processes, usability, and perceived relevance of each attribute.

The final AINCRA is comprised of five core dimensions and 69 attributes.

Regulatory Requirements and Aspects (9 attributes): The focus of this dimension is on legal compliance, with specific reference to the EU AI Act and the GDPR. In addition, data protection and data-sharing models are discussed.

Processual and Translational Requirements and Aspects (40 attributes): This part addresses the most significant innovation of AINCRA. This dimension encompasses the frequently disregarded process of AI integration, incorporating strategic planning, digital readiness, resource allocation, stakeholder engagement, and the practical value of AI systems.

Technical Requirements and Aspects (6 attributes): The evaluation process encompasses a comprehensive assessment of the IT infrastructure deemed essential, the data security protocols, the interoperability standards that are in place, and the AI-specific compute capabilities that are currently available.

Social and Ethical Requirements and Aspects (11 attributes): The dimension focuses on trust, acceptance, fairness, transparency, and the ethical implications of AI on nursing practice and patient care.

Community Building Requirements and Aspects (3 attributes): The promotion of collaborative networks for knowledge exchange and shared development beyond the scope of individual projects is of paramount importance.

Each of the 69 attributes is described across five maturity levels, providing a clear progression from initial awareness to optimised, sustainable integration.

The AINCRA provides support to researchers, clinicians and health scientists in the planning, assessment and improvement of AI projects. The AINCRA's primary objective is to encourage more effective and sustainable AI integration in nursing [26]. The value of the latter is attributable to its practical application across the entire project lifecycle:

Project Planning & Initiation: AINCRA serves as a critical self-assessment tool to determine if a project is viable. The process facilitates the identification of critical "deal-breaker" criteria (e.g., poor data quality, lack of regulatory clarity) prior to the allocation of substantial resources. It provides a framework for the selection of suitable partners and the delineation of project scope.

Project Management & Monitoring: By enabling the repeated administration of assessments, AINCRA facilitates the monitoring of team progress over time, the identification of areas requiring improvement, and the proactive adjustment of strategies. The platform fosters transparent, data-driven discussions among a diverse range of stakeholders, including clinicians, researchers, IT specialists and ethicists.

Grant Applications & Funding: A high AINCRA score has the potential to enhance the credibility of grant proposals by evidencing a meticulously formulated, pragmatic and judicious approach to AI integration, thereby rendering projects more appealing to funding bodies.

Evaluation and Learning: AINCRA provides a structured framework for summative evaluation, thereby facilitating comprehension of the factors that influenced the success or failure of a project. The primary function of the system is to facilitate the documentation and comparison of disparate AINC projects, thereby contributing to the collective knowledge base.

Interprofessional Collaboration: The provision of a common language and framework is pivotal in AINCRA, as it bridges the so-called "domain language" gap between healthcare professionals, researchers and technologists, thereby fostering a shared understanding and more effective teamwork.

AINCRA is more than just an assessment tool; it is a transformative framework for responsible innovation in nursing care. It addresses a critical gap in the literature by offering a scientifically validated, context-specific, and practical solution to the complex challenges of implementing AI in healthcare. By systematically reflecting on all dimensions of readiness – from the technical to the ethical – AINCRA empowers practitioners, researchers, and clinicians to move beyond technological enthusiasm and build sustainable, effective, and ethically sound AI solutions for the future of nursing.

5. Conclusion

The field of artificial intelligence is undergoing a transition from experimental tools to practical applications in nursing, thereby demonstrating clear potential to reduce nurse workload and enhance care quality. However, the realisation of this potential requires a shift beyond technological solutions alone. A significant gap exists between research and real-world implementation, hindering the development of user-centred, needs-driven solutions.

To bridge this gap, a strategic, interdisciplinary approach is essential. This approach must harmonise technological innovation with ethical principles, clinical practice, and workforce needs. It is crucial that nurses play an active role in the design, evaluation, and implementation of AI systems. Their frontline expertise is vital to ensure that tools are trustworthy, explainable, safe, and truly aligned with patient care and professional values.

The future of AI in nursing is defined by the delivery of predictive, personalised, efficient, and human-centred care. This vision is already becoming manifest through evidence-based applications:

Predictive and Preventive Care: The utilisation of artificial intelligence (AI) facilitates the analysis of real-time data from wearables, electronic health records (EHRs), and vital signs. This analysis enables the anticipation of potential risks, including sepsis, falls, and chronic disease exacerbations. Consequently, proactive interventions can be initiated in a timely manner.

Personalised and Precision Nursing: The care plans will be tailored to individual patients based on genetics, lifestyle, and continuous health data, especially for those with chronic conditions.

Advanced Decision Support: The integration of AI into clinical workflows will facilitate the delivery of real-time, evidence-based recommendations for interventions and care prioritisation, thereby enhancing clinical reasoning without compromising the autonomy of nurse judgment.

Assistive Technologies: In the context of long-term care, the utilisation of AI-powered robots to assist with physical tasks, such as patient lifting, has the potential to address the challenges posed by staffing shortages.

Virtual Nursing and Remote Care: The integration of telehealth with artificial intelligence (AI) has the potential to enhance accessibility to healthcare, particularly in regions characterised by limited resources. The utilisation of virtual assistants to facilitate triage, education, and follow-up is a promising development.

Workforce Optimization: AI will optimize staffing, predict workload peaks, reduce administrative burden, and provide digital tools to support nurse well-being.

In order to achieve this future, stakeholders must prioritise participatory design, real-world integration,

workforce training, and the development of evidence-based AI applications. In order to address this issue, it is essential to invest in robust ethical frameworks, to integrate AI literacy into nursing education, and to build the necessary digital infrastructure. The overarching objective is not to substitute for nurses, but rather to empower them. The utilisation of artificial intelligence as a strategic partner enables nurses to deliver care of a higher standard, characterised by compassion and sustainability, for all patients, thereby ensuring that technology aligns with the fundamental values of the profession.

References

[1]. State of the world's nursing report 2025: Investing in education, jobs, leadership and service delivery, *World Health Organization*, 2025.

[2]. Clinician of the future 2026: Nurses edition, *Elsevier*, 2026, https://www.elsevier.com/insights/clinician-of-the-future/2026

[3]. T. Krick, O. Huter, D. Domhoff, A. Schmidt, et al., Digital technology and nursing care: A scoping review on acceptance, effectiveness and efficiency studies of informal and formal care technologies, *BMC Health Services Research*, Vol. 19, 2019, 400.

[4]. C. Grosan, A. Abraham, Rule-based expert systems, Chapter 7, in Intelligent Systems: A Modern Approach, *Springer*, Berlin, 2011, pp. 149-185.

[5]. B. Wahl, et al., Artificial intelligence and global health, *BMJ Global Health*, Vol. 3, Issue 4, 2018, e000798.

[6]. K. Seibert, D. Domhoff, D. Bruch, et al., Application scenarios for artificial intelligence in nursing care: *Rapid review, Journal of Medical Internet Research*, Vol. 23, Issue 11, 2021, e26522.

[7]. S. Hassanein, N. Elkheir, M. Al-Yami, Artificial intelligence in nursing: An integrative review of clinical and operational impacts, *Frontiers in Digital Health*, Vol. 7, 2025, 1552372.

[8]. S. O'Connor, et al., Artificial intelligence in nursing and midwifery: A systematic review, *Journal of Clinical Nursing*, Vol. 32, Issue 13-14, 2023, pp. 2951-2968.

[9]. K. Seibert, et al., Exploring needs and challenges for AI in nursing care, *BMC Digital Health*, Vol. 1, 2023, 13.

[10]. D. Hunstein, M. Fiebig, Staff management with AI: Predicting the nursing workload, *Studies in Health Technology and Informatics*, Vol. 315, 2024, pp. 231-235.

[11]. D. Sommer, et al., Nurses' perceptions, experience and knowledge regarding artificial intelligence: Results from a cross-sectional online survey in Germany, *BMC Nursing*, Vol. 23, Issue 1, 2024, 205.

[12]. L. Hung, et al., Ethical considerations in the use of social robots for supporting mental health and wellbeing in older adults in long-term care, *Frontiers in Robotics and AI*, Vol. 12, 2025, 1560214.

[13]. A. Lukas, et al., 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, *Zeitschrift für Gerontologie und Geriatrie*, Vol. 54, Issue 7, 2021, pp. 685-694 (in German).

[14]. K. Majjouti, KIADEKU: Identification of wound types with AI, in *Proceedings of the 1st International Conference on AI in Medicine and Healthcare (AiMH)*, 2025, pp. 83-85.

[15]. K. Schwabe, et al., Reducing nurses' workload with an AI speech assistant for documentation, in *Proceedings of the 1st International Conference on AI in Medicine and Healthcare (AiMH)*, 2025, pp. 91-95.

[16]. R. J. Gallo, et al., Effectiveness of an artificial intelligence-enabled intervention for detecting clinical deterioration, *JAMA Internal Medicine*, Vol. 184, Issue 5, 2024, pp. 557-562.

[17]. Y. Liu, et al., Evaluation of a wearable wireless device with artificial intelligence, iThermonitor WT705, for continuous temperature monitoring for patients in surgical wards: A prospective comparative study, *BMJ Open*, Vol. 10, Issue 11, 2020, e039474.

[18]. D. Ferizaj, S. Neumann, Assessing perceptions and experiences of an AI-driven speech assistant for nursing documentation: A qualitative study in German nursing homes, *Lecture Notes in Computer Science*, Vol. 14688, 2024, pp. 19-32.

[19]. N. G. da Rosa, et al., Nursing workload: Use of artificial intelligence to develop a classifier model, *Revista Latino-Americana de Enfermagem*, Vol. 32, 2024, e4239.

[20]. A. Cho, et al., Development of an artificial intelligence-based tailored mobile intervention for nurse burnout: Single-arm trial, *Journal of Medical Internet Research*, Vol. 26, 2024, e54029.

[21]. H. von Gerich, et al., Artificial intelligence-based technologies in nursing: A scoping literature review of the evidence, *International Journal of Nursing Studies*, Vol. 127, 2022, 104153.

[22]. K. Wolf-Ostermann, et al., Concept for embedding AI systems in nursing care, *University of Bremen*, 2021.

[23]. K. Wolf-Ostermann, H. Rothgang, Digital technologies in care – What can they achieve?, *Bundesgesundheitsblatt – Gesundheitsforschung – Gesundheitsschutz*, Vol. 67, Issue 3, 2024, pp. 324-331 (in German).

[24]. R. El Arab, et al., The role of AI in nursing education and practice: Umbrella review, *Journal of Medical Internet Research*, Vol. 27, 2025, e69881.

[25]. K. Seibert, K. Wolf-Ostermann, et al., ProKIP: Rationale and development of the AI-Nursing-Care-Readiness-Assessment (AINCRA), in *Proceedings of the 1st International Conference on AI in Medicine and Healthcare (AiMH)*, 2025, pp. 56-60.

[26]. K. Seibert, et al., A readiness assessment for AI in nursing care projects (AINCRA): A multimethods study, *JMIR Nursing*, 2026, 10.2196/84148 (in print).

(031)

SAVE & SAFE: An AI-Supported Assistive System for Fall Prevention and Nursing Workload Reduction in Acute Geriatric Care

E. Mena [1,2], T. Schultz [3] and K. Wolf-Ostermann [1,2]
[1] University of Bremen, Institute for Public Health and Nursing Research (IPP), PF 330440, D-28334 Bremen, Germany
[2] Health Sciences Bremen, University of Bremen, Bremen, Germany
[3] University of Bremen, Cognitive Systems Laboratory, Enrique-Schmidt-Str. 5, 28359 Bremen, Germany
Tel.: + 49 421 218 60560
E-mail: e.mena@uni-bremen.de

Summary: Falls and fall-related consequences are highly prevalent among older people and substantially burden acute geriatric care and nursing staff. The SAVE & SAFE project addresses this challenge by developing and implementing a novel care model that combines an AI-supported assistive system with organisational innovation to improve fall prevention while reducing psychosocial and time-related workload among nurses. In three acute geriatric hospitals in North Rhine-Westphalia, sensor-based monitoring, automated detection of "hard" and "soft" falls, bed-exit alarms, and partially automated mobility assessment based on the Short Physical Performance Battery are integrated into routine care. Specially trained "fall prevention professionals" coordinate preventive measures across care settings during the three-month high-risk period following hospitalisation. The primary evaluation focus is on nursing staff and uses a cluster-randomised stepped-wedge design with COPSOQ surveys, time-action analyses of fall-related activities, biosignal-based case analyses and semi-structured interviews. The project is expected to provide evidence on feasibility, organisational requirements and the impact of AI-enabled fall prevention on nurses' workload and care quality in real-world geriatric practice.

Keywords: Fall prevention, Acute geriatrics, Artificial intelligence in nursing, Assistive monitoring system, Nursing workload, Implementation study, Digital health intervention.

1. Introduction

Falls and fall-related consequences are highly prevalent among older people and substantially burden acute geriatric care and nursing staff [1-3]. The SAVE & SAFE project addresses this challenge by developing and implementing a novel care model that combines an AI-supported assistive system with organisational innovation to improve fall prevention while reducing psychosocial and time-related workload among nurses [4, 5]. In three acute geriatric hospitals in North Rhine-Westphalia, sensor-based monitoring, automated detection of "hard" and "soft" falls, bed-exit alarms, and partially automated mobility assessment based on the Short Physical Performance Battery are integrated into routine care. Additionally, specially trained "fall prevention professionals" coordinate and support preventive measures across care settings during the three-month high-risk period following hospitalisation.

The primary evaluation focus is on nursing staff and uses a cluster-randomised stepped-wedge design with COPSOQ surveys, time-action analyses of fall-related nursing activities, biosignal-based case analyses of mental strain, and semi-structured interviews. The project is expected to provide evidence on feasibility, organisational requirements, and the impact of AI-enabled fall prevention on nurses' workload and care quality in real-world geriatric practice [4].

2. Methods

2.1. Intervention

The intervention combines a sensor-based intelligent care system (intelligentes Pflegesystem, iPS) with new organisational roles and processes for fall prevention. In participating acute geriatric wards, the iPS is sensor-based and designed to detect hard falls (sudden falls to a lower level), soft falls (gradual sliding to the floor), and bed-exit events, with alerts forwarded either to the ward's standard alarm system or to nurses' tablets, depending on local implementation. In addition, a partially automated Short Physical Performance Battery (SPPB) assessment supports standardized evaluation of patients' mobility and fall risk.

Specially trained "fall prevention professionals" act as key persons for coordinating the intervention across professional groups and sectors, ensuring continuity of fall-preventive care during the three-month high-risk period from acute hospitalisation into subsequent care settings [4]. The stepped-wedge implementation is organised in three

crossover cohorts at three-month intervals; the first cohort, comprising two wards, entered the intervention phase in April 2026 and the systems have now been activated.

2.2. Study Design and Setting

The evaluation of the primary effect focuses on nursing staff and follows a cluster-randomised stepped-wedge design. Three hospitals in North RhineWestphalia with in total six acute geriatric wards participate as clusters that sequentially transition from control to intervention conditions. All nursing staff on the participating wards are eligible. An open cohort approach and staggered implementation are used to reflect real-world conditions [4].

2.3. Data Collection and Outcomes

Psychosocial working conditions and strain are assessed using the Copenhagen Psychosocial Questionnaire (COPSOQ) [6] at baseline (control phase) and at the end of the intervention phase. To capture time expenditure related to fall-preventive and fall-related nursing activities, a structured time-action analysis [7] is conducted, comprising 3024 hours of observation across the six wards over the course of the study. In addition, biosignal-based single-case analyses explore patterns of mental strain and relief among a subsample of nurses. Semi-structured interviews with nursing staff provide qualitative insights into experiences with the new care model, perceived usefulness and barriers to implementation [4].

3. Expected Results

We expect that the SAVE & SAFE intervention will reduce psychosocial strain and documentation related workload among nurses on acute geriatric wards. In line with the project's logic model, a more targeted, information-rich and continuous fall preventive care process is anticipated to increase perceived control over critical situations and improve coordination within the team. Quantitatively, this should be reflected in more favourable COPSOQ scores and a reduction in time spent on fall-related activities per shift. At the same time, the intervention is expected to support earlier detection of mobility decline and fall risk, thereby facilitating timely preventive action and potentially reducing the incidence or severity of falls.

4. Discussion and Conclusions

The project addresses several known challenges in AI in-nursing implementations, including limited data availability outside intensive care, insufficient involvement of nursing practice and the high-risk character of many AI applications in healthcare. By embedding the AI-supported assistive system in an explicitly designed care model with strong participation of nursing staff and clear organisational structures, SAVE & SAFE aims to enhance both feasibility and acceptance in routine practice. The project also revealed an important implementation challenge, as the data volume generated by the AI function exceeded the capacity of the hospital infrastructure. To address this, an additional sensor was implemented to relay alerts, while the iPS sensor was limited to documenting the fall trajectory.

The stepped-wedge design allows robust evaluation under real-world conditions while ensuring that all participating wards eventually receive the intervention. Overall, the project is expected to contribute evidence on how AI-enabled monitoring and decision support can be leveraged to improve fall prevention and working conditions in acute geriatrics, and may serve as a model for broader implementation in geriatric care. The presentation will highlight the practical and organisational challenges of the transition phase preceding and accompanying system deployment in the participating hospitals.

Acknowledgements

We sincerely thank the staff and patients of Alexianer Krefeld GmbH, Alexianer St. Martinus Düsseldorf GmbH and the Johanniter GmbH for their invaluable support during data collection. The study was funded by the German Federal Joint Committee (G-BA) - Innovation Fund (grant: 01NVF23105) and was prospectively registered in the German Clinical Trials Register (DRKS00035598). The official G-BA project page provides project information, including the funding number and project period:https://innovationsfonds.g-ba.de/projekte/save-safe.629.

References

[1]. M. Montero-Odasso, N. van der Velde, F. C. Martin, M. Petrovic, et al., World guidelines for falls prevention and management for older adults: A global initiative, *Age and Ageing*, Vol. 51, Issue 9, 2022, afac205.

[2]. L. Clemson, S. Stark, A. C. Pighills, N. J. Fairhall, et al., Environmental interventions for preventing falls in older people living in the community, *Cochrane Database of Systematic Reviews*, Issue 3, 2023, CD013258.

[3]. Techniker Krankenkasse, Zum Internationalen Tag der Pflegenden: Krankenstand bei Pflegekräften auf Rekordhoch, *Techniker Krankenkasse*, Hamburg, 2023 (in German).

[4]. E. Mena, M. Schlacke, M. Puhlemann, M. Fünfstück, et al., Innovative Versorgungsform zur Entlastung von Pflegefachkräften in der Akutgeriatrie: Eine cluster-randomisierte Studie zur Evaluation psychosozialer und zeitlicher Entlastungseffekte, *Zenodo*, 10.5281/zenodo.17064244, 2025 (in German).

[5]. K. Wolf-Ostermann, D. Fürstenau, S. Theune, L. Bergmann, et al., Konzept zur Einbettung von KI-Systemen in der Pflege: Sondierungsprojekt (SoKIP), *Universität Bremen*, Bremen, 2021 (in German).

[6]. M. Nübling, M. Vomstein, S. G. Schmidt, S. Gregersen, et al., Psychosocial work load and stress in the geriatric care, *BMC Public Health*, Vol. 10, 2010, 428.

[7]. M. Lopetegui, P.-Y. Yen, A. Lai, J. Jeffries, et al., Time motion studies in healthcare: What are we talking about?, *Journal of Biomedical Informatics*, Vol. 49, 2014, pp. 292-299.

(032)

AI-Driven Services for Care Facilities: Results from a Longitudinal Field Study

R. E. Paul [1] and K. Wolf-Ostermann [2] and T. Schultz [1]
[1] University of Bremen, Cognitive Systems Laboratory, Enrique-Schmidt-Str. 5,
28359 Bremen, Germany
[2]University of Bremen, Institut für Public Health und Pflegeforschung, Universitätsallee 1B,
28359 Bremen, Germany
Tel.: + 49 421 218 64270
E-mail: tanja.schultz@uni-bremen.de

Summary: This paper presents results from a longitudinal field study focused on AI-driven models and methods for reliably recognizing the traits and activities of older adults. Unobtrusive, privacy-protecting depth sensors were installed in three nursing homes, capturing depth data from 45 volunteer residents aged 60+ years over 21 months. The dataset comprises over 58000 hours of 16-bit depth data recorded at 10 frames per second, with a resolution of 640 × 480 pixels. We describe the pipeline we implemented to fully automatically (a) process the incoming data streams, (b) remove artifacts from timeouts, frozen sensors, and alike, (c) detect room occupancy, (d) extract human skeleton information, (e) discriminate helpers from residents by age classification, and (f) recognize traits and activities of the residents, such as sitting, walking, and bed exit. The resulting models and methods are evaluated on an annotated subset using cross-validation, with accuracy and F-score measured against manually annotated ground truth. Our experimental results emphasize the potential of AI for personalized care.

Keywords: AI in patient care, AI for predictive analysis, Human activity recognition.

1. Introduction

The number of older adults requiring long-term care in Germany has increased over the years, with projections indicating it could reach 6.3 million by 2035 [1]. The growing population necessitates additional caregivers in long-term care facilities. Owing to demographic trends, Germany is facing a severe and escalating shortage of caregivers [2]. This situation underscores the need for automated monitoring systems capable of detecting routine activities and supporting personalized clinical decision-making. Although wearable-based monitoring is effective, it is impractical for continuous use due to obtrusiveness, battery limitations, signal degradation from body-mounted electrodes, and fluid-conductivity issues [3, 4]. Furthermore, wearables often cause discomfort for older adults and are frequently forgotten to put on, lost, or misplaced.

In contrast, environmental sensors such as pressure, temperature, and motion sensors can be installed remotely while still monitoring a person's activities. However, these sensors are typically not ubiquitous or continuous and are often triggered at set intervals, possibly requiring several units for comprehensive monitoring. Vision-based monitoring provides a non-wearable alternative, where cameras are mounted at a distance and powered to enable continuous monitoring of human behavior. RGB cameras, in particular, are associated with privacy concerns [5], whereas depth cameras offer a privacy-preserving solution. Depth images capture the structure of subjects without revealing distinguishing features, such as facial details, making them well-suited for privacy-conscious monitoring applications [6].

1.1. Related Work

Previous studies have conducted continuous monitoring of younger and older adults using wearable, vision-based, and environmental sensor-based systems. Studies show that fitness trackers were used to detect mild cognitive impairment [7] and to assess mobility [8] using short-term longitudinal data collected from older participants. For example, CASAS (Context-Aware Smart Ambient Systems), introduced by Cook et al. [9], is a platform for activity recognition in natural living environments that uses ZigBee sensor mesh to monitor daily activities. It was validated in 32 homes with 20 young participants and demonstrated the feasibility of unobtrusive ambient sensing for recognizing daily living activities in smart-home environments. In vision-based approaches, pose estimation is required to get information on the active human body. The preferred method is to extract human-skeleton information while ignoring background details, as this is most relevant to human-activity models [10], thereby enabling privacy-preserving, computationally efficient activity analysis. Zouba et al. [11] conducted a study in an apartment living lab, collecting data from video (RGB), pressure, and temperature sensors, involving nine older volunteers. Their findings showed that combining multimodal sensing improved the reliability of activity and behavior monitoring in assisted living environments. Skeleton information

extracted from both the RGB and depth cameras (Kinect) was used to discriminate between daily activities in [12]; the study used data from four younger adults in natural environments.

1.2. Summary of Contributions

Most existing studies concentrate on short-term data collected in controlled environments with younger adults. To overcome these limitations, we developed a monitoring system for older adults using an AI-driven approach and continuous longitudinal monitoring across three nursing homes employing privacy-preserving depth cameras. This paper presents the results from this longitudinal field study and extends our previous work [13]. The current study introduces new downstream models, including artifact removal, image enhancement, and age-group recognition. The dataset used herein is an expansion of our earlier study, with models trained on this extended data. We present a system that filters irrelevant depth frames, extracts joint positions, applies an age-group recognition model to distinguish staff from residents, and ultimately uses an activity recognition model to detect bed-exit attempts.

2. Methodology

Depth data were collected using a wall-mounted DHC (Digital HealthCare System GmbH) sensor box, which includes an integrated depth sensor installed in resident rooms at three nursing homes [14]. The resulting dataset, ETAP-DID (ETAP-Depth Imaging Dataset), includes over 58000 hours collected over 21 months from 45 residents aged 60+ years. Each resident room has a single resident; hence each sensor observes a single resident. Data were recorded at 10 frames per second (fps) with a resolution of 640×480. Each pixel in a depth frame denotes the distance from the sensor.

The proposed pipeline (Fig. 1) consists of:

(1) Artifact removal for handling sensor failures;
(2) Occupancy detection to identify presence in rooms;
(3) Skeleton extraction from depth frames;
(4) Classification of helpers versus residents using age recognition;
(5) Activity recognition, including sitting, walking, and bed exit.

Frozen frames are artifacts in the data, in which depth images are frozen due to sensor or network errors. We identify frozen frames by comparing frame-to-frame variations. If the variation is close to zero, it indicates no change between the frames, meaning they are frozen [14]. Another artifact appears as white or black patches in the images; one reason is reflection from mirrors or glasses. These patches can be covered by identifying the pixel values of black (0) and white (65535), and replacing them with the average non-artifact pixel value from the adjacent pixels. Occupancy detection is performed using a Decision Tree (DT) model trained on contours extracted via background subtraction [15] to classify frames as occupied or unoccupied [14].

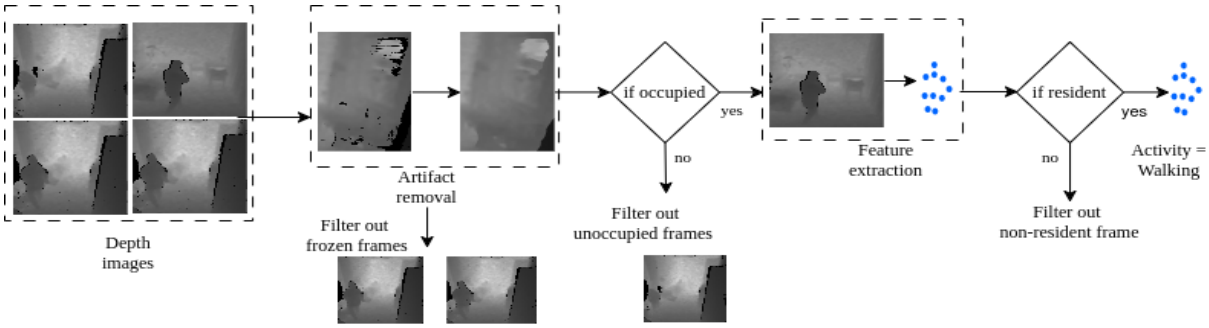


Fig. 1. The implemented AI-driven pipeline with all the modules and results.

We used skeleton extraction to obtain skeleton information, as done in prior work. ResNet model [16] performed best for skeleton extraction using depth data in prior studies [17]. We extracted 15 2D joint positions from the depth images using the ResNet18 model [18]. As our dataset included 31 sensors, we used relative joint positions with respect to the torso joint to remove sensor bias in the models. Using the extracted skeleton features in the downstream tasks reduces the hardware requirements by a large margin. Depth sensors capture both staff and residents. Since monitoring is meant for residents, we use age-group recognition to distinguish the two groups, as staff is, on average, significantly younger than residents. For discrimination, we used relative joint positions with a Spatio-Temporal Graph Convolutional Network (ST-GCN). For daily activity recognition, we applied a Long Short-Term Memory (LSTM) model on joint positions to classify sitting, standing, walking, transition (sit-to-stand, stand-to-sit), and bed exit.

The full pipeline and results for the given reference depth images are shown in Fig. 1. We start with several depth images, including frames with older adults and staff, unoccupied frames, and frozen frames. Our first artifact removal module removes the frozen frames and enhances the depth images by detecting artifact

patches. Our occupancy detection model filtered out the unoccupied frames. The skeleton joints are extracted from depth images in the feature extraction module. The extracted features are used in the age-group recognition module to filter out the frames with staff. Then, in our final module activity recognition, the activity performed by the older adult is identified. This identified activity serves as an indicator of bed-exit attempts or falls.

3. Results and Discussion

Frozen frames, identified as artifacts in the data, are excluded from processing, while image patches are corrected to enhance overall depth-image quality. No data augmentation was applied within this pipeline. For the occupancy detection model, 17 hours of unoccupied-room data and 13 hours of occupied-room data, both manually labeled, were used, yielding an accuracy of 89.42 % compared to a majority-class baseline of 56.70 % on this set. Occupancy detection operates on a per-frame basis, comparing each frame with 15 adjacent frames. Hyperparameters were selected via grid search, using a maximum depth of 10 and the entropy criterion. The model was subsequently trained. The current evaluation uses sensor-dependent splits, which allow samples from the same sensor to appear in both the training and test sets. The final detector is used to label the entire dataset and filter out the usage of unoccupied rooms.

We used the Computer Vision Annotation Tool (CVAT) to label 15 2D skeleton joints and activities performed, and to distinguish between staff and residents. We used 13 hours of data with a single person in the frame (resident = 11 hours; staff = 2 hours) and split it into train, validation, and test sets at 80:10:10. We trained a model ResNet from scratch on our annotated dataset rather than using ImageNet-pretrained weights. This choice was motivated by the domain gap between RGB images and depth data. The model was trained for 50 epochs with the Adam optimizer. L1 loss (Eq. (1)) and Percentage of Correct Keypoints (PCK) (Eq. (2)), where keypoints denote skeleton joints, were used as evaluation metrics, achieving a PCK of 72.54 % and a normalized L1-based score (derived from L1 loss) of 77.62 % on the test set.

$$L1 = \frac{1}{N}\sum_{i=1}^{N} norm\, K_i\text{-}J_i, \quad (1)$$

$$PCK = \frac{100}{N}\sum_{i=1}^{N} \begin{matrix} 1 & if norm K_i\text{-}J_i < 15\ cm \\ 0 & otherwise \end{matrix}, \quad (2)$$

where K_i and J_i denote predicted and ground-truth joint positions, respectively, and N is the total number of keypoints (skeleton joints).

For both age-group and activity recognition, we used a window length of 12 frames with a 6-frame overlap to capture temporal dynamics. An 80:20 training-testing split was used. A smaller dataset to maintain a moderate class balance across activities and age groups, ensuring reliable classification performance and avoiding dominance by the most frequent classes. To ensure robustness, we employed 5-fold cross-validation and averaged results across folds. For age-group recognition, we used 4 hours of resident data and 2 hours of staff data for binary classification between older and younger adults. The model was trained for 30 epochs, and we used the Adam optimizer for both age-group and activity recognition models. We used relative joint positions for the ST-GCN, achieving an average accuracy of 87.70 % and an average F1-score of 87.51 %, with a majority-class guess baseline of 66.70 % for discriminating between young and old.

For activity recognition, we used 4 hours of resident data, including the following activities: stand, walk, sit in a chair, sit in bed, sit-to-stand, and stand-to-sit. Sitting in bed was considered an activity that corresponded to bed-exit. The dataset is moderately imbalanced, with the most frequent class (sitting in a chair), and the majority-class guess baseline accuracy is 27 %. We used relative joint positions as inputs to the LSTM, achieving an average accuracy of 82.37 % and an average F1-score of 81.68 % to discriminate the 6 activities.

Our results demonstrate that AI-driven monitoring using privacy-preserving depth sensors is feasible for long-term care facilities. The modular design, in which each model is trained separately, allows for flexible deployment and simplifies updating or retraining individual components, but error generated in one module can be propagated to downstream stages. For example, false negatives in occupancy detection led to dropping frames with residents, and noisy skeleton joints extracted affect the age-group and activity recognition. To mitigate this, we incorporated partial sanity checks for the temporal consistency of joint positions and classified activities. We used a smaller set from our test data (30 minutes of resident, 10 minutes of staff, and 10 minutes of unoccupied frame from the same sensor, and 2 activities) to test the full pipeline. The accuracy of activity recognition results was 72.64 %.

Another limitation was that unoccupied frames were erroneously classified as occupied during feature extraction, leading to pseudo-joints. This happens because we trained our pose estimation model with occupied frames; the model was forced to generate joints. Our next step would be to add such frames to the pose estimation validation set and further fine-tune the model. In this study, age group was used as a distinguishing factor because residents were older than 65 years, while staff members were comparatively younger. This assumption may not always hold in every nursing home setting, where some residents may be younger than staff members. In such cases, incorporating more robust person identification methods may provide a more reliable alternative for distinguishing between residents and staff. The activity recognition focuses only on bed-exit attempt detection; we plan to use it for behavioral analysis of each activity.

Classification tasks achieved significant improvements over the majority-class guess baseline. Skeleton extraction performed robustly with several sensors. The models need further improvement, as errors in one model can propagate to subsequent stages in the pipeline. Even though we collected a lot of data, we could use only a very small subset due to annotation difficulties. Future work will focus on self-supervised models and sanity checks to create a larger weakly supervised dataset. Class imbalance might have reduced the model's sensitivity to the least frequent classes; a larger dataset could help address this. Future work will employ sensor-independent splits for realistic generalization.

4. Conclusions

This paper presents initial results from a privacy-preserving AI pipeline for monitoring older adults in three nursing homes over 21 months. Our pipeline combined image enhancement and filtering of corrupt images, removing images from unoccupied rooms using occupancy detection, pose estimation for skeleton extraction, and skeleton-based age-group and activity recognition. The pipeline demonstrates that unobtrusive, continuous monitoring is feasible without compromising residents' privacy. Future research will focus on detecting events such as falls and on identifying gait changes and mobility decline over longer periods. We will also focus on scenarios with multiple people in a single image and on fusing multiple sensors within a single room to enhance the field of view.

Acknowledgements

The research reported in this paper was partly supported by the German Ministry of Health (BMG), as part of the research project ETAP - Evaluation von teilautomatisierten Pflegeprozessen in der Langzeitpflege am Beispiel von KI-basiertem Bewegungsmonitoring (etap-projekt.de/). We sincerely thank the staff and residents of AWO Karlsruhe GmbH, Stift Tilbeck GmbH Havixbeck, and S. Fabian und S. Rosendahl for their invaluable support during data collection.

References

[1]. Zahl der Pflegebedürftigen steigt bis 2070 deutlich an, *Statistisches Bundesamt*, 2026 (in German).

[2]. D. Heger, et al., Personnel shortages and the provision of long-term care: An empirical analysis of German nursing homes, *The European Journal of Health Economics*, Vol. 26, Issue 9, 2025, pp. 1539-1555.

[3]. E. Murata, T. Osa, T. Yoshikawa, T. Maeno, Behavior monitoring with non-wearable sensors for precision nursing, in *Proceedings of the AHFE International Conference on Safety Management and Human Factors*, 2018, pp. 384-392.

[4]. S. Patel, H. Park, P. Bonato, L. Chan, et al., A review of wearable sensors and systems with application in rehabilitation, *Journal of NeuroEngineering and Rehabilitation*, Vol. 9, 2012, 21.

[5]. E. Chou, et al., Privacy-preserving action recognition for smart hospitals using low-resolution depth images, *arXiv*, 2018, arXiv:1811.09950.

[6]. A. Boulemtafes, A. Derhab, Y. Challal, Privacy-preserving deep learning for pervasive health monitoring: A study of environment requirements and existing solutions adequacy, *Health Technology*, Vol. 12, Issue 2, 2022, pp. 285-304.

[7]. Q. Xu, Y. Jiang, K. Wang, Z. Wang, et al., Prediction of mild cognitive impairment status: Pilot study of machine learning models based on longitudinal data from fitness trackers, *JMIR Formative Research*, Vol. 8, 2024, e48719.

[8]. K. M. Manning, et al., Longitudinal analysis of physical function in older adults: The effects of physical inactivity and exercise training, *Aging Cell*, Vol. 23, Issue 1, 2024, e14029.

[9]. D. J. Cook, A. S. Crandall, B. L. Thomas, N. C. Krishnan, CASAS: A smart home in a box, *Computer*, Vol. 46, Issue 7, 2013, pp. 62-69.

[10]. M. Oczak, et al., Skeleton-based image feature extraction for automated behavioral analysis in human-animal relationship tests, *Applied Animal Behaviour Science*, Vol. 277, 2024, 106338.

[11]. N. Zouba, F. Bremond, M. Thonnat, A. Anfosso, An activity monitoring system for real elderly at home: Validation study, in *Proceedings of the 7th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, 2010, pp. 278-285.

[12]. H. Wu, et al., Human activity recognition based on the combined SVM & HMM, in *Proceedings of the IEEE International Conference on Information and Automation (ICIA)*, 2014, pp. 219-224.

[13]. R. E. Paul, et al., Depth sensor based AI-services for nursing homes, in *Proceedings of the First International Joint Conference on Artificial Intelligence for Healthcare, and Hybrid Models for Coupling Deductive and Inductive Reasoning (HC@AIxIA+HYDRA)*, 2026, pp. 323-335.

[14]. R. E. Paul, et al., Longitudinal data acquisition for AI services in long-term care facilities for older adults, in *Proceedings of the 18th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC)*, 2025, pp. 1099-1110.

[15]. B. Karasulu, S. Korukoglu, Moving object detection and tracking by using annealed background subtraction method in videos: Performance optimization, *Expert Systems with Applications*, Vol. 39, Issue 1, 2012, pp. 33-43.

[16]. X. Sun, J. Xiao, S. Liang, Y. Wei, Compositional human pose regression, in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2602-2611.

[17]. Y. Hartmann, et al., Gait parameter estimation from a single depth sensor, *Journal of Smart Cities and Society*, Vol. 4, Issue 1, 2025, pp. 35-61.

[18]. R. E. Paul, et al., Automated assessment tests with depth sensors in older adults, in *Proceedings of the 47th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2025, pp. 1-7.

(033)

Bibliometric Analysis of Nursing Research Related to Artificial Intelligence in Nursing Care

Şengül Akdeniz [1] and K. Wolf-Ostermann [2]
[1] Akdeniz University, Vocational School of Health Services, 07070 Antalya, Turkey
[2] University of Bremen, Institute for Public Health and Nursing Research (IPP), PF 330440, D-28334 Bremen, Germany
Tel.: +90 242 227 45 37
E-mail: sengulakdeniz@akdeniz.edu.tr

Summary: The aim of this study is to conduct a bibliometric analysis of nursing research related to Artificial Intelligence (AI) in Nursing Care published in the Web of Science Core Collection (WoSCC) database from 1995 to 2025. A retrospective descriptive study was conducted in August 2025, analyzing 490 documents from 111 journals. Data were processed using the Biblioshiny package for RStudio, VOSviewer, and WoSCC metrics. The findings revealed that the United States holds a central position in global collaboration and publication output. The most productive journals identified were the Journal of Nursing Management, BMC Nursing, and CIN-Computers Informatics Nursing. Frequently used keywords such as "care," "artificial-intelligence," and "challenges" reflect the field's progression. This analysis demonstrates that nursing research has evolved from exploring technological possibilities to focusing on the safe and effective integration of AI into clinical practice. These results indicate a maturing field that increasingly prioritizes practical, in-depth approaches to technological implementation in healthcare.

Keywords: Artificial intelligence, Nursing care, Bibliometric analysis, Web of science, RStudio, Scientific mapping, Healthcare technology.

1. Introduction

The integration of artificial intelligence (AI) into nursing care represents a transformative shift in healthcare delivery. AI technologies, ranging from machine learning algorithms to natural language processing tools, are increasingly utilized to enhance clinical decision-making, automated administrative tasks, and optimize patient outcomes [1].

As the volume of scientific literature in this field grows, it becomes essential to map the research landscape to identify trends, key contributors, and future directions. Recent evidence suggests that while AI offers immense potential, its implementation requires a focus on ethical integration and the maintenance of human-centered care [2]. This study provides a comprehensive bibliometric analysis of the global research output on AI in nursing care over the past three decades [3]. Unlike previous bibliometric reviews that predominantly focus on general AI trends in medicine, this study offers a unique contribution by specifically mapping the structural evolution of AI through the lens of nursing care integration and professional workload. By utilizing a longitudinal dataset without chronological restrictions, our research identifies a critical paradigm shift: the transition from theoretical technological potential to practical, human-centered implementation challenges. Furthermore, this study uniquely links emerging digital trends with the specific ethical and security needs of the nursing workforce, providing a strategic roadmap that distinguishes it from broader healthcare AI analyses.

2. Methods

This study employed a retrospective descriptive design. Data collection and analysis were performed in August 2025 using a computer-based environment. The study adhered to the BIBLIO checklist for transparency and reproducibility [4].

2.1. Search Strategy and Criteria

The research data were obtained from the WoSCC database. The search strategy employed the following topic terms: (TS = "Artificial Intelligence" OR "artificial intelligence" OR "artificial-intelligence" OR "AI") AND (TS = "Nursing Care"). To ensure relevance and quality, the following inclusion criteria were applied: (1) articles or review papers published in English between 1995 and 2025, and (2) research specifically categorized under the 'Nursing' field within the WoSCC categories. Exclusion criteria included early access articles, books, book chapters, conference papers, editorials, and case studies. Following the screening process, a total of 490 documents were identified for analysis.

2.2. Data Analysis Tools

The Bibliometrix R-package (Biblioshiny) was used for performance analysis and science mapping. VOSviewer was utilized to visualize co-occurrence and collaboration networks.

2.3. Ethical Considerations

As this study has no direct impact on human subjects, ethical committee approval is not required.

3. Results

3.1. General Publication Trends

The dataset includes 490 articles published in 111 different journals, containing 15890 references. The field shows a robust annual growth rate of 19.49 %, with an average article age of 1.48 years, indicating a highly current and rapidly evolving body of literature (Fig. 1).

3.2. Leading Contributors

The United States is the most productive country (177 publications), maintaining strong collaboration networks with China (57) and Türkiye (43). Notably, Türkiye has recently emerged as a significant contributor to the global scientific landscape in this field. Columbia University (19 articles) and Duke University (16 articles) are the leading institutions in terms of both volume and citation impact (Fig. 2).

3.3. Publication Trends and Journals

The analysis identified the Journal of Nursing Management, BMC Nursing, and CIN-Computers Informatics Nursing as the most productive journals. There has been a significant upward trend in publication volume, particularly over the last decade (Fig. 3).

Fig. 1. Basic Information About Nursing Research Related to the Use of Artificial Intelligence in Nursing Care.

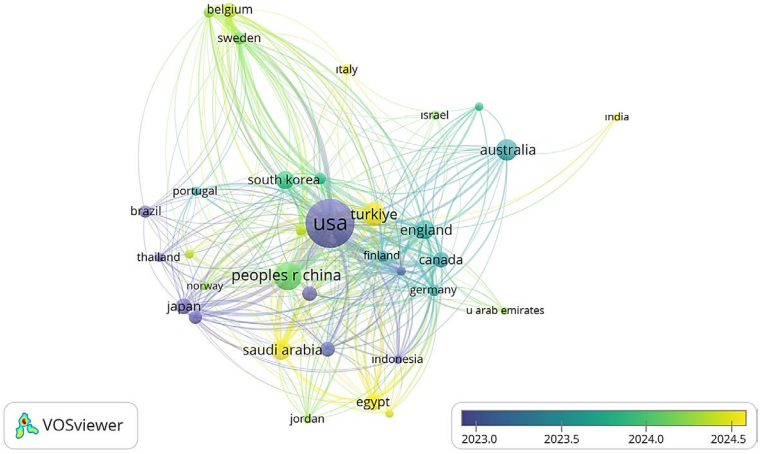


Fig. 2. Countries' Contributions to Publications.

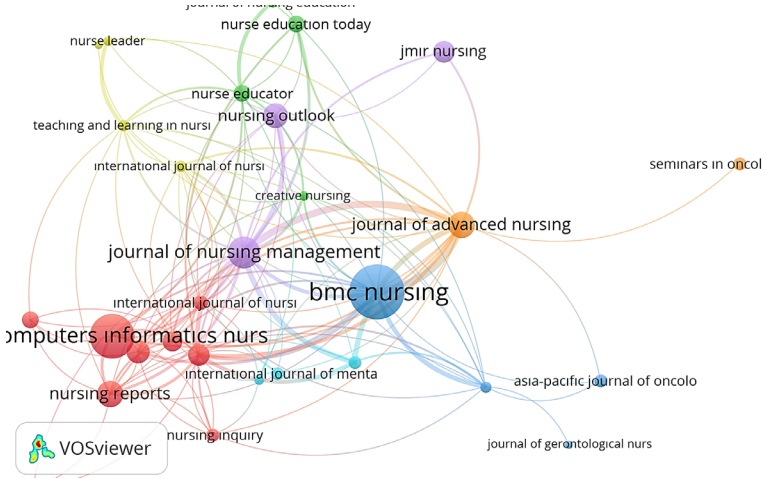


Fig. 3. Citation Analysis of the Most Productive Journals.

3.4. Conceptual Structure and Evolution

Keyword analysis shows that "care," "artificial intelligence," and "health" are the most frequent concepts. Chronological mapping reveals a shift in research focus: early studies (2019-2021) focused on general ethical and philosophical implications, whereas recent research (2023-2024) specifically examines generative AI tools like ChatGPT and their impact on nursing education and patient communication. The field is moving toward addressing practical challenges, such as "risk," "management," and "health equity" (Fig. 4).

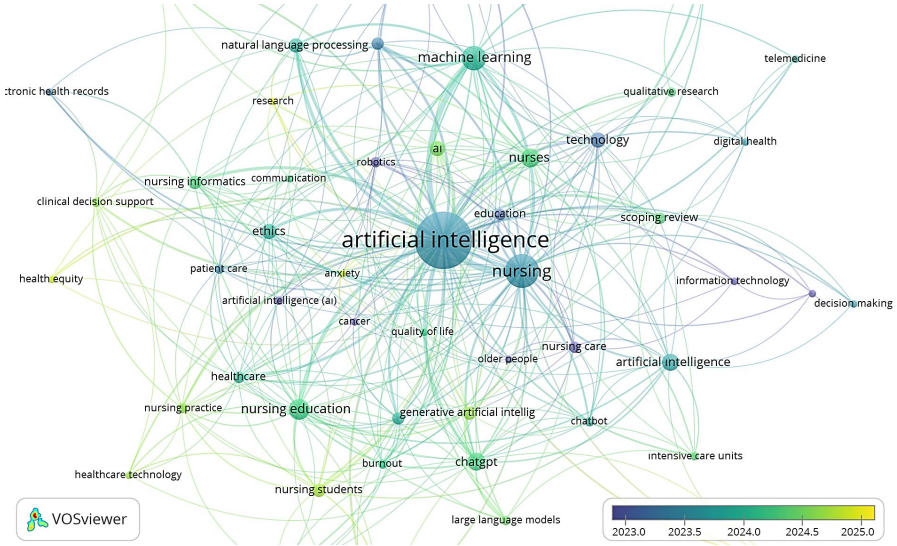


Fig. 4. Chronological Network of Author Keywords.

The words in the word cloud, such as 'care' and 'artificial intelligence', indicate the main themes of the text, the most emphasized concepts or key terms. These words usually form the focus of the text. Furthermore, related words are visually positioned close to each other. This may indicate that certain concepts are mentioned together or complement each other (Fig. 5).

Fig. 5. Word Cloud.

3.5. Most Influential Authors

Fig. 6 visualizes author relationships within the field under study. Instead of including all authors, this visualization focuses on authors who exceed a certain threshold of link strength. Using this method, instead of displaying a cluster of 46 authors with at least three publications each, the seventeen different authors who constitute the largest cluster of connected elements in the network are presented. In Fig. 6, 17 authors form 4 clusters with a total of 36 connections and 70 total connection strength. This methodological choice allows the analysis to identify and focus on the most influential and central authors. Authors with the strongest connections are grouped into four clusters, represented by red, blue, green, and yellow. Authors in the same color cluster share similar themes and indicate a co-authorship relationship. The closeness of the circles indicates a co-authorship relationship between authors, while the thickness of the connection lines reflects the strength of this relationship. Maxsim Topaz was identified as the author with the highest overall connection strength and has strong co-authorship ties. The thicker line between Maxim Topaz and J. Song indicates a stronger relationship between them. The network map in Fig. 6 reveals a co-authorship between Maxsim Topaz and Maria Peltonen, although they are located in different groups.

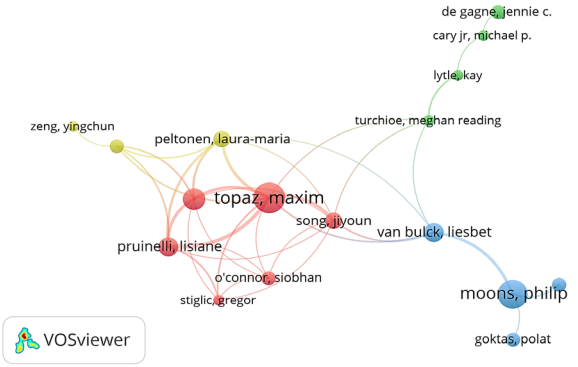


Fig. 6. Most Influential Authors.

3.6. Institutions' Contributions to Publications

According to the data examined, Columbia University stands out as the institution with the most publications, with 19 articles. It is followed by Duke University with 16 articles, the University of Minnesota with 12 articles, and Alexandria University with 11 articles. This ranking proves that these institutions are the centers that quantitatively contribute the most to the scientific literature in the field. In addition to the number of publications, the number of citations also reflects the impact of these institutions. Columbia University is the most influential institution in this field with 395 citations. The University of Minnesota ranks second with 346 citations. These high citation counts show that the work of these universities not only produces a large number of publications but is also considered valuable by the academic community and makes significant contributions to the accumulation of scientific knowledge. The University of Turku (341 citations) and the University of Toronto (337 citations) also have a global impact with their high citation counts. Moreover, Koç University in Turkey's contribution to this important field with 5 publications demonstrates that national academic efforts are finding their place on the international stage (Fig. 7).

The bibliometric indicators in Table 1 reveal a significant shift in the thematic focus of nursing literature.

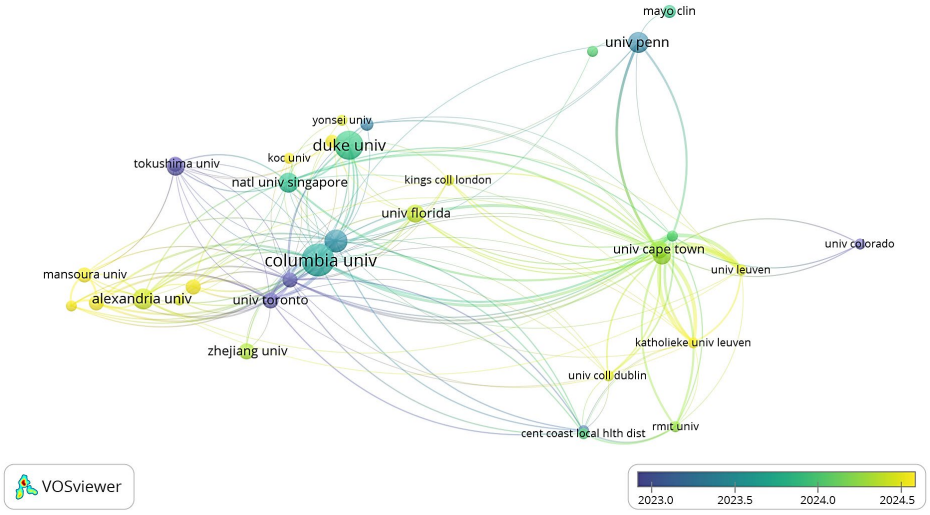


Fig. 7. Institutions' Contributions to Publications.

Table 1. Top 10 Most Influential Articles.

Rank	Article Title	TC	TC/ Year
1	Artificial intelligence in nursing: Priorities and opportunities...	146	29.20
2	Artificial Intelligence-based technologies in nursing: A scoping review	134	33.50
3	The ChatGPT Storm and What Faculty Can Do	118	39.33
4	Artificial intelligence in nursing and midwifery: A systematic review	92	30.67
5	Can nurses remain relevant in a technologically advanced future?	92	13.14
6	Ethics of Caring as a Guide to Dividing Tasks Between AI and Humans	82	13.67
7	What if your patient switches from Dr. Google to Dr. ChatGPT?	81	40.50
8	Nursing education in the age of AI-powered Chatbots (AI-Chatbots)	77	25.67
9	AI in virtual reality simulation for interprofessional communication	73	24.33
10	Using ChatGPT and Google Bard to improve patient information	70	35.00

TC: Total Citations

4. Discussion

The concentration of publications in high-impact nursing journals suggests that AI is no longer viewed merely as a technical novelty but as a strategic element of nursing administration and excellence. The bibliometric findings of this study, encompassing 490 articles with a robust annual growth rate of 19.49 %, signify that the field has transitioned into a period of rapid maturation. The dominance of the United States (177 publications) indicates a well-established infrastructure for digital health research, while the rising contribution from other regions such as China (57) and Turkey (43) reflects increasing global relevance and a shift toward multi-centered scientific production. This geographical expansion suggests that AI in nursing is being adapted to diverse healthcare systems, supported by national strategic investments in digital transformation [1, 4]. Recent integrative reviews [5] and scoping studies [3] provide strong evidence for this trend, illustrating that AI's role in nursing is evolving beyond simple administrative support; These studies highlight a shift from operational efficiency toward

profound clinical transformations, such as enhanced patient monitoring and predictive care.

The analysis reveals that Columbia University stands out as the most productive institution with 19 articles. This quantitative superiority suggests that the institution has established a primary research focus on artificial intelligence within nursing and has allocated significant resources to this domain. Furthermore, the presence of Duke University, the University of Minnesota, and Alexandria University among the top institutions confirms their roles as pivotal centers enriching the scientific literature.

Beyond mere publication volume, citation counts serve as a critical indicator of scientific impact. The finding that Columbia University is the most influential institution (395 citations), followed by the University of Minnesota (346 citations), demonstrates that their contributions are not only quantitatively significant but also profoundly shape the field's knowledge base. These high citation rates indicate that their research is widely accepted by the academic community and provides a foundational framework for future studies.

Another significant finding is the geographical distribution of publications. Although the majority of leading institutions are based in the United States, the overall landscape reflects a diverse global scope. This emphasizes that the integration of AI into nursing care is not confined to a specific region or culture; Rather, it holds universal importance for healthcare systems worldwide. This diversity suggests that research in this area is increasingly enhanced by international collaborations. Such global participation carries great potential for adapting AI applications to varied healthcare environments and developing culturally sensitive solutions, which will be essential for the inclusive evolution of the field.

The evolution of keywords signifies a shift toward implementation science, specifically addressing the complexities of integrating tools like ChatGPT into nursing education and clinical workflows [6]. While early seminal works (2019-2021) focused on the ethical, philosophical, and visionary dimensions of AI [1, 7], the current landscape (2023-2024) is dominated by the practical implications of generative AI. Recent research highlights that tools like ChatGPT are redefining patient communication shifting the paradigm from 'Dr. Google 'to 'Dr. ChatGPT' and raising new questions about the trustworthiness and value of AI-generated health information [8, 9]. This rapid technological shift underscores the urgency for a fundamental revision of nursing curricula; a need that has evolved from early calls for general digital literacy [10] to the current necessity of managing generative AI in clinical practice, as advocated by professional organizations like the American Association of Colleges of Nursing (AACN).

While AI offers immense potential for optimizing nursing workload, the literature emphasizes specific functional applications: for instance, AI-driven speech assistants [11] address the critical administrative burden where documentation can occupy up to 30 % of a nurse's shift by enabling real-time, bedside data entry. Our findings regarding authorship trends, where 97.4 % of authors publish as co-authors, further prove that these complex technological integrations require a collective, interdisciplinary effort rather than isolated expertise. Similarly, predictive staffing models [12] enhance resource allocation by forecasting patient care demands. Research indicates that nursing professionals face multifaceted challenges regarding technological integration, necessitating a deeper understanding of user needs, system usability, and data security [13].

A detailed evaluation of author impact highlights the central roles of specific researchers. Topaz M. emerges as the most influential figure with a high Total Link Strength, an h-index of 6, and 405 citations, confirming his leadership in AI-supported nursing decision-making [14]. Furthermore, the emergence of researchers like Moons P. and Van Bulck L., who achieved high impact scores in a very short time, along with the high m-index values of Agaoglu FO and Almagharbeh WT, demonstrates a dynamic and competitive research environment. These metrics suggest that the field is not only led by established scholars but is also being rapidly reshaped by young researchers focusing on niche, groundbreaking topics.

Furthermore, the introduction of social robots and intelligent emergency systems in elderly care introduces critical dimensions of user acceptance and safety [15]. As the literature matures, it is clear that successful AI integration requires not only technical accuracy, such as in automated wound identification [16], but also a rigorous focus on ethical considerations to maintain the huma n-centered essence of nursing [17].

The analysis of the most influential articles (Table 1) provides a comprehensive roadmap of how artificial intelligence has evolved within nursing scholarship. The citation metrics reveal a distinct transition from foundational, visionary frameworks to immediate, practical explorations of generative AI technologies.

The study by Ronquillo et al. (2021), which holds the highest total number of citations (146), serves as a foundational pillar for the field. This work established the "invitational think-tank" priorities, framing AI as a strategic necessity for nursing leadership [1]. Similarly, von Gerich et al. (2022) provided a crucial scoping review that summarized the evidence base prior to the generative AI explosion [2]. These works represent a period where the nursing community was primarily concerned with defining the scope, ethics, and overarching potential of AI [7].

The most striking finding in this analysis is the emergence of "citation velocity" as a key indicator of current trends. Although foundational papers have higher total citations, the highest annual impact is observed in 2023 and 2024 publications. Van Bulck and Moons (2024) achieved an unprecedented citation average of 40.50 per year [9], followed closely by Sun and Hoelscher (2023) with 39.33 [8]. This "citation burst" reflects what has been termed the "ChatGPT

Storm," where the accessibility of Large Language Models (LLMs) has forced an immediate re-evaluation of nursing education and patient communication.

The prominence of Van Bulck and Moons (2024) [9] and Moons and Van Bulck (2024) [6] in the top 10 list highlights a critical shift in the patient-provider dynamic. The transition from "Dr. Google" to "Dr. ChatGPT" poses new challenges for health literacy and the trustworthiness of information a concern that was less prevalent in earlier AI literature. Furthermore, the high impact of studies focused on nursing education [8, 18] suggests that the academic sector is currently the most proactive in responding to AI advancements, driven by the need to maintain academic integrity and prepare students for an AI-augmented workforce.

In summary, the transition from the ethical and philosophical inquiries of Stokes and Palmer (2020) to the practical, simulation-based, and chatbot-focused studies of 2023–2024 demonstrates a maturing field [19]. Nursing research is no longer merely asking if AI should be used, but is now rigorously investigating how it is currently redefining the boundaries of care, documentation, and education.

Strengths and Limitations of the Study

This study is a pioneering bibliometric analysis that systematically examines the publication characteristics and structural development of nursing research on artificial intelligence (AI). A key strength is the absence of specific date restrictions, allowing for a holistic assessment of the field's evolution from its inception to the present. By capturing the longitudinal trajectory of the literature, this work provides a solid foundational reference for researchers and practitioners.

Despite these strengths, several limitations must be acknowledged. First, the study relied solely on the WoSCC. While WoSCC is a premier database for high-impact scholarly literature, integrating other databases like Scopus or PubMed could provide additional coverage, particularly regarding clinical nursing informatics and interdisciplinary health sciences. Second, the study was limited to English-language publications, which may introduce language bias and exclude significant regional research published in other languages.

Finally, the search strategy is primarily focused on "Artificial Intelligence" as an overarching framework within the "Nursing" category. Although AI acts as a comprehensive umbrella term in academic indexing, this approach may not fully capture niche studies that are exclusively indexed under specific technical sub-terms such as "machine learning," "deep learning," or "robotics" without the broader "AI" descriptor. While this ensures high thematic precision regarding the integration of AI into nursing care, future studies could employ a more granular, multi-term Boolean string to provide a comparative analysis of these specific technological sub-sectors.

5. Conclusions

This bibliometric analysis demonstrates that nursing research on AI has matured from asking "What can technology do?" to "How can we integrate it most safely and effectively?" While the US remains the leader, global participation is increasing, leading to a more inclusive and multi-centered research environment. Future research should continue to focus on ethical integration, data security, and the impact of AI on nursing students and patient outcomes.

References

[1]. C. E. Ronquillo, L. M. Peltonen, L. Pruinelli, et al., Artificial intelligence in nursing: Priorities and opportunities from an international invitational think-tank of the Nursing and Artificial Intelligence Leadership Collaborative, *Journal of Advanced Nursing*, Vol. 77, Issue 9, 2021, pp. 3707-3717.

[2]. H. von Gerich, H. Moen, L. J. Block, et al., Artificial intelligence-based technologies in nursing: A scoping literature review of the evidence, *International Journal of Nursing Studies*, Vol. 127, 2022, 104153.

[3]. S. O'Connor, Y. Yan, F. J. Thilo, et al., Artificial intelligence in nursing and midwifery: A systematic review, *Journal of Clinical Nursing*, Vol. 32, Issue 13-14, 2023, pp. 2951-2968.

[4]. A. Montazeri, S. Mohammadi, P. M. Hesari, et al., Preliminary guideline for reporting bibliometric reviews of the biomedical literature (BIBLIO): A minimum requirements, *Systematic Reviews*, Vol. 12, Issue 1, 2023, 239.

[5]. S. Hassanein, R. A. El Arab, A. Abdrbo, et al., Artificial intelligence in nursing: An integrative review of clinical and operational impacts, *Frontiers in Digital Health*, Vol. 7, 2025, 1552372.

[6]. P. Moons, L. Van Bulck, Using ChatGPT and Google Bard to improve the readability of written patient information: A proof of concept, *European Journal of Cardiovascular Nursing*, Vol. 23, Issue 2, 2024, pp. 122-126.

[7]. J. A. Pepito, R. Locsin, Can nurses remain relevant in a technologically advanced future?, *International Journal of Nursing Sciences*, Vol. 6, Issue 1, 2019, pp. 106-110.

[8]. G. H. Sun, S. H. Hoelscher, The ChatGPT storm and what faculty can do, *Nurse Educator*, Vol. 48, Issue 3, 2023, pp. 119-124.

[9]. L. Van Bulck, P. Moons, What if your patient switches from Dr. Google to Dr. ChatGPT? A vignette-based survey of the trustworthiness, value, and danger of ChatGPT-generated responses to health questions, *European Journal of Cardiovascular Nursing*, Vol. 23, Issue 1, 2024, pp. 95-98.

[10]. T. Risling, Educating the nurses of 2025: Technology trends of the next decade, *Nurse Education in Practice*, Vol. 22, 2017, pp. 89-92.

[11]. K. Schwabe, D. Ferizaj, S. Neumann, Reducing nurses' workload with an AI speech assistant for documentation, in *Proceedings of the 1st International Conference on AI in Medicine and Healthcare (AiMH)*, 2025, pp. 119-122.

[12]. D. Hunstein, M. Fiebig, Personnel management with AI: Predicting the nursing workload, *Studies in Health Technology and Informatics*, Vol. 315, 2024, pp. 231-235.

[13]. A. Lukas, I. Maucher, S. Bugler, et al., Security and user acceptance of an intelligent home emergency call system for older people living at home, *Zeitschrift für Gerontologie und Geriatrie*, Vol. 54, Issue 7, 2021, pp. 685-694 (in German).

[14]. M. Hu, Y. Wang, Y. Liu, et al., Artificial intelligence in nursing decision-making: A bibliometric analysis of trends and impacts, *Nursing Reports*, Vol. 15, Issue 6, 2025, 198.

[15]. K. Seibert, D. Domhoff, D. Fürstenau, et al., Exploring needs and challenges for AI in nursing care results of an explorative sequential mixed methods study, *BMC Digital Health*, Vol. 1, Issue 1, 2023, 13.

[16]. L. Hung, Y. Zhao, H. Alfares, et al., Ethical considerations in the use of social robots for supporting mental health and wellbeing in older adults in long-term care, *Frontiers in Robotics and AI*, Vol. 12, 2025, 1560214.

[17]. K. Majjouti, M. Tapp-Herrenbrueck, et al., KIADEKU: Identification of wound types with AI, in *Proceedings of the 1st International Conference on AI in Medicine and Healthcare (AiMH)*, 2025, pp. 114-118.

[18]. W. Tam, T. Huynh, et al., Nursing education in the age of artificial intelligence powered chatbots (AI-chatbots): Are we ready yet?, *Nurse Education Today*, Vol. 129, 2023, 105917.

[19]. F. Stokes, A. Palmer, Artificial intelligence and robotics in nursing: Ethics of caring as a guide to dividing tasks between AI and humans, *Nursing Philosophy*, Vol. 21, Issue 4, 2020, e12306.

(034)

Dependence on Fragile AI Systems: Rethinking the Collingridge Dilemma in Clinical AI

A. Gerdes
University of Southern Denmark, Department of Design, Media and Educational Science,
Universitetsparken 1, Kolding, Denmark
Tel.: + 0045 65501323
E-mail: gerdes@sdu.dk

Summary: We may have created technology that grows more fragile as it evolves. Consequently, the healthcare sector risks a form of technological regression, marked by continued reliance on fragile, no longer fully functional tools. This introduces a fragility–entrenchment dilemma: when human expertise is strong, AI's technical challenges are manageable. However, as clinicians increasingly rely on AI, professional judgment and clinical skills come under pressure. Simultaneously, the structural fragility of AI persists and even grows. Hence, the concern is not merely that AI systems are fragile, but that their fragility may leave healthcare worse off than before adoption.

Keywords: Clinical judgment, LLM, Model collapse, Hidden technical debt, Cognitive debt, Beyond the Collingridge dilemma.

1. Introduction

Society increasingly integrates AI systems into critical fields such as healthcare. However, AI systems are inherently fragile and may not exhibit the same pattern of technological stabilization that has typically characterized most engineered systems. Consequently, AI in healthcare introduces a competence problem due to a fragility–entrenchment dilemma. When human expertise is strong, the technical challenges posed by AI remain manageable. However, as clinicians increasingly rely on AI – outsourcing trivial tasks to AI, seeking clinical advice from large language models (LLMs), and depending on AI-assisted decision-making tools – their clinical competencies and professional judgment risk eroding, yet at the same time, the fragility of the technology is intensified, resulting in a dilemma meaning that we increasingly depend on systems whose built-in structural fragility may not diminish and even grow over time.

In what follows, the paper presents selected examples of structural fragility outlining a fragility–entrenchment dilemma inherent in complex AI systems. The concern is not merely that AI systems are fragile, but that their fragility may leave society, institutions, businesses, and the healthcare sector worse off than before the AI adoption.

1.1. Structural Fragility: Climate Crisis

Many digital technologies – particularly AI and language models – have high environmental costs. Developing and operating AI consumes large amounts of electricity and water.

Research is underway to develop energy-efficient algorithms and computer chips that could make AI and data center operations more sustainable. However, it will likely be a long time before such solutions are widely implemented. Meanwhile, the development and unrestricted use of AI and LLMs show no signs of slowing. As a result, data centers' electricity and water consumption are likely to rise dramatically in the coming decades.

A conservative estimate from 2023 suggests that globally, AI will consume 85–134 TWh of electricity in 2027 [1]. Meanwhile, the U.S. Department of Energy's 2024 report on data center energy use estimates that in the United States alone, AI server electricity consumption may rise to 325-580 TWh by 2028, or 6.7 % to 12 % of total U.S. electricity consumption [2].

The vast electricity consumption also drives enormous demand for water to cool servers. Water is drawn from groundwater or surface sources. Since water is a limited resource that we all rely on, AI's enormous thirst is, to say the least, concerning.

When discussing water abstraction, both the water that evaporates and the water returned to the source are counted. It is estimated that in 2027, AI servers' global water demand will require the abstraction of 4.2–6.6 billion cubic meters of water annually, even accounting for the portion returned to the sources [3]. For context, in 2024, Denmark's total water abstraction was 814.96 million cubic meters [4].

The actual global water consumption of data centers – that is, the water evaporated during cooling – is estimated at 0.38–0.6 billion cubic meters annually in 2027 [3]. However, this is a conservative estimate; the U.S. Department of Energy's 2024 report [2] predicts even higher water use in the U.S. alone by 2028.

These numbers and their variability illustrate how complex it is for energy agencies to project resource needs. The U.S. naturally accounts for a large share of global consumption because it hosts the most data

centers. Yet AI's insatiable resource demands remain deeply concerning, as AI development and deployment continue to grow. AI companies promise that AI can solve the climate crisis, even as AI itself contributes heavily to that crisis [5].

The preferred method for developing AI, reinforcement learning, is particularly energy-intensive. Additionally, the painstaking fine-tuning of language models – often delivering only marginal improvements [6] – has a substantial climate footprint. Whether artificial intelligence can help solve the climate crisis before it is too late remains an open question.

1.2. Structural Fragility: Model Collapse in LLMs

Model collapse follows when internet data is polluted with AI-generated content that unintentionally ends up in training datasets used to generate future LLMs. Originally, ChatGPT was trained predominantly on human-generated internet data, although some synthetic data from bots and similar sources likely slipped in. In the future, however, much of the internet's content will be AI-generated. This pollutes the internet's data ecosystem and gives rise to an inbreeding problem.

Because LLMs generate content based on probabilities, they tend to reproduce what is most frequent and overlook or forget rarer occurrences. As a result, synthetic training data reflects a probability distribution that underrepresents rare occurrences ("outliers"), meaning that variation and diversity are not fully preserved. In diagnostic applications, this could lead LLMs to overlook signals associated with rare diseases.

When future generations of LLMs are trained on such synthetic data, these distortions may be amplified, contributing to model collapse. Moreover, the increasing use and reuse of synthetic data in medical AI to enhance and scale training data or preserve privacy (by using synthetic data as a proxy for real patient records) comes at a cost. As [7] warns, such data may lead to "unwarranted confidence in models trained on artificially generated datasets that fail to preserve clinical validity or demographic realities". Consequently, the authors call for tools to enhance data integrity and emphasize the need to ensure the integration of real data during training to prevent model collapse, which occurs when a model produces progressively narrower results and ultimately becomes completely unintelligible [8].

Model collapse is not only a risk when LLMs are fine-tuned for medical applications, as discussed above. From an overall perspective, the internet could become so polluted with synthetic AI-generated data that it may no longer be possible to produce language models.

OpenAI has a business advantage because GPT was created primarily from authentic human-generated internet data. OpenAI also has a large user base for ChatGPT, which could enable the company to identify and use authentic user interactions as training data. However, this does not solve the fundamental problem of internet-wide data pollution. Currently, there are no obvious solutions beyond ensuring the availability of high-quality human-generated data – precisely what is being done in the Danish Language Model Consortium [9]. Here, open Danish language models are based on high-quality Danish data and respect copyright. These models are not large-scale language models, but even smaller models tailored according to ethical principles and adapted to a specific cultural context can achieve significant results.

However, model collapse remains a serious threat to the future development of LLMs. [10] characterizes this challenge with reference to the Kessler Syndrome, which describes how space debris has polluted Earth's orbit to the point that future space travel may become impossible.

1.3. Structural Fragility: Hidden Technical Debt in Machine Learning Systems and Programmers' Cognitive Debt

Technical debt is generally considered manageable and acceptable in traditional IT systems, as long as it can be identified, controlled, and serviced. In contrast, AI systems consist not only of a relatively small amount of code but also of (sometimes dynamically changing) data embedded within highly complex infrastructures. Under these conditions, *hidden* technical debt can be difficult to detect, as noted in [11], it is "inappropriately easy to build large data dependency chains that can be difficult to untangle". Similarly, analysis debt may arise, since machine learning models can influence their own performance over time during updates, subtle changes that are not easily observable. Moreover, the field of maintainable AI remains underdeveloped. While [11] provides additional examples, the cases discussed here are sufficient to illustrate the substantial resources required to deploy AI effectively, pointing to a problem that will grow as programmers gradually lose competencies due to the outsourcing of programming to AI agents. Within computer science, professional judgment may also erode as coding skills become outsourced. Programming is not just about coding; programming is theory building, as stressed by Naur in 1985 [12]: "(…) the primary aim of programming is to have the programmers build a theory of the way the matters at hand may be supported by the execution of a program. Such a view leads to a notion of program life that depends on the continued support of the program by programmers having its theory. Further, on this view the notion of a programming method, understood as a set of rules of procedure to be followed by the programmer, is based on invalid assumptions and so has to be rejected" [12].

If future system developers cannot understand, contest, or evaluate the systems or no longer have an overarching theory behind the systems they built, this could have far-reaching consequences for an AI-fueled healthcare sector.

2. Beyond the Collingridge Dilemma: The Fragility–Entrenchment Dilemma

In raising a problem of technology control, Collingridge's dilemma states that it is not possible to anticipate the consequences of technology in its early phases, and once we see the negative consequences, "the technology is often so much a part of the whole economic and social fabric that its control is extremely difficult" [13]. Collingridge's dilemma operates across three intersecting dimensions: epistemological, political, and temporal. Consequently, we must have knowledge to act, but without power, we lack impact. Furthermore, we need to know *when* to control.

Collingridge's dilemma is widely discussed in the literature on early ethics assessment and responsible innovation, and within the field of AI in healthcare. As an example, [14] suggests applying the medico-legal method's criteria to Collingridge's dilemma. Consequently, [14] identifies proactive and reactive strategies to prevent the negative effects of AI in healthcare. As a proactive strategy, the authors highlight health technology assessments, such as the MAS-AI model for assessing the value of AI in medical imaging, which provide opportunities for informed decision-making [15]. As a reactive strategy, the authors emphasize regulating the innovation process through responsible AI development to address, e.g., bias-related problems, and here they point to frameworks and tools for datasets and models, e.g., the CLAIM guideline for AI in medical imaging.

Yet the traditional framing of Collingridge's dilemma, exemplified with [14], remains tied to the dilemma's original temporal logic, i.e., the question of *when* to intervene. Whether proactive or reactive, both strategies assume that the core challenge is one of timing: acting early enough to shape the technology before it becomes entrenched, or building responsible development practices into the innovation process itself to prevent harm from arising.

Alternatively, this paper argues that clinical AI poses a structurally different problem, one that the framing of Collingridge's dilemma does not fully capture. The concern is not only that it is hard to proactively identify consequences before a technology is embedded, and that later, when trouble arises, it becomes hard to address them because the technology is entrenched in society. Rather, the concern is that the very process of embedding AI erodes professional judgment, while the technology itself remains structurally fragile. This is the fragility–entrenchment dilemma, and it introduces a competence problem: the healthcare sector may become dependent on AI technologies that remain structurally fragile, whilst original practices and professional judgment are weakened.

This contrasts with the traditional governance issue identified in the Collingridge dilemma. In many technological fields, systems typically become stable over time through engineering improvements, regulation, and accumulated operational experience. This paper argues that modern AI systems might behave in the opposite way due to phenomena such as unacceptable environmental costs, AI model collapse, and technical and cognitive debt. Against this backdrop, we ought to be concerned as systems become increasingly embedded in clinical practice, even though their underlying technological fragility may persist or intensify rather than diminish over time.

3. Concluding Remarks

Expertise and professional judgment are traditionally cultivated through theoretical knowledge and practice-based experience. Thus, expertise develops through repeated engagement with clinical cases, a point also acknowledged in medical training and quality assurance frameworks [16]. As clinical tasks are increasingly delegated to AI, the clinical practices and professional judgment may gradually erode. However, as AI becomes more entrenched, its structural fragility does not diminish; instead, it may intensify. Thus, the process of embedding AI in clinical practice is doubly destabilizing: it weakens the professional judgment needed to deliver holistic clinical care, while the technology itself, far from stabilizing through use, may become more fragile over time. Consequently, if structurally fragile AI systems fail, the healthcare sector may thus face a form of technological regression: dependence on AI tools that have ceased to function, combined with diminished capacity to manage without them.

References

[1]. A. de Vries, The growing energy footprint of artificial intelligence, *Joule*, Vol. 7, Issue 10, 2023, pp. 2191-2194.

[2]. A. Shehabi, et al., United States data center energy usage report 2024, Technical Report, *Lawrence Berkeley National Laboratory*, Berkeley, 2024.

[3]. P. Li, J. Yang, M. A. Islam, S. Ren, Making AI less 'thirsty', *Communications of the ACM*, Vol. 68, Issue 7, 2025, pp. 54-61.

[4]. Danmarks Statistik, Vand og spildevand, 2026, https://www.dst.dk/da/Statistik/emner/miljoe-og-energi/groent-nationalregnskab/vand-og-spildevand (in Danish).

[5]. K. Joshi, The AI climate hoax: Behind the curtain of how big tech greenwashes impacts, Beyond Fossil Fuels, Berlin, 2026, https://beyondfossilfuels.org/wp-content/uploads/2026/02/AI-for-climate-claims-Report_FEB-2026_FINAL-2-16.pdf

[6]. E. M. Bender, T. Gebru, A. McMillan-Major, S. Shmitchell, On the dangers of stochastic parrots: Can language models be too big?, in *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2021, pp. 610-623.

[7]. A. Koul, D. Duran, T. Hernandez-Boussard, Synthetic data, synthetic trust: Navigating data challenges in the digital revolution, *The Lancet Digital Health*, Vol. 7, Issue 11, 2025, 100924.

[8]. I. Shumailov, Z. Shumaylov, Y. Zhao, N. Papernot, et al., AI models collapse when trained on recursively generated data, *Nature*, Vol. 631, 2024, pp. 755-759.

[9]. D. L. M. Consortium, Development of Danish open language models: Ethical principles and quality data, Dansk Sprogmodelkonsortium, https://sprogmodel.dk

[10]. A. D. Lautrup, Generation and evaluation of realistic tabular synthetic data, PhD Thesis, *Syddansk Universitet*, Odense, 2025.

[11]. D. Sculley, et al., Hidden technical debt in machine learning systems, in Advances in Neural Information Processing Systems 28 (C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, et al., Eds.), *Curran Associates, Inc*., Red Hook, 2015, pp. 2503-2511.

[12]. P. Naur, Programming as theory building, *Microprocessing and Microprogramming*, Vol. 15, Issue 5, 1985, pp. 253-261.

[13]. D. Collingridge, The Social Control of Technology (Reprint Ed.), *Pinter*, London, 1982.

[14]. R. Cecchi, T. M. Haja, F. Calabrò, I. Fasterholdt, et al., Artificial intelligence in healthcare: Why not apply the medico-legal method starting with the Collingridge dilemma?, *International Journal of Legal Medicine*, Vol. 138, Issue 3, 2024, pp. 1173-1178.

[15]. I. Fasterholdt, et al., Model for ASsessing the value of Artificial Intelligence in medical imaging (MAS-AI), *International Journal of Technology Assessment in Health Care*, Vol. 38, Issue 1, 2022, e74.

[16]. European Commission, European Guidelines for Quality Assurance in Breast Cancer Screening and Diagnosis (4th Ed., supplements), *Publications Office of the European Union*, Luxembourg, 2013.

(035)

Policy-Focused Evaluation of Individualized Vasopressor Effects with Robustness to Irrelevant Covariates in MIMIC-III ICU Data

A. S. Khan [1], E. Schaffernicht [2] and J. A. Stork [1]
[1] School of Science and Technology, Örebro University, Sweden
[2] Technical University of Applied Sciences Würzburg-Schweinfurt, Germany
E-mail: {ahmad-saeed.khan, johannesandreas.stork}@oru.se, erik.schaffernicht@thws.de

Summary: Estimating individualized treatment effects (ITEs) from observational intensive care unit (ICU) data is challenging because treatment assignment is strongly confounded by patient severity, overlap between treatment groups may be limited, and high-dimensional clinical covariates often contain irrelevant measurements that degrade treatment-effect models. We study individualized vasopressor effects using the MIMIC-III ICU database, where treatment is defined as vasopressor administration within the first 24 hours of ICU admission. The outcome $y_i$is a clinically motivated lactate-change endpoint defined as the difference between mean serum lactate measured in a late post-admission window (60–84h) and a baseline window around ICU admission (−6h to +6h), where negative values indicate improvement. Because counterfactual outcomes are unobserved, we adopt a policy-focused evaluation protocol combining model-implied average treatment effect (ATE) estimation with bootstrap uncertainty, plug-in policy-risk evaluation under a treat-if-benefit rule $\pi_f(x)$, heterogeneity diagnostics, subgroup analysis across baseline lactate strata, and robustness analysis under controlled irrelevant-covariate inflation. We compare DRI-ITE, which learns disentangled latent factors corresponding to treatment-predictive (Γ), confounding (Δ), adjustment (Υ), and irrelevant (Ω) variation, against standard meta-learners (T-, S-, and X-learners), a causal forest, and disentanglement-based baselines (DRCFR and RLOCFR). While model-implied ATE estimates are broadly similar across methods, policy evaluation reveals substantial differences in decision quality. DRI-ITE achieves the largest estimated improvement over a treat-all policy ($\Delta_{\text{all}} = -0.132 \pm 0.009$) while maintaining a balanced treatment rate, whereas several baselines collapse toward near-universal treatment with negligible policy gain. Under controlled irrelevant-feature inflation via appended Gaussian covariates ($|\ \Omega\ | \in \{5,10,15,20\}$), DRI-ITE consistently retains the strongest policy improvement and stable treatment behavior, while causal forest, S-learners, and weaker disentanglement-based baselines deteriorate toward more aggressive or less informative policies. These results demonstrate that evaluating ITE models through decision-focused criteria reveals failure modes that are not apparent from average-effect estimates alone, and that explicitly separating irrelevant covariate variation leads to more robust and clinically meaningful treatment policies in real-world observational settings.

Keywords: Individualized treatment effects, ICU, MIMIC-III, Vasopressors, Causal inference, Policy learning.

1. Introduction

Vasopressors are a cornerstone of haemodynamic resuscitation in the intensive care unit (ICU): they increase arterial pressure and help maintain end-organ perfusion in critically ill patients. However, their benefit is highly heterogeneous. In some patients with severe vasodilatory shock, early vasopressor administration is essential for stabilizing circulation, whereas in others aggressive vasopressor use may be unnecessary or even harmful. Average treatment effect (ATE) analyses therefore provide limited guidance for clinical decision-making. Instead, clinicians need models that can estimate *individualized treatment effects* (ITEs) and identify which patients are most likely to benefit from treatment.

ITE estimation from observational data has been widely studied in causal machine learning, building on the potential outcomes framework [1, 2] representation learning approaches [3] and doubly robust estimation techniques [4, 5]. Applying these methods to ICU data introduces several challenges.

Confounding

Treatment assignment in clinical practice depends strongly on patient covariates such as hemodynamic instability, laboratory measurements, and clinician judgment. Consequently, the propensity $P(T = 1 \mid X)$ varies substantially across the covariate space, introducing selection bias relative to a randomized treatment setting.

Limited overlap

Vasopressors are typically administered to severely ill patients and less frequently to stable ones. Regions of the covariate space with near-zero or near-one treatment propensity can lead to unstable estimation, especially for methods relying on propensity-based reweighting or adjustment.

High-dimensional nuisance covariates

ICU datasets contain many physiological measurements and laboratory variables, some of which may be weakly related or irrelevant to treatment assignment and outcome prediction. Such nuisance dimensions can inflate estimator variance, increase overfitting, and induce overly aggressive or degenerate treatment policies.

Existing empirical evaluations of ITE methods on clinical data often emphasize average-effect recovery or proxy prediction criteria that do not directly reflect the quality of individualized treatment decisions. In practice, however, the central question is whether a model can induce a useful treatment policy. Because counterfactual outcomes are not observed in observational data, the true value of a policy cannot be directly evaluated. Instead, we adopt a *policy-focused*

evaluation strategy based on model-implied treatment rules, where policy value is estimated using predicted counterfactual outcomes. These quantities should therefore be interpreted as plug-in diagnostics that reflect the internal consistency and decision behavior of a model, rather than unbiased estimates of real-world clinical utility.

In this work, we study individualized vasopressor treatment effects using the real-world MIMIC-III intensive care unit database [6]. We construct a cohort of adult ICU patients, define treatment as vasopressor administration within the first 24 hours of ICU admission, and use a clinically motivated lactate-change endpoint comparing baseline lactate measurements around admission to mean lactate levels measured 60–84 hours later. Using this cohort, we evaluate heterogeneous treatment effect estimators under several complementary criteria.

Our analysis focuses on the empirical behaviour of DRI-ITE [7], a disentangled representation learning framework for ITE estimation that separates treatment-predictive, confounding, outcome-adjustment, and irrelevant covariate variation. We compare DRI-ITE not only against widely used meta-learners (T-, S-, and X-learners) and causal forests, but also against related disentanglement-based baselines.

While a wide range of ITE estimation methods have been proposed, we focus in particular on DRI-ITE because its design explicitly targets a key challenge in observational ICU data: the presence of high-dimensional covariates that include both clinically relevant and irrelevant information. In contrast to standard meta-learners and tree-based approaches, DRI-ITE separates covariate information into treatment-predictive, confounding, outcome-related, and nuisance components through a disentangled representation. This structure is particularly well suited to clinical settings, where many recorded variables may be weakly related or irrelevant to the treatment decision and outcome, and can therefore degrade model reliability and lead to unstable or overly aggressive treatment policies. The MIMIC-III ICU database provides a realistic and challenging setting for evaluating such models. Treatment assignment is strongly confounded by patient severity, overlap between treated and untreated patients is limited, and the covariate space includes a wide range of physiological measurements and laboratory variables with varying relevance. These characteristics make MIMIC-III an appropriate testbed for studying whether ITE models can produce stable and clinically meaningful treatment policies under realistic data conditions.

Our study makes four contributions:

- We provide a reproducible real-world evaluation of individualized vasopressor treatment effects using the MIMIC-III ICU database with a clinically motivated outcome based on lactate dynamics.
- We introduce a structured, policy-focused evaluation protocol for individualized treatment effect models in observational clinical data. The protocol integrates model-implied ATE estimation, plug-in policy evaluation, subgroup analysis, and robustness assessment under controlled irrelevant-covariate inflation, enabling a more comprehensive assessment of decision quality beyond standard predictive metrics.
- We demonstrate empirically that DRI-ITE produces more stable and informative treatment policies than standard meta-learners, causal forests, and related disentanglement-based baselines, particularly under high-dimensional nuisance variation, highlighting the importance of separating irrelevant covariates in real-world settings.
- We provide an interpretable analysis of the learned representations by examining how different groups of covariates contribute to treatment prediction and outcome estimation. This analysis shows that the model separates clinically meaningful patient characteristics from less relevant variation, supporting the interpretability of the learned structure.

Taken together, these results highlight the importance of explicitly accounting for irrelevant covariate structure when estimating individualized treatment effects from high-dimensional observational ICU data.

2. Related Work

Heterogeneous treatment effects from observational data

Early approaches to heterogeneous treatment effect estimation relied on matching and propensity score adjustments within the potential outcomes framework [1, 8]. More recently, machine learning methods have been developed to estimate individualized treatment effects from high-dimensional data. Meta-learners such as the S-, T-, and X-learners [9] provide flexible plug-in estimators that can be implemented with arbitrary supervised learning models. Tree-based approaches such as causal trees and causal forests [10], [11] estimate local treatment effects using adaptive partitioning with asymptotic guarantees. Representation-learning methods aim to reduce confounding by learning balanced latent representations prior to outcome prediction, including TARNet and CFRNet [3, 12], variational models such as CEVAE [13], and architectures such as DragonNet [14]. While these approaches achieve strong performance on semi-synthetic benchmarks, their behavior in real-world clinical settings with high-dimensional nuisance variation remains less well understood.

Disentangled representations for causal inference

Recent work has explored disentangled representation learning as a strategy for improving treatment effect estimation. These approaches aim to separate covariate information related to treatment assignment from information predictive of outcomes

in order to better control confounding [3, 15, 16]. By structuring latent representations into interpretable components, such models seek to isolate confounding variation and improve counterfactual prediction. However, empirical evaluations of disentangled causal representations have largely focused on benchmark datasets or proxy metrics, leaving open questions about their practical behavior under real-world data conditions. In this work, we study the empirical behavior of DRI-ITE [7], which explicitly factorizes treatment-predictive, confounding, outcome-related, and irrelevant covariate variation, and evaluate its performance under clinically motivated and robustness-focused criteria.

Causal inference with electronic health records

Causal machine learning has increasingly been applied to electronic health records and ICU datasets. Publicly available resources such as the MIMIC-III critical care database [6] have enabled empirical studies of treatment strategies and counterfactual prediction in real-world clinical settings. However, evaluation remains challenging because counterfactual outcomes are not observed. Prior work therefore relies on proxy metrics such as PEHE [3], off-policy evaluation methods [17], or policy learning frameworks [18, 19]. Our work follows a similar policy-focused perspective but emphasizes practical evaluation of individualized treatment policies derived from observational ICU data, rather than relying solely on proxy prediction metrics.

Irrelevant covariates and robustness in causal estimation

Including irrelevant or weakly related covariates can increase estimator variance and, in some cases, amplify bias in causal estimation [20, 21]. This issue becomes particularly pronounced in high-dimensional observational datasets, where many measured variables may have only weak relationships with treatment or outcomes. Recent work has emphasized the role of regularization and representation learning in improving robustness of causal estimators in such settings [22-24]. However, systematic empirical evaluation of robustness to irrelevant covariates – especially in real-world clinical data – remains limited. In this paper, we address this gap by introducing a controlled irrelevant-feature stress test and examining how treatment effect estimators behave under increasing nuisance variation.

3. Problem Formulation

We formulate individualized treatment effect (ITE) estimation for vasopressor administration in observational ICU data using the potential outcomes framework. This section defines the data structure, identification assumptions, policy evaluation setup, and the latent-factor perspective on covariates that motivates the disentangled representation used by DRI-ITE.

3.1. Potential Outcomes Framework and Observational Data

We consider an observational dataset $\mathcal{D} = \{(x_i, t_i, y_i)\}_{i=1}^{N}$, where each instance (e.g. an ICU patient) has covariates $x_i \in \mathcal{X} \subseteq \mathbb{R}^K$ representing pre-treatment variables, a binary treatment assignment $t_i \in \mathcal{T} = \{0,1\}$indicating whether vasopressors were administered, and an observed outcome $y_i \in \mathcal{Y}$. In our application, y_i denotes a clinically motivated lactate-change endpoint measuring the change in serum lactate levels during the early ICU stay.

Under the potential outcomes framework [1, 2], [8], each patient has two potential outcomes: y_i^0 if untreated and y_i^1 if treated. However, for each individual we only observe the *factual* outcome

$$y_i = y_i^{t_i},$$

while the counterfactual outcome $y_i^{1-t_i}$ remains unobserved. In observational clinical datasets, treatment assignment depends on patient covariates such as physiological measurements, laboratory values, and clinician judgment. Consequently, the treatment propensity $P(T = 1 \mid X)$ varies across the covariate space, introducing selection bias and making the data fundamentally different from randomized experiments.

We adopt the following standard and widely used assumptions from causal inference.

Stable Unit Treatment Value (SUTVA): There is no interference between units and no hidden versions of treatment, so that each unit's potential outcomes depend only on its own treatment assignment.

Consistency: The observed outcome corresponds to the potential outcome under the received treatment:

$$y_i = y_i^{t_i}$$

Unconfoundedness: The potential outcomes are conditionally independent of treatment given covariates:

$$(y_i^0, y_i^1) \perp\!\!\!\perp t_i \mid x_i$$

Overlap: Every treatment has nonzero probability for all covariate values:

$$0 < P(t = 1 \mid x) < 1 \forall\, x \in \mathcal{X}$$

Unconfoundedness and overlap together imply *strong ignorability* [8]. Under these conditions, the conditional average treatment effect (CATE) is identifiable as

$$\tau(x) = \mathbb{E}[y^1 - y^0 \mid x] = \mathbb{E}[y \mid x, t = 1] - \mathbb{E}[y \mid x, t = 0]$$

At the individual level the treatment effect is defined as

$$\tau\tau_i = y_i^1 - y_i^0$$

The goal of individualized treatment effect estimation is therefore to learn a predictive model

$$f: \mathcal{X} \times \mathcal{T} \to \mathcal{Y},$$

that estimates counterfactual outcomes $\hat{y}^0(x)$ and $\hat{y}^1(x)$.

Although the formulation follows the standard potential outcomes framework, its application to ICU data introduces domain-specific challenges. Treatment assignment reflects clinician-driven decisions under severe and heterogeneous patient conditions, leading to strong confounding and potentially limited overlap between treatment groups. In addition, high-dimensional physiological measurements may include substantial nuisance variation that is weakly related to treatment or outcomes. These characteristics distinguish the present setting from standard benchmark scenarios and motivate the use of representation learning and robustness-focused evaluation.

3.2. Policy Evaluation

Given predicted counterfactual outcomes $\hat{y}^0(x)$ and $\hat{y}^1(x)$from a model f, we define the treatment policy [18, 19]

$$\pi_f(x) = \mathbf{1}\{\hat{y}^1(x) - \hat{y}^0(x) < 0\},$$

which recommends treatment when the predicted individual treatment effect is beneficial, i.e., when the predicted lactate change is negative.

The value of a treatment policy [19] is defined as

$$V(\pi_f) = \mathbb{E}[y^{\pi_f(x)}],$$

representing the expected outcome if the policy were deployed. Because both potential outcomes are not observed simultaneously, we compute a *plug-in estimate* of policy value by substituting model predictions:

$$\hat{V}(\pi_f) = \mathbb{E}[\pi_f(x)\hat{y}^1(x) + \\ +(1 - \pi_f(x))\hat{y}^0(x)]$$

This provides a decision-focused diagnostic for comparing models, although it should not be interpreted as an unbiased estimate of real-world clinical utility. To compare policies, we evaluate improvement relative to the treat-all policy:

$$\Delta_{\text{all}} = \hat{V}(\pi_f) - \hat{V}(\mathbf{1})$$

Because the outcome $y_i$represents lactate change where negative values correspond to improvement, more negative values of Δ_{all}indicate larger improvements relative to treating every patient.

3.3. Latent Factor View of Covariates

To motivate the representation-learning approach used in DRI-ITE [7], we adopt a latent-factor perspective on covariates in which different components of x may influence treatment assignment and outcomes in different ways. In practice, these factors are not directly observable and may only be approximately separable [15].

We conceptually distinguish four types of latent spaces:

- **Γ (treatment-predictive):** covariate information that primarily influences treatment assignment;
- **Δ (confounding):** variables that influence both treatment assignment and outcomes;
- **ϒ (outcome-related):** variables that affect outcomes but are not strong drivers of treatment decisions;
- **Ω (irrelevant or nuisance variation):** covariate information that is weakly related or unrelated to treatment and outcome prediction for the task at hand.

In high-dimensional clinical datasets, irrelevant covariates can inflate variance and degrade treatment effect estimation. Explicitly learning an Ωlatent space provides a mechanism to absorb such variation while encouraging the remaining representations to focus on treatment-relevant structure.

4. DRI-ITE: Deep Disentanglement with Irrelevant Factors

DRI-ITE was originally introduced in [7]. In this work, we apply the method to observational ICU data and briefly summarize its architecture and objective for completeness.

4.1. Architecture

DRI-ITE learns disentangled representations that separate different roles of covariates in treatment and outcome prediction. The model consists of four encoders producing latent factors $\Gamma(\cdot), \Delta(\cdot), \Upsilon(\cdot), \Omega(\cdot)$ representing treatment-predictive, confounding, outcome-related, and irrelevant covariate variation respectively.

Each encoder is implemented as a three-layer MLP with ELU activations and is trained jointly with three downstream components: two outcome regression networks $h_y^0(\cdot), h_y^1(\cdot)$, a treatment classifier $h_c(\cdot)$, and a reconstruction decoder $h_{\text{recon}}(\cdot)$. The regression networks predict potential outcomes and help disentangle Δand ϒthrough supervised outcome prediction. The classification network predicts treatment assignment and encourages separation between treatment-predictive (Γ) and confounding (Δ) representations. Since the nuisance latent space Ω receives no direct supervision signal, the decoder reconstructs the original covariates from all latent

factors, encouraging residual covariate information to be absorbed by Ω.

4.2. Objective Function

The overall objective function is:

$$L_{\text{main}} = L_{\text{reg}} + \alpha L_{\text{class}} + \beta L_{\text{disc}} + \\ + \gamma L_{\text{recons}} + \lambda L_{\text{orth}} + \mu \,\text{Reg}(h_y^1, h_y^0, h_c, h_{\text{recon}}),$$

where Regdenotes L_2weight decay and $\alpha, \beta, \gamma, \lambda, \mu \geq 0$ are hyperparameters controlling the relative contribution of each loss.

Regression loss

Outcome prediction relies on confounding and outcome-related representations:

$$L_{\text{reg}} = \mathcal{L}[y_i,\, h_y^{t_i}(\Delta(x_i), \Upsilon(x_i))]$$

Two outcome networks are trained, one for each treatment group, following [3, 15]. We use mean squared error (MSE).

Classification loss

Treatment prediction uses treatment-predictive and confounding factors:

$$L_{\text{class}} = \mathcal{L}\,[t_i,\, h_c(\Gamma(x_i), \Delta(x_i))]$$

The classifier is trained using binary cross-entropy (BCE).

Discrepancy loss

To reduce selection-induced distribution shift, the model encourages treatment invariance in the Υrepresentation by minimizing a discrepancy between treatment groups:

$$L_{\text{disc}} = \text{disc}[\Upsilon(x_i)\,|_{t_i=0}\,,\, \Upsilon(x_i)\,|_{t_i=1}],$$

where disc(·)denotes a distribution discrepancy measure (implemented as Wasserstein distance in our experiments), following the discrepancy-based balancing approach of [15].

Reconstruction loss

The decoder reconstructs the input covariates from all latent factors:

$$L_{\text{recons}} = \\ = \mathcal{L}[x_i,\, h_{\text{recon}}(\Gamma(x_i), \Delta(x_i), \Upsilon(x_i), \Omega(x_i))]$$

Because Γ, Δ, and Υare anchored by supervised objectives, the reconstruction task encourages remaining covariate information to be captured by the nuisance latent space Ω.

Orthogonality loss

To reduce information leakage between latent spaces, we encourage approximate orthogonality among encoder representations. Let $W_\Gamma, W_\Delta, W_\Upsilon$, and W_Ω denote the effective encoder weight matrices, and let $\bar{W}$ denote the row-wise average of absolute weights [16]. The orthogonality penalty is

$$L_{\text{orth}} = \bar{W}_\Gamma^\top \bar{W}_\Delta + \bar{W}_\Delta^\top \bar{W}_\Upsilon + \bar{W}_\Upsilon^\top \bar{W}_\Gamma + \\ + \bar{W}_\Omega^\top \bar{W}_\Gamma + \bar{W}_\Omega^\top \bar{W}_\Delta + \bar{W}_\Omega^\top \bar{W}_\Upsilon$$

The orthogonality constraints involving Ω are particularly important: without them the nuisance latent space may encode information already captured by other latent factors, weakening isolation of irrelevant variation.

4.3. Prediction

At inference time, individualized treatment effects are estimated using only the confounding and outcome-related representations:

$$\hat{\tau}(x_i) = h_y^1(\Delta(x_i), \Upsilon(x_i)) - h_y^0(\Delta(x_i), \Upsilon(x_i))$$

The treatment-predictive representation $\Gamma(x_i)$ is excluded from outcome prediction because it primarily captures treatment assignment signals, while $\Omega(x_i)$ represents nuisance covariate variation for the current task and is likewise excluded from outcome prediction.

5. Data and Cohort

We construct an observational cohort from the MIMIC-III critical care database in order to study individualized treatment effects of vasopressor administration. The cohort extraction follows a fixed temporal protocol separating baseline covariates, treatment assignment, and outcome measurement windows relative to ICU admission. Fig. 1 illustrates the temporal structure used to define covariates, treatment, and outcome variables.

5.1. MIMIC-III and Cohort Construction

We use the MIMIC-III v1.4 database [6], which contains detailed clinical data for more than 40000 ICU admissions at Beth Israel Deaconess Medical Center between 2001 and 2012. We construct a cohort of adult ICU stays ($\geq$ 18 years) following a fixed extraction protocol. Only the first ICU stay per patient is included to avoid repeated-measure bias. Patients are required to have lactate measurements in both the baseline window (-6 to $+6$hours around ICU admission) and the follow-up window (60--84 hours), and an ICU stay of at least 72 hours to ensure that the outcome window can be observed. These windows are chosen to capture the initial physiological state and subsequent short-term response while allowing for variability in measurement timing.

Treatment

The treatment variable $t_i$indicates whether a vasopressor was administered within the first 24 hours of ICU admission:

$$t_i = \begin{cases} 1 & \text{vasopressor administered} \\ 0 & \text{otherwise.} \end{cases},$$

binary indicator reflecting whether vasopressor was administered during the window, without imposing dosage thresholds, to avoid introducing additional modeling assumptions.

Outcome

The outcome variable y_i is defined as the change in serum lactate between a baseline window and a delayed follow-up window:

$$y_i = \bar{\ell}_i^{(60\text{–}84h)} - \bar{\ell}_i^{(-6\text{–}+6h)}$$

Negative values correspond to lactate reduction and therefore improved metabolic status. The baseline window captures the initial physiological state around ICU admission, while the delayed window reflects short-term response to treatment while reducing sensitivity to measurement noise. Patients with missing lactate measurements in either window are excluded from the analysis.

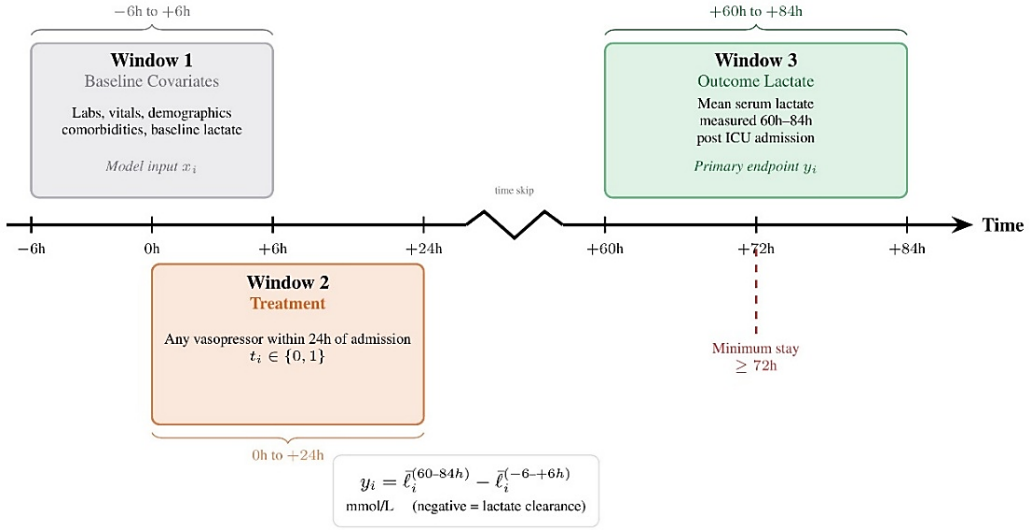


Fig. 1. Temporal structure of cohort extraction from MIMIC-III relative to ICU admission. Window 1 (−6 h to +6 h) defines baseline covariates and reference lactate. Window 2 (0 h to +24 h) defines treatment, where $t_i = 1$ if any vasopressor is administered. Window 3 (+60 h to +84 h) defines the outcome lactate. The outcome is the lactate change between the late and baseline windows, where negative values indicate lactate clearance. A minimum ICU stay of 72 hours is required to observe the outcome window.

Covariates

Each instance is represented by $K = 27$ pre-treatment covariates $x_i \in \mathcal{X} \subseteq \mathbb{R}^K$ including:

- Vital signs: heart rate, systolic and diastolic blood pressure, respiratory rate, temperature, SpO_2;
- Laboratory measurements: baseline lactate, creatinine, bilirubin, sodium, potassium, bicarbonate, hemoglobin, platelet count, white blood cell count, blood urea nitrogen, chloride;
- Neurological status: Glasgow Coma Scale (GCS);
- Comorbidities: diabetes, congestive heart failure, COPD, renal failure, liver disease, hypertension, arrhythmia;
- Demographics: age and sex.

Continuous covariates are standardized using training set statistics and missing covariate values are imputed using median imputation computed from the training data, and the same statistics are applied to the test set.

Dataset Statistics

After cohort construction, the dataset is split into training and test sets using an 85/15 split stratified by treatment assignment.

$$N_{\text{train}} = 6562, N_{\text{test}} = 1159,$$
$$P(t_i = 1 \mid \text{test}) \approx 0.138$$

Outcome statistics on the test set are

$$E[y_i] = -0.411 \pm 1.337,$$
$$E[y_i \mid t_i = 1] = -0.558 \pm \pm 1.435, E[y_i \mid t_i = 0] = -0.387 \pm 1.319$$

The crude treated-minus-control difference

$$E[y_i \mid t_i = 1] - E[y_i \mid t_i = 0] \approx -0.172$$

is confounded by clinical treatment selection and should not be interpreted causally. Table 1 summarizes the main characteristics of the resulting cohort.

6. Evaluation Protocol

Because counterfactual outcomes $y_i^{\neg t_i}$ are unobserved in observational data, no single evaluation metric can fully characterize the quality of individualized treatment effect estimates. In particular,

commonly used metrics such as PEHE are not available, and average-effect measures alone may fail to capture whether a model produces meaningful treatment decisions. We therefore adopt a multi-faceted evaluation protocol designed to assess complementary aspects of model behavior in a clinically relevant setting.

Table 1. Summary statistics of the constructed vasopressor cohort.

Quantity	Value
Total ICU stays	7721
Training set size	6562
Test set size	1159
Treated proportion	13.8 %
Number of covariates (K)	27
Mean outcome E[y_i]	−0.411
Outcome std. dev.	1.337

Specifically, we evaluate models along five dimensions: treatment-effect estimation (via model-implied ATE), decision quality (via policy risk and improvement), heterogeneity (via the distribution of predicted treatment effects), subgroup behavior (via stratified analysis across baseline lactate levels), and robustness (via irrelevant-covariate stress tests). This combination of criteria is intended to capture not only whether a model predicts reasonable average effects, but also whether it induces stable, non-degenerate, and clinically meaningful treatment policies under realistic observational conditions.

6.1. Policy Evaluation

Beyond average treatment effects, we evaluate decision quality using the treat-if-benefit policy. This policy recommends treatment when the predicted effect indicates improved outcome (negative lactate change) [18, 19].

Policy performance is measured using the plug-in policy value $\hat{V}(\pi_f)$ defined in Section 3. Because counterfactual outcomes are not observed, policy value is estimated using model-predicted potential outcomes $\hat{y}^1(x)$ and $\hat{y}^0(x)$.

We report policy improvement relative to the treat-all and treat-none policies:

$$\Delta_{\text{all}} = \hat{V}(\pi_f) - \hat{V}(\mathbf{1}), \Delta_{\text{none}} = \hat{V}(\pi_f) - \hat{V}(\mathbf{0})$$

More negative values indicate larger improvements over the baseline policy. These metrics provide a decision-focused comparison of individualized treatment effect models.

6.2. Model-Implied ATE

We first report the model-implied average treatment effect (ATE) [1, 25]:

$$\widehat{\text{ATE}}_{\text{model}} = \frac{1}{N_{\text{test}}} \sum_{i \in \mathcal{D}_{\text{test}}} (\hat{y}^1(x_i) - \hat{y}^0(x_i))$$

Bootstrap confidence intervals are computed using 500 resamples of the test set to quantify sampling uncertainty.

6.3. Subgroup Analysis

To examine treatment heterogeneity, we stratify test patients by quartiles of baseline lactate measured in the baseline window. For each stratum we compute the mean predicted treatment effect

$$E[\hat{y}^1(x_i) - \hat{y}^0(x_i) \mid \text{stratum}],$$

and the corresponding policy treat rate

$$E[\pi_f(x_i) \mid \text{stratum}],$$

both with bootstrap 95 % confidence intervals [9].

6.4. Heterogeneity of Predicted Treatment Effects

To evaluate heterogeneity in predicted treatment effects, we analyze the empirical distribution of $\hat{\tau}(x) = \hat{y}^1(x) - \hat{y}^0(x)$ on the test set. Because true individual treatment effects are not observed, these statistics provide a diagnostic characterization of model-implied heterogeneity rather than a direct measure of estimation accuracy. We report summary statistics including mean(τ), standard deviation, and extreme percentiles (p1, p99), which capture both overall dispersion and tail behavior of predicted treatment effects.

6.5. Irrelevant-Covariate Stress Test

To evaluate robustness to irrelevant covariates, we augment the covariate vector x_i with $|\Omega| \in \{5,10,15,20\}$ synthetic features drawn i.i.d. from $\mathcal{N}(0,1)$. These synthetic features are appended to both the training and test sets. All baseline models are retrained on the augmented datasets [22]. For DRI-ITE, separate checkpoints are trained for each noise level $|\Omega|$. We report policy improvement Δ_{all} and policy treat rate across noise levels. Together, these evaluation criteria assess whether treatment-effect models produce stable and clinically plausible decision policies in observational ICU data where counterfactual outcomes are not observed.

7. Baseline Methods

We compare DRI-ITE against several standard heterogeneous treatment effect estimators widely used in causal machine learning.

S-learner

A single outcome model is trained to predict y_i from the augmented feature vector (x_i, t_i). Counterfactual outcomes are obtained by intervening on the treatment indicator t_i. We implement the

S-learner using both Ridge regression and a Random Forest (RF) regressor [9].

T-learner

Separate outcome models $\hat{\mu}_0$and $\hat{\mu}_1$are trained on the control ($t_i = 0$) and treated ($t_i = 1$) subsets, respectively. Counterfactual outcomes are predicted using the model corresponding to the desired treatment condition. We use Ridge and RF regressors [9].

X-learner

The X-learner [9] is a two-stage procedure designed for settings with treatment imbalance. First, outcome models $\hat{\mu}_0$ and $\hat{\mu}_1$ are estimated. Pseudo-treatment effects are then imputed for each unit using the observed outcome and the predicted counterfactual outcome. Separate effect models $\hat{\tau}_0(x)$ and $\hat{\tau}_1(x)$ are trained on the control and treated groups. The final treatment effect estimate is a propensity-weighted combination:

$$\hat{\tau}(x) = g(x)\,\hat{\tau}_0(x) + (1 - g(x))\,\hat{\tau}_1(x),$$

where $g(x) = P(t = 1 \mid x)$ denotes the propensity score. We implement the X-learner using Ridge regression.

Causal Forest (CF)

We use the Generalized Random Forest (GRF) implementation [11], which estimates heterogeneous treatment effects using adaptive tree-based partitioning and provides asymptotic inference procedures.

DRCFR

DRCFR is a disentangled representation-learning approach for treatment effect estimation that separates treatment-related and outcome-related latent factors while encouraging balanced representations across treatment groups [15].

RLOCFR

RLOCFR extends DRCFR by incorporating regularization mechanisms that promote more structured latent representations and improved robustness to irrelevant covariates [16].

8. Results

We evaluate DRI-ITE on the MIMIC-III vasopressor cohort against standard meta-learners, causal forests, and disentanglement-based baselines. Throughout this section, policy quantities should be interpreted as *plug-in* estimates computed from model-predicted potential outcomes rather than estimates of deployed clinical utility.

8.1. Policy-Based Evaluation

We next evaluate decision quality under the treat-if-benefit rule $\pi_f(x) = \mathbf{1}\{\hat{y}^1(x) - \hat{y}^0(x) < 0\}$. Table 2 reports estimated plug-in policy risks and improvements relative to the treat-all policy. Under this criterion, DRI-ITE achieves the largest improvement over treat-all, with $\Delta_{\text{all}} = -0.132 \pm 0.009$ and CI $[-0.147, -0.117]$. It also attains the lowest policy risk $\mathcal{R}_{\text{pol}}(\pi_f) = -0.671 \pm 0.029$among all methods. The strongest meta-learning baselines, X-learner and T-learner with Ridge regression, achieve $\Delta_{\text{all}} \approx -0.073$and -0.072, respectively. The representation-based baselines also improve over several standard learners, with DRCFR at $-0.068 \pm 0.007$and RLOCFR at -0.062 ± 0.003, but both remain substantially below DRI-ITE. DRI-ITE maintains a non-degenerate treatment recommendation rate of 0.567 ± 0.014, avoiding collapse to trivial treat-all or treat-none policies, whereas several baselines exhibit near-degenerate behavior. S-learner (Ridge) recommends treatment for all patients, yielding no gain over treat-all, while Causal Forest remains highly aggressive with a treatment rate of 0.965 ± 0.005. These results highlight that models with similar ATE magnitudes can induce qualitatively different treatment policies when evaluated from a decision perspective.

8.2. Average Treatment Effect Estimates

Table 3 reports model-implied ATE estimates together with bootstrap confidence intervals. DRI-ITE estimates ATE $= -0.168 \pm 0.032$ with 95 % CI $[-0.212, -0.124]$, which is similar in magnitude to the T- and X-learners, Causal Forest, and the disentanglement baselines. Overall, ATE estimates alone provide only weak separation between methods. Several models yield broadly similar average effects while differing much more substantially in their implied treatment policies, motivating a decision-focused evaluation.

Table 2. Plug-in policy evaluation under the treat-if-benefit rule. DRI-ITE achieves the largest improvement over the treat-all policy ($\boldsymbol{\Delta_{\text{all}}}$) and the lowest policy risk $\boldsymbol{\mathcal{R}_{\text{pol}}(\pi_f)}$, while several baselines collapse toward near-degenerate policies (e.g., treat-all). This shows that models with similar ATE can induce substantially different decision policies. Values are reported as mean ± standard deviation over bootstrap resamples; best values are in bold.

Method	$\mathcal{R}_{\text{pol}}(\mathbf{0})$	$\mathcal{R}_{\text{pol}}(\mathbf{1})$	$\mathcal{R}_{\text{pol}}(\pi_f)$	Δ_{all}	CI	Δ_{none}	Treat rate
X-learner (Ridge)	-0.381±0.000	-0.586±0.000	-0.659±0.000	-0.073±0.000	[-0.081, -0.066]	-0.279±0.000	0.645±0.000
T-learner (Ridge)	-0.381±0.000	-0.587±0.000	-0.659±0.000	-0.072±0.000	[-0.080, -0.065]	-0.279±0.000	0.645±0.000
T-learner (RF)	-0.381±0.000	-0.561±0.002	-0.596±0.001	-0.035±0.001	[-0.041, 0.029]	-0.215±0.002	0.710±0.004
S-learner (RF)	-0.401±0.001	-0.456±0.001	-0.456±0.001	0.000±0.000	[0.000, 0.000]	-0.054±0.002	0.876±0.016
S-learner (Ridge)	-0.384±0.000	-0.615±0.000	-0.615±0.000	0.000±0.000	[0.000, 0.000]	-0.232±0.000	1.000±0.000
Causal Forest	-0.381±0.000	-0.546±0.008	-0.548±0.008	-0.001±0.000	[-0.002, -0.001]	-0.167±0.008	0.965±0.005
DRCFR	-0.358±0.015	-0.536±0.042	-0.603±0.037	-0.068±0.007	[-0.074, -0.061]	-0.245±0.031	0.600±0.016
RLOCFR	-0.358±0.018	-0.551±0.035	-0.613±0.032	-0.062±0.003	[-0.069, -0.056]	-0.255±0.027	0.621±0.017
DRI-ITE (ours)	-0.372±0.015	-0.539±0.036	**-0.671±0.029**	**-0.132±0.009**	**[-0.147, -0.117]**	-0.299±0.027	0.567±0.014

Table 3. Model-implied ATE estimates with bootstrap confidence intervals. Most methods yield similar ATE magnitudes, including DRI-ITE, indicating that average effects alone provide limited discrimination between models and motivating policy-based evaluation.

Method	Est.	CI_{low}	CI_{hi}
T-learner (Ridge)	-0.207 ±0.000	-0.234±0.000	-0.182±0.001
T-learner (RF)	-0.180 ±0.002	-0.205±0.003	-0.155±0.003
S-learner (Ridge)	-0.232 ±0.000	-0.232±0.000	-0.232±0.000
S-learner (RF)	-0.054 ±0.002	-0.063±0.002	-0.046±0.002
X-learner (Ridge)	-0.206 ±0.000	-0.234±0.000	-0.181±0.001
Causal Forest	-0.166±0.008	-0.183±0.010	-0.150±0.006
DRCFR	-0.177±0.035	-0.209±0.040	-0.145±0.029
RLO-DRCFR	-0.192±0.029	-0.225±0.033	-0.160±0.023
DRI-ITE	-0.168±0.032	-0.212±0.037	-0.124±0.027

8.3. Subgroup Analysis by Baseline Lactate

To examine how treatment decisions vary with baseline severity, we stratify patients by baseline lactate quartiles and report policy treatment rates in Table 4. DRI-ITE recommends treatment for roughly 60 % of patients across quartiles, with only modest variation across severity groups. This pattern is more balanced than the highly aggressive policies learned by several baselines: S-learner (Ridge) recommends treatment for all patients in every quartile, and Causal Forest remains above 0.91 even in the highest-lactate quartile. The representation-based baselines also produce non-degenerate policies, but with higher treatment rates than DRI-ITE across all quartiles. DRCFR and RLOCFR remain in the range of roughly 0.70--0.77, whereas DRI-ITE stays closer to a balanced 0.60. These patterns are consistent with the main policy results: DRI-ITE does not collapse to treat-all while maintaining a stable decision rule across severity strata.

8.4. Heterogeneity of Predicted Treatment Effects

To further characterize the heterogeneity captured by each model, we examine the empirical distribution of predicted individual treatment effects $\hat{\tau}(x) = \hat{y}^1(x) - \hat{y}^0(x)$on the test set. Table 5 summarizes the mean, standard deviation, and extreme percentiles of the predicted treatment-effect distributions. Several baselines exhibit limited variability. In particular, S-learner (Ridge) produces a nearly constant effect for all patients, with essentially zero dispersion, which is fully consistent with its degenerate treat-all policy. S-learner (RF) also exhibits comparatively narrow dispersion. The T- and X-learners produce broader distributions, with standard deviations around 0.45 and positive upper tails. The representation-based baselines further increase treatment-effect dispersion. DRCFR and RLO-DRCFR both produce wider distributions than the meta-learners, but DRI-ITE yields the broadest distribution overall, with std $= 0.775$, p1 $= -3.572$, and p99 $= 1.227$.

Table 4. Policy treatment rates stratified by baseline lactate quartiles. DRI-ITE maintains a stable, non-degenerate treatment rate across severity levels, while several baselines exhibit near-uniform or highly aggressive policies with limited variation across subgroups. This indicates differences in how models adapt treatment decisions to patient severity.

Baseline lactate quartile	S(RF)	S(Ridge)	T(RF)	T(Ridge)	X(Ridge)	CF	DRCFR	RLOCFR	DRI-ITE
Q1 (≤ -0.636)	0.475	1.000	0.528	0.419	0.419	0.996	0.702	0.741	0.606
Q2 (-0.636, -0.297]	0.990	1.000	0.845	0.491	0.491	0.993	0.711	0.752	0.615
Q3 (-0.297, 0.340]	0.997	1.000	0.748	0.748	0.745	0.954	0.753	0.771	0.603
Q4 (> 0.340)	0.975	1.000	0.722	0.931	0.935	0.917	0.703	0.722	0.606

Table 5. Distribution of predicted treatment effects $\hat{\tau}(x)$ on the test set. DRI-ITE exhibits the largest spread (Std, p1, p99), indicating richer heterogeneity, while some baselines (e.g., S-learner with Ridge) produce near-constant predictions. This suggests that models differ substantially in the degree of patient-level variation they capture.

Method	Mean $\hat{\tau}$	Std($\hat{\tau}$)	p1	p99
S-learner (RF)	-0.053	0.145	-0.787	0.002
S-learner (Ridge)	-0.232	0.000	-0.232	-0.232
T-learner (RF)	-0.181	0.450	-2.185	0.538
T-learner (Ridge)	-0.207	0.453	-1.841	0.620
X-learner (Ridge)	-0.206	0.456	-1.863	0.623
Causal Forest	-0.181	0.310	-1.476	0.014
DRCFR	-0.180	0.571	-2.754	0.494
RLO-DRCFR	-0.195	0.567	-2.697	0.495
DRI-ITE	-0.172	0.775	-3.572	1.227

This indicates that DRI-ITE captures both strongly beneficial and potentially harmful responses, rather than collapsing toward a near-constant effect estimate. Importantly, this broader predicted heterogeneity does not translate into an extreme treatment policy: DRI-ITE maintains a balanced treatment rate while achieving the largest policy improvement in Table 2. Taken together, these results suggest that the learned disentangled representation captures richer patient-level variation without inducing degenerate or overly aggressive treatment policies.

8.5. Robustness to Irrelevant Covariates

We assess robustness to nuisance variation by appending synthetic Gaussian covariates to the feature set. Table 6 reports policy improvement Δ_{all} and policy treatment rates as irrelevant features are added. Several baselines degrade noticeably under this perturbation.

Causal Forest remains close to treat-all throughout, and S-learner (Ridge) remains fully degenerate at every noise level. The Ridge-based meta-learners are more stable but continue to achieve substantially smaller policy improvements than DRI-ITE. The representation-based baselines remain non-degenerate under feature inflation, but their improvements are consistently weaker than those of DRI-ITE. DRCFR ranges from -0.068 to -0.041 and RLOCFR from -0.062 to -0.040, while both shift toward more aggressive treatment policies as irrelevant covariates are added. In contrast, DRI-ITE maintains strong policy improvement at every noise level, ranging from -0.110 to -0.150, while its treatment rate remains tightly controlled between 0.552 and 0.571. These results suggest that the disentanglement objective used by DRI-ITE helps provide a more robust separation between treatment-relevant structure and nuisance variation, reducing the tendency of the policy to drift toward indiscriminate treatment when irrelevant covariates are introduced. Fig. 2 complements Table 6 by visualizing these trends. The left panel shows that DRI-ITE consistently outperforms competing methods across all noise levels, whereas other methods deteriorate more substantially as $|\Omega|$ increases. The right panel shows that DRI-ITE preserves a stable and balanced treatment rate, while several baselines drift toward more aggressive treatment behavior.

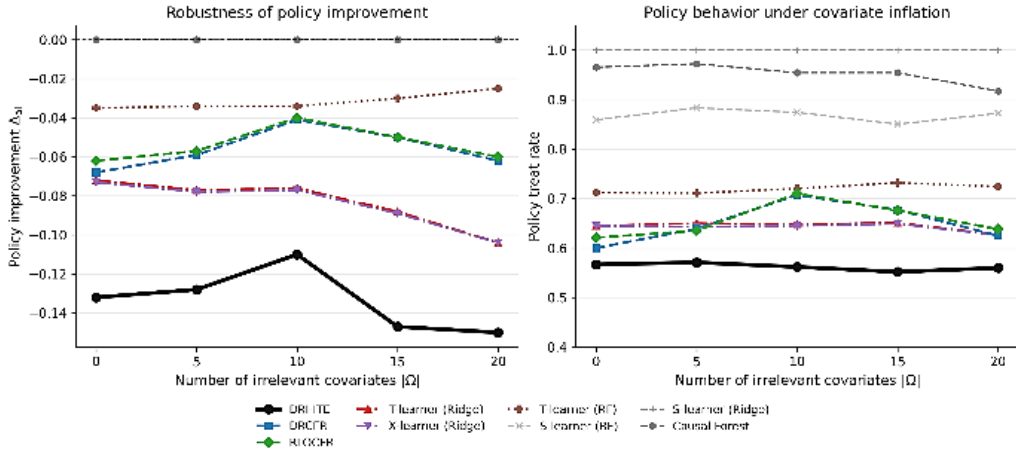


Fig. 2. Robustness under irrelevant-feature inflation. Left: policy improvement $\boldsymbol{\Delta_{all}}$ as the number of appended nuisance covariates increases (more negative is better). Right: corresponding treatment rate. DRI-ITE maintains the strongest and most stable policy improvement across noise levels, while several baselines degrade or drift toward more aggressive, near-degenerate policies.

Table 6. Robustness under irrelevant-feature inflation. Left: policy improvement $\boldsymbol{\Delta_{all}}$ as the number of appended nuisance covariates increases (more negative is better). Right: corresponding treatment rate. DRI-ITE maintains the strongest and most stable policy improvement across noise levels, while several baselines degrade or drift toward more aggressive, near-degenerate policies.

Method	Δ_{all}					Treat rate				
	0	**5**	**10**	**15**	**20**	**0**	**5**	**10**	**15**	**20**
S-learner (RF)	0.000	0.000	0.000	0.000	0.000	0.859	0.883	0.874	0.850	0.872
S-learner (Ridge)	0.000	0.000	0.000	0.000	0.000	1.000	1.000	1.000	1.000	1.000
T-learner (RF)	−0.035	−0.034	−0.034	−0.030	−0.025	0.712	0.711	0.720	0.732	0.724
T-learner (Ridge)	−0.072	−0.077	−0.076	−0.088	−0.104	0.645	0.650	0.648	0.651	0.627
X-learner (Ridge)	−0.073	−0.078	−0.077	−0.089	−0.104	0.645	0.643	0.645	0.649	0.626
Causal Forest	0.000	0.000	0.000	0.000	0.000	0.965	0.972	0.954	0.954	0.917
DRCFR	−0.068	−0.059	−0.041	−0.050	−0.062	0.600	0.639	0.707	0.677	0.625
RLOCFR	−0.062	−0.057	−0.040	−0.050	−0.060	0.621	0.635	0.710	0.676	0.638
DRI-ITE	**−0.132**	**−0.128**	**−0.110**	**−0.147**	**−0.150**	0.567	0.571	0.562	0.552	0.560

8.6. Role Identification via Disentangled Representations

Finally, we examine the learned feature-role decomposition. Fig. 3 visualizes absolute feature-contribution profiles for the four latent factors. The confounding factor Δis primarily associated with acute physiological indicators including baseline lactate, multiple vital signs, metabolic laboratory measurements, and neurological status. This is clinically plausible, as such variables can influence both vasopressor administration and subsequent lactate trajectory. The adjustment factor Υis strongly influenced by renal failure, suggesting that renal dysfunction carries outcome-related information that is not primarily routed through the treatment-predictive factor. The factor denoted Γis best interpreted as a *model-implied treatment-predictive latent space* rather

than a formally validated set of instrumental variables. It is driven by features such as diabetes, COPD, liver disease, and systolic blood pressure, indicating that the learned treatment-assignment pathway reflects both chronic comorbidity burden and hemodynamic state. The nuisance factor Ωcaptures variables such as gender, SpO_2, bicarbonate, and CHF that exhibit comparatively lower contribution to the supervised prediction objectives under the imposed factorization. Overall, the separation between factors is broadly consistent with the intended behavior of the orthogonality and reconstruction objectives.

This separation improves interpretability and robustness by isolating nuisance variation and distinguishing confounding from outcome-related factors, ensuring that predicted effects reflect clinically meaningful signals. At the same time, these assignments should be interpreted as *model-implied* rather than externally validated causal categories.

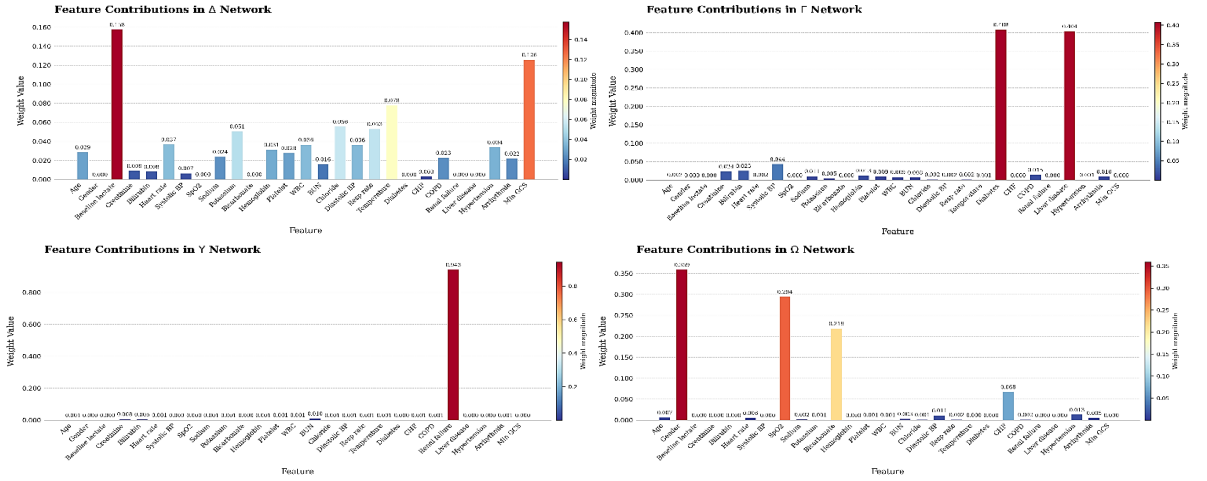


Fig. 3. Feature-contribution profiles for the four latent factors. Higher values indicate stronger reliance on a covariate within each factor. The confounding factor **Δ** concentrates on acute severity markers, **Υ** captures outcome-related variation (e.g., renal failure), **Γ** reflects treatment-predictive structure, and **Ω** absorbs nuisance variation. Overall, the separation illustrates how the model organizes covariates into distinct functional roles in the learned representation.

Summary: Across these complementary diagnostics, DRI-ITE produces the strongest policy improvements while maintaining a balanced treatment rate and remaining comparatively robust to irrelevant feature inflation. Although several baselines yield broadly similar average treatment effects, policy-based evaluation, heterogeneity diagnostics, and robustness analyses reveal substantial differences in the quality and stability of individualized treatment decisions.

9. Discussion

Why ATE comparisons are insufficient

Most methods produce ATE estimates of similar magnitude, yet their policy behavior differs sharply. In particular, S-learner (Ridge) attains a competitive model-implied ATE while recommending treatment for every patient, yielding no policy improvement over treat-all. This disconnect illustrates why evaluating individualized treatment effect estimators solely through average-effect recovery can be misleading in observational clinical settings.

Degenerate policies as a failure mode

Near-universal treatment emerges as a systematic failure mode for several baselines, especially under irrelevant feature inflation. Such models may still report reasonable average effects, but they fail to produce actionable individualization. Explicitly reporting policy treatment rates and subgroup behavior is therefore essential for diagnosing whether a model is genuinely personalizing or merely reproducing an almost constant treatment recommendation.

Why DRI-ITE is more robust

A possible explanation for DRI-ITE's robustness is that the orthogonality and reconstruction objectives discourage nuisance information from entering the latent spaces used for treatment effect prediction. The orthogonality loss L_{orth} limits overlap between learned latent spaces, while the reconstruction loss L_{recons} provides an explicit sink for irrelevant variation. Together, these mechanisms may help preserve meaningful heterogeneity in predicted treatment effects $\hat{\tau}_i$ under irrelevant-feature inflation.

Clinical implications

The DRI-ITE policy treats roughly half of the patients, and this recommendation rate remains moderate across baseline lactate strata. This suggests that the model identifies a sizable subgroup that may not benefit from vasopressor administration under the current observational treatment pattern. At the same time, these findings are derived entirely from observational data and plug-in counterfactual predictions; they should be viewed as hypothesis-generating rather than deployment-ready.

Limitations

Unmeasured confounding: The unconfoundedness assumption is untestable, and unrecorded severity markers or clinical decision factors may bias all estimates.

Plug-in policy evaluation: Policy value $\mathcal{R}_{\text{pol}}(\pi_f)$ is computed from model-predicted counterfactual outcomes rather than observed outcomes under deployment. Consequently, it reflects model behavior rather than realized clinical utility.

Limited support and extrapolation: Despite substantial overall propensity-score overlap, the learned policy recommends treatment more frequently than observed, including in low-propensity regions. This reflects broadly negative predicted treatment effects but also indicates reliance on extrapolation, making estimates in such regions inherently more uncertain and requiring cautious interpretation.

Outcome proxy: Lactate change is an intermediate physiological endpoint and may not fully capture longer-term clinical outcomes such as mortality, organ recovery, or length of ICU stay.

External validity: MIMIC-III is derived from a single tertiary-care center in the United States, which may limit generalizability to other hospitals or patient populations.

Role interpretation: Feature-role assignments in the disentangled representation are model-implied and should not be interpreted as externally validated causal categories.

Together, these results suggest that explicitly separating irrelevant covariate variation may improve the stability and interpretability of individualized treatment policies derived from high-dimensional observational clinical data.

10. Conclusion

We presented a policy-focused evaluation framework for individualized treatment effect (ITE) models on MIMIC-III vasopressor data, integrating model-implied ATE estimation, plug-in policy risk evaluation, subgroup analysis, and an irrelevant-covariate stress test. Applying this framework, we showed that standard meta-learners and causal forests, while competitive in ATE recovery, can approach near-degenerate treat-all policies and lose robustness as Ω increases. DRI-ITE, by explicitly disentangling treatment-predictive (Γ), confounding (Δ), outcome-related (Υ), and irrelevant (Ω) covariate variation through reconstruction and orthogonality objectives, maintains heterogeneous treatment effect predictions and achieves larger policy improvements across experimental conditions.

References

[1]. D. B. Rubin, Estimating causal effects of treatments in randomized and nonrandomized studies, *Journal of Educational Psychology*, Vol. 66, Issue 5, 1974, pp. 688-701.

[2]. J. Neyman, On the application of probability theory to agricultural experiments, *Statistical Science*, Vol. 5, Issue 4, 1990, pp. 465-480.

[3]. U. Shalit, F. D. Johansson, D. Sontag, Estimating individual treatment effect: Generalization bounds and algorithms, in *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 2017, pp. 3076-3085.

[4]. J. M. Robins, A. Rotnitzky, L. P. Zhao, Estimation of regression coefficients when some regressors are not always observed, *Journal of the American Statistical Association*, Vol. 89, Issue 427, 1994, pp. 846-866.

[5]. H. Bang, J. M. Robins, Doubly robust estimation in missing data and causal inference models, *Biometrics*, Vol. 61, Issue 4, 2005, pp. 962-973.

[6]. A. Johnson, T. Pollard, R. Mark, MIMIC-III clinical database, *PhysioNet*, 2016, 10.13026/C2XW26.

[7]. A. S. Khan, E. Schaffernicht, J. A. Stork, On the effects of irrelevant variables in treatment effect estimation with deep disentanglement, in *Proceedings of the 27th European Conference on Artificial Intelligence (ECAI)*, 2024, pp. 416-423.

[8]. P. R. Rosenbaum, D. B. Rubin, The central role of the propensity score in observational studies for causal effects, *Biometrika*, Vol. 70, Issue 1, 1983, pp. 41-55.

[9]. S. R. Künzel, J. S. Sekhon, P. J. Bickel, B. Yu, Metalearners for estimating heterogeneous treatment effects using machine learning, *Proceedings of the National Academy of Sciences*, Vol. 116, Issue 10, 2019, pp. 4156-4165.

[10]. S. Athey, G. W. Imbens, Recursive partitioning for heterogeneous causal effects, *Proceedings of the National Academy of Sciences*, Vol. 113, Issue 27, 2016, pp. 7353-7360.

[11]. S. Wager, S. Athey, Estimation and inference of heterogeneous treatment effects using random forests, *Journal of the American Statistical Association*, Vol. 113, Issue 523, 2018, pp. 1228-1242.

[12]. F. Johansson, U. Shalit, D. Sontag, Learning representations for counterfactual inference, in *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, 2016, pp. 3020-3029.

[13]. C. Louizos, U. Shalit, J. M. Mooij, D. Sontag, et al., Causal effect inference with deep latent-variable models, in Advances in Neural Information Processing Systems 30 (I. Guyon, U. von Luxburg, S. Bengio, H. Wallach, et al., Eds.), *Curran Associates, Inc*., Red Hook, 2017, pp. 6449-6459.

[14]. C. Shi, D. M. Blei, V. Veitch, Adapting neural networks for the estimation of treatment effects, in Advances in Neural Information Processing Systems 32 (H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, et al., Eds.), *Curran Associates, Inc*., Red Hook, 2019, pp. 2507-2517.

[15]. N. Hassanpour, R. Greiner, Learning disentangled representations for counterfactual regression, in *Proceedings of the 8th International Conference on Learning Representations (ICLR)*, 2020.

[16]. M. Cheng, X. Liao, Q. Liu, B. Ma, et al., Learning disentangled representations for counterfactual regression via mutual information minimization, in *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2022, pp. 1802-1806.

[17]. M. Dudík, J. Langford, L. Li, Doubly robust policy evaluation and learning, in *Proceedings of the 28th International Conference on Machine Learning (ICML)*, 2011, pp. 1097-1104.

[18]. T. Kitagawa, A. Tetenov, Who should be treated? Empirical welfare maximization methods for treatment choice, *Econometrica*, Vol. 86, Issue 2, 2018, pp. 591-616.

[19]. S. Athey, S. Wager, Policy learning with observational data, *Econometrica*, Vol. 89, Issue 1, 2021, pp. 133-161.

[20]. J. Pearl, On a class of bias-amplifying covariates that endanger effect estimates, in *Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence (UAI)*, 2010, pp. 425-432.

[21]. C. Cinelli, C. Hazlett, Making sense of sensitivity: Extending omitted variable bias, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, Vol. 82, Issue 1, 2020, pp. 39-67.

[22]. P. R. Hahn, J. S. Murray, C. M. Carvalho, Bayesian regression tree models for causal inference: Regularization, confounding, and heterogeneous effects, *Bayesian Analysis*, Vol. 15, Issue 3, 2020, pp. 965-1056.

[23]. V. Chernozhukov, et al., Double/debiased machine learning for treatment and structural parameters, *The Econometrics Journal*, Vol. 21, Issue 1, 2018, pp. C1-C68.

[24]. M. J. Vowels, N. C. Camgoz, R. S. Sherratt, Targeted VAE: Variational and targeted learning for causal inference, in *Proceedings of the 21st IEEE International Conference on Data Mining Workshops (ICDMW)*, 2021, pp. 1051-1060.

[25]. G. W. Imbens, D. B. Rubin, Causal Inference for Statistics, Social, and Biomedical Sciences, *Cambridge University Press*, Cambridge, 2015.

(039)

Evaluating Artificial Intelligence in Generating Biochemistry Knowledge Assessments for Medical Education

Veljkovic Andrej [1], Aleksandar Mitic [1], Ognjen Radovic [2], Monika Simjanoska Misheva [3] and Stevo Lukic [1]
[1] Faculty of Medicine, University in Nis, Serbia
[2] Faculty of Economics, University of Nis, Serbia
[3] Faculty of Computer Science and Engineering in Skopje, Serbia
Tel.: + 381 63654293
E-mail: andrej.veljkovic@medfak.ni.ac.rs

Summary: This study evaluates the quality and educational usability of AI generated multiple choice tests for practical biochemistry courses at the Faculty of Medicine, University of Niš. We have compared model × input-format effects under expert professor evaluation. Two AI systems, Google Gemini and ChatGPT, were used to generate four test variants based on plain text and PDF input formats. Five professors assessed the tests using six criteria, including content coverage, difficulty balance, and professional validity. Results showed that the Gemini model with PDF input achieved the highest ratings, closely followed by ChatGPT with PDF input, while ChatGPT with plain text performed the weakest. Statistical analysis (Friedman and Wilcoxon tests) confirmed significant differences among test variants, although effect sizes were small. The findings suggest that input format plays a critical role in the quality of AI generated assessments. Overall, AI demonstrates strong potential as a supportive tool in test development, but expert evaluation remains essential to ensure pedagogical accuracy, reliability, and alignment with learning objectives.

Keywords: Artificial intelligence, Input format, Multiple choice tests, Biochemistry education, Supportive tool.

1. Introduction

The possibility of the integration of Artificial Intelligence (AI) into the education process has opened new possibilities for automated test questions generation for student assessment [1]. In the last few years, advances in large language models and AI driven educational tools have enabled the rapid creation of learning materials and adaptive assessments. Traditional teaching paradigms are now transforming [2]. In medical education the development of high-quality assessment materials remains a time consuming task for educators. This is especially important in biochemistry, where complex molecular mechanisms require deep conceptual understanding. There is a need to design questions that ensure adequate content coverage with balanced difficulty [3]. AI based systems have the potential to assist educators by generating optimal assessment content aligned with learning objectives.

However, despite these advantages, there are concerns regarding the pedagogical validity and consistency of AI generated assessment materials [1]. Therefore, professor evaluation of AI generated assessments is very important to determine their applicability. It is particularly essential in medical education where accuracy and quality standards are crucial.

2. Aim

This study aimed to evaluate the quality and educational usability of AI generated multiple choice tests designed for practical biochemistry courses at the Faculty of Medicine, University of Niš. In addition, the study aimed to compare model × input-format effects under expert professor evaluation in a practical biochemistry course, as well as to determine their potential role in supporting educators in the development of reliable, pedagogically appropriate evaluation tools.

3. Material and Methods

In this study we have used two AI systems: Google Gemini (Version 1.5 Flash) and ChatGPT (GPT-5.5, OpenAI), to generate tests consisting of 10 questions. Each MCQ had a single correct answer. The source material was the official practicum used by medical students. We have used two different input formats: direct text input and PDF based input. Both plain text and PDF contained formatting, images and tables Both AI systems got an identical prompt. The only variation was the format of the input material, either plain text or a PDF document. So, we had: Test Gemini with plain text input (Test 1), Test Gemini with PDF input (Test 2), Test ChatGPT with plain text input (Test 3), and Test ChatGPT with PDF input (Test 4).

The generated tests were independently evaluated by five professors using a structured scoring system (1–5 scale) across six criteria. Criteria were: content coverage, difficulty balance, cognitive level diversity, redundancy avoidance, professional validity, and overall quality and usability in teaching. All tests were without factual errors, and all the tests aligned with specific learning objectives.

4. Results

In Table 1 descriptive statistics of the test scores is represented.

Table 1. Descriptive statistics of test scores.

Test	Minimum	Maximum	Mean	St. dev
Test 1	3.00	5.00	3.9	0.96
Test 2	3.00	5.00	4.43	0.62
Test 3	2.00	5.00	4.03	0.89
Test 4	3.00	5.00	4.36	0.66

Descriptive analysis showed that Test 2 input achieved the highest mean teacher rating (M = 4.43; SD = 0.62), followed by Test 4 (M = 4.36; SD = 0.66), then Test 3 (M = 4.03; SD = 0.89), while Test 1 had the lowest mean score (M = 3.90; SD = 0.96).

In the next set of analyses, we performed a Friedman test at the level of individual assessments (30 blocks) in order to more precisely evaluate differences among the tests. The Friedman test indicated a statistically significant difference among the four tests, $\chi^2(3, N = 30) = 12.366$, $p = 0.006$, with a small effect size (Kendall's W = 0.137).

Fig. 1 shows the distribution of scores for four tests.

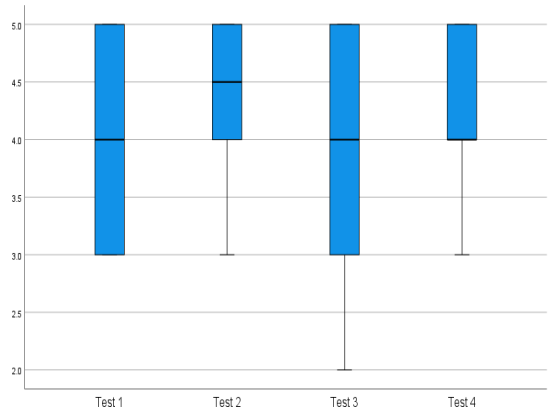


Fig. 1. Box plot showing the distribution of scores for four AI generated tests.

The box plot (Fig. 1) shows that Test 2 and Test 4 exhibited a more favorable distribution of scores compared to Test 3. Test 2 had a higher median and a narrower concentration of scores in the upper range of the scale, while Test 4 demonstrated a similar, stable pattern. In contrast, Test 3 showed the widest range and was the only test to include a score of 2, indicating lower and less consistent quality. This visual finding is consistent with the descriptive statistics and Friedman rankings, according to which Test 2 and Test 4 were rated more favorably than Test 3.

To enable a more detailed examination of differences between individual tests, pairwise Wilcoxon signed-rank tests were conducted. Before correction, statistically significant differences were observed between Test 2 and Test 1 ($p = 0.005$), Test 3 and Test 2 ($p = 0.004$), and Test 4 and Test 3 ($p = 0.023$). To control the familywise Type I error rate across six pairwise comparisons, Holm's sequential correction procedure was applied. After adjustment, only the differences between Test 2 and Test 1 (Holm-adjusted $p = 0.005$) and between Test 2 and Test 3 (Holm-adjusted $p = 0.004$) remained statistically significant. The difference between Test 4 and Test 3 did not remain significant after correction (Holm-adjusted $p = 0.092$).

5. Discussion

The integration of artificial intelligence in education is increasingly recognized as a necessity in contemporary teaching. The frequent creation of assessments, particularly in courses where tests are used for formative rather than summative evaluation, may lead to educator fatigue and an increased likelihood of errors. Recent studies suggest that the quality of AI generated questions is influenced by the model employed, prompt design, and the extent of expert supervision [4]. This is particularly critical in medical and biomedical education, where high standards of accuracy and responsibility are required [5].

Statistical analysis was conducted on four distinct test sets, as described earlier. The most robust finding of the study is the advantage of Test 2. This variant achieved the highest mean score, the highest mean rank in the Friedman analysis at the level of 30 observations, and a statistically significant advantage over Test 1 and Test 3 in pairwise comparisons. In other words, Test 2 was the most successful in balancing content coverage, difficulty level, cognitive diversity, and overall pedagogical acceptability.

Test 4 represents a very close alternative. Its mean score is nearly identical to that of Test 2, and the Wilcoxon comparison between these two variants did not reveal a statistically significant difference. A large number of ties in the Test 4 – Test 2 comparison further confirms that these two variants are highly similar in quality. In practical terms, if the goal is the rapid selection of a pedagogically acceptable AIgenerated test, Test 2 and Test 4 represent the two most rational candidates for further use or refinement.

Test 1 occupies a middle position. Its ratings are not low; however, compared to Test 2, it demonstrated a statistically significantly lower performance. This suggests that Test 1 is not a failed AI generated output, but rather a variant that likely requires more extensive teacher revision prior to direct implementation.

Test 3 is consistently the lowest rated variant. It has the lowest mean score, the lowest mean rank, and is the only test to include the minimum score of 2. Compared to Test 2 and Test 4, it demonstrates a less favorable pattern across multiple criteria, suggesting that in its current form, it is not an optimal choice for immediate instructional use without substantial expert revision.

The Friedman analysis revealed that differences among the test sets become more discernible and statistically testable at a finer level of evaluation. Nevertheless, the low value of Kendall's W coefficient suggests that these differences are relatively small, reflecting moderate variations distributed across individual criteria. The Wilcoxon test show us the best results for Test 2. But, the results of the Wilcoxon tests should be considered exploratory. When applying a more stringent control of Type I error, the most robust contrasts are those favoring Test 2 over Test 1 and Test 3, whereas the comparison between Test 4 and Test 3 should be interpreted with caution [6].

Due to its interactive capabilities, generative AI can serve as a supportive tool in the learning process, enhancing both student engagement and teaching effectiveness [7].

So, AI should not be regarded as an autonomous author of final assessments, but rather as a tool for generating multiple candidate drafts from which the most appropriate version can be selected and refined [4]. This approach offers several advantages: it enhances the likelihood of achieving an optimal balance between difficulty, content coverage, and cognitive complexity. It facilitates the identification of weaknesses in individual AI generated outputs prior to implementation. It also mitigates the risk that formally correct tests remain pedagogically misaligned with intended learning outcomes. Consequently, expert evaluation remains a critical intermediary step between AI generated content and its effective use in educational practice.

The findings should be interpreted in light of several limitations. The number of evaluators was small (N = 5), which limits the statistical power of the conservative analysis of aggregated scores. Also, the assessments were based on expert judgment and an ordinal scale, which is appropriate for evaluating teaching materials but still involves a certain degree of subjectivity. Future studies should involve a larger pool of evaluators, a wider range of instructional units, and multiple AI models or prompt engineering strategies.

6. Conclusion

Overall, the findings suggest that Test Gemini with PDF input and Test ChatGPT with PDF input are the highest quality AI generated variants, with Test Gemini with PDF input showing the most consistent performance advantage. These results indicate that AI generated assessments hold substantial practical potential in biochemistry education, although their reliable implementation necessitates systematic expert validation.

Acknowledgements

This project is funded by the European Union under Horizon Europe (project ChatMED grant agreement ID: 101159214)

References

[1]. R. Luckin, W. Holmes, M. Griffiths, L. B. Forcier, Intelligence unleashed: An argument for AI in education, *Pearson*, London, 2016.

[2]. L. Chen, P. Chen, Z. Lin, Artificial intelligence in education: A review, *IEEE Access*, Vol. 8, 2020, pp. 75264-75278.

[3]. E. J. Topol, High-performance medicine: The convergence of human and artificial intelligence, *Nature Medicine*, Vol. 25, Issue 1, 2019, pp. 44-56.

[4]. Y. Artsi, Y. Bichovsky, E. Manisterski, S. Dvorkin, et al., Large language models for generating medical examinations: Systematic review, *BMC Medical Education*, Vol. 24, Issue 1, 2024, 354.

[5]. K. W. Chan, F. Ali, J. Park, K. S. B. Sham, et al., Automatic item generation in various STEM subjects using large language model prompting, *Computers and Education: Artificial Intelligence*, Vol. 8, 2025, 100344.

[6]. S. Holm, A simple sequentially rejective multiple test procedure, *Scandinavian Journal of Statistics*, Vol. 6, Issue 2, 1979, pp. 65-70.

[7]. F. Wu, Y. Dang, A systematic review of responses, attitudes, and utilization behaviors on generative AI for teaching and learning in higher education, *Behavioral Sciences*, Vol. 15, Issue 4, 2025, 467.

(040)

AI-Based Bacterial Detection Using Multiple Biosensing Technologies under Data Scarcity

Felipe Yamada [1,2], António Cardoso [1], Flávia Barbosa [1,3] and Luís Guimarães [1,2]
[1] INESC TEC, Porto, Portugal
[2] Faculty of Engineering, University of Porto, Porto, Portugal
[3] School of Economics and Management, University of Porto, Porto, Portugal
Tel.: + 351 222094000
E-mail: felipe.yamada@inesctec.pt

Summary: This work presents a methodological overview of artificial intelligence-assisted biosensing approaches for antimicrobial resistance monitoring under constrained experimental conditions. Three complementary sensing platforms are investigated: a graphene-based lab-on-a-chip, an artificial nose for volatile organic compound detection, and a phage-based biosensor for protein identification. For graphene sensors, the study evaluates the limitations of conventional Dirac voltage shift calibration under temporal drift and device variability, demonstrating that analysis of full transfer characteristics combined with drift-aware preprocessing improves classification performance under temporally structured evaluation. The artificial nose platform applies transfer learning and few-shot learning to classify bacterial volatile organic compound signatures under limited-data conditions. Similarly, the phage-based biosensor employs interpolation-based synthetic augmentation to improve classification robustness from limited experimental datasets. Overall, the results illustrate how combining device-aware signal interpretation with data-efficient machine learning methods can improve robustness and reliability in prototype-stage biosensing systems.

Keywords: Machine learning, Biosensing, Few-shot learning, Data augmentation, Graphene sensors, Antimicrobial resistance.

1. Introduction

Antimicrobial resistance represents a growing global health challenge requiring rapid and reliable diagnostic technologies capable of detecting pathogens across diverse environments [1]. Recent advances in biosensing and artificial intelligence (AI) have enabled the development of portable sensing platforms capable of extracting clinically relevant information from complex biological and chemical signals [2, 3]. However, prototype-stage biosensing systems frequently operate under practical constraints such as limited labeled datasets, sensor variability, and unstable measurement conditions. Under these conditions, establishing robust and generalizable machine learning models remains challenging, particularly during early-stage device validation.

Within the SMARTgNOSTICS consortium, AI methods are being investigated as supporting tools for the interpretation and validation of multiple biosensing technologies targeting antimicrobial resistance-related applications. In this work, three complementary sensing platforms are considered: a graphene-based lab-on-a-chip, an artificial nose targeting volatile organic compounds, and a phage-based biosensor for protein detection. Although these platforms rely on different sensing principles, they share common methodological challenges associated with signal variability, data scarcity, and reliable feature extraction from complex sensor responses.

To address these limitations, different data-efficient AI strategies were explored according to the characteristics of each sensing modality. For graphene sensors, the study focuses on interpreting full transfer-characteristic signals under temporal drift conditions instead of relying exclusively on scalar calibration metrics. For the artificial nose platform, transfer learning and few-shot learning approaches were applied to improve bacterial volatile organic compound classification under limited-data conditions. For the phage-based biosensor, interpolation-based augmentation techniques were investigated to improve the robustness of supervised classification from limited experimental datasets.

Overall, this work presents a methodological overview of AI-assisted approaches for biosensing systems operating under constrained experimental conditions.

2. Graphene-Based Device Interpretation

Graphene field-effect transistor (GFET) biosensors are promising platforms for antimicrobial-resistance monitoring due to their high electrostatic sensitivity and compatibility with portable lab-on-a-chip architectures [2, 4]. Graphene sensors are being investigated as experimental biosensing devices whose electrical responses require robust AI-assisted interpretation under conditions of temporal drift and limited experimental datasets.

Conventional GFET sensing strategies typically rely on the Dirac voltage shift (ΔVDirac) as a scalar calibration metric [4]. However, the relationship between analyte concentration and ΔVDirac may become non-ideal under certain sensing conditions, leading to poor concentration estimation performance. In the evaluated datasets, classification using ΔVDirac

alone remained near chance level, with accuracies of approximately 26–28 %. In contrast, machine learning models using full transfer characteristics achieved substantially higher performance by leveraging the complete transfer-curve information instead of a single scalar descriptor.

The study evaluated convolutional neural networks (1D CNN), multilayer perceptrons (MLP), and Random Forest models using full transfer-curve representations extracted from electrolyte-gated graphene sensors. Under random train–test splits, neural models achieved accuracies above 92 % for functionalized GFET datasets and above 98 % for non-functionalized graphene devices. However, temporally ordered evaluation revealed a significant degradation in performance for functionalized sensors, with accuracy decreasing to approximately 59 % for the 1D CNN model, demonstrating the impact of temporal drift and distribution shift on predictive generalization.

To address this limitation, drift-aware preprocessing based on transfer-curve alignment was applied prior to classification. The procedure compensates for systematic offsets between complete transfer-curve readings before model evaluation. After preprocessing, classification performance under temporal evaluation increased to approximately 97 % accuracy.

3. Artificial Nose Platform

The artificial nose platform investigates bacterial identification through detection of VOCs emitted by bacteria. The system is based on a wavelength-multiplexed photoionization detector composed of four ionization lamps with distinct photon energies, enabling the acquisition of multiple temporal current-response curves associated with bacterial VOC profiles. Because different bacterial species emit distinct VOC mixtures, the resulting signals provide characteristic response patterns that can be explored for bacterial differentiation.

Within the SMARTgNOSTICS consortium, this platform is being investigated as an example of AI-assisted biosensing under limited-data conditions, where collecting large labeled datasets is experimentally expensive and time-consuming. To address this limitation, the sensor signals were transformed into image-based representations and analyzed using transfer learning and few-shot learning strategies. A pre-trained ResNet-18 convolutional neural network was used for feature extraction within a Prototypical Networks framework, enabling classification from a small number of labeled examples [5-8].

The platform demonstrated real-time differentiation of four clinically relevant bacterial species, namely Escherichia coli, Staphylococcus aureus, Pseudomonas aeruginosa, and Klebsiella pneumoniae, while also detecting bacterial concentrations as low as 10^2 CFU [8]. Under the original imbalanced dataset, the few-shot learning approach achieved approximately 89 % accuracy for bacterial identification. After balancing the dataset using interpolation-based synthetic signal generation, classification accuracy increased to approximately 96 % [8-10]. In addition, the system demonstrated the ability to distinguish between low and high bacterial concentrations, supporting the feasibility of AI-assisted VOC sensing under constrained experimental conditions.

These results illustrate how combining portable sensing platforms with data-efficient AI techniques can support bacterial detection systems, particularly in scenarios where data acquisition is limited and signal variability is significant.

4. Phage-Based Biosensor

The phage-based biosensing platform investigates protein classification using spectral signatures generated by a prototype sensing device functionalized with bacteriophage-derived receptor-binding proteins. The system was evaluated using fluorescent proteins and their mixtures as representative analytes for biosensing validation. A rule-based classifier had previously been developed for signal interpretation. In this work, machine learning approaches were investigated to improve classification accuracy and sensitivity under lower concentration conditions.

To improve signal consistency, a background-correction procedure was applied using multiple baseline measurements acquired in the absence of protein samples. Each protein signal was corrected against five distinct background profiles, generating differential response curves used as input features for classification. Given the limited amount of experimental data available, interpolation-based synthetic signal augmentation was additionally employed to generate realistic training samples while preserving the structural characteristics of the original biosensor responses. The augmentation procedure incorporated controlled Gaussian noise and low-frequency jitter to simulate biologically and instrumentally plausible variability [9, 10].

Two classification approaches were evaluated. The first consisted of a rule-based peak-detection model based on heuristic thresholding, while the second employed a supervised XGBoost classifier trained on the synthetically augmented dataset [11]. The parametric approach achieved approximately 90 % accuracy on confidently assigned predictions but produced classifications for only about 56 % of the samples due to conservative thresholding under low-signal conditions. In contrast, the supervised model demonstrated strong classification performance, achieving approximately 99.98 % accuracy on the synthetic holdout test set. These results suggest that supervised learning approaches combined with controlled synthetic augmentation can support classification under limited-data conditions.

5. Conclusions

This work presented a methodological overview of AI-assisted biosensing approaches developed within the SMARTgNOSTICS consortium for antimicrobial-resistance monitoring under constrained experimental conditions. Across the three sensing platforms investigated, namely graphene-based sensors, an artificial nose for VOC detection, and a phage-based biosensor, the results demonstrated that combining device-aware signal interpretation with data-driven modeling can substantially improve robustness and classification performance in limited-data scenarios.

For graphene sensors, the analysis of complete transfer characteristics together with drift-aware preprocessing improved classification performance under temporally structured evaluation, highlighting the importance of accounting for temporal variability in biosensing workflows. In the artificial nose platform, transfer learning and few-shot learning enabled accurate bacterial differentiation from VOC response patterns despite limited labeled datasets. For the phage-based platform, interpolation-based synthetic augmentation combined with supervised learning improved classification robustness under noisy and low-signal measurement conditions.

Although the sensing modalities investigated rely on different physical principles, all platforms exhibited common challenges related to data scarcity, signal variability, and experimental instability during prototype-stage validation. The results therefore support the use of data-efficient AI methods as complementary tools for improving interpretation and reliability in emerging biosensing systems.

Future work will focus on expanding dataset diversity, evaluating model robustness under more realistic operating conditions, and improving validation under temporal and experimental variability. Additional efforts will investigate adaptive preprocessing strategies, datasets acquired across multiple experimental sessions, and broader testing across sensing conditions and analyte classes to support the development of more reliable biosensing platforms for practical AMR monitoring applications.

Acknowledgements

This work is co-financed by Component 5 – Capitalization and Business Innovation, integrated in the Resilience Dimension of the Recovery and Resilience Plan within the scope of the Recovery and Resilience Mechanism (MRR) of the European Union (EU), framed in the Next Generation EU, for the period 2021–2026, within project SMARTgNOSTICS, with reference 5. This work was also supported by the Portuguese Foundation for Science and Technology – FCT through PhD research grant under reference 2025.04109.BD.

References

[1]. Antimicrobial Resistance, *World Health Organization*, 2024.

[2]. A. Pannone, A. Raj, H. Ravichandran, S. Das, et al., Robust chemical analysis with graphene chemosensors and machine learning, *Nature*, Vol. 634, 2024, pp. 572-578.

[3]. M. Xue, C. Mackin, W.-H. Weng, J. Zhu, et al., Integrated biosensor platform based on graphene transistor arrays for real-time high-accuracy ion sensing, *Nature Communications*, Vol. 13, Issue 1, 2022, 5064.

[4]. A. Purwidyantri, J. Mouro, P. Alpuim, J. Borme, Graphene-based field-effect transistor biosensors for DNA detection, *ACS Sensors*, Vol. 5, Issue 6, 2020, pp. 1736-1745.

[5]. K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770-778.

[6]. J. Deng, W. Dong, R. Socher, L.-J. Li, et al., ImageNet: A large-scale hierarchical image database, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009, pp. 248-255.

[7]. J. Snell, K. Swersky, R. Zemel, Prototypical networks for few-shot learning, in Advances in Neural Information Processing Systems 30 (I. Guyon, U. von Luxburg, S. Bengio, H. Wallach, et al., Eds.), *Curran Associates, Inc*., Red Hook, 2017, pp. 4077-4087.

[8]. S. P. Costa, A. Cardoso, H. Mahmoodnia, F. Gonçalves, et al., Bacterial species differentiation via real-time detection of microbial volatile organic compounds using a wavelength multiplexed photoionization detector and AI image-based analysis, *Scientific Reports*, 2026, 10.1038/s41598-026-46818-x.

[9]. C. Oh, S. Han, J. Jeong, Time-series data augmentation based on interpolation, *Procedia Computer Science*, Vol. 175, 2020, pp. 64-71.

[10]. G. Forestier, F. Petitjean, P. Gancarski, A. Termier, Generating synthetic time series to augment sparse datasets, in *Proceedings of the IEEE International Conference on Data Mining (ICDM)*, 2017, pp. 865-870.

[11]. T. Chen, C. Guestrin, XGBoost: A scalable tree boosting system, in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2016, pp. 785-794.

(044)

The Ethical Implications of Artificial Intelligence in Orthopedic Surgery: A Systematic Review

Tobi Kamoru, Rebbeca Alemu, Nuhame Mulugeta, Muna Jalani, John Cyrus and Lauren A. Barber
Virginia commonwealth University School of Medicine, USA
E-mail: Tobikams@gmail.com, laurenbarbermd@gmail.com

Summary: Artificial intelligence (AI) is increasingly integrated into orthopedic surgery. However, its adoption raises important ethical concerns. Existing literature primarily emphasizes technical performance, with limited attention to issues that could impact patients and surgeons. A PRISMA-guided systematic review was conducted to identify studies published between 2015-2025 addressing ethical considerations of AI in orthopedic surgery. Of 1393 records screened, 95 underwent full-text review, and 38 studies met inclusion criteria. Privacy, consent, and data security (92 %) and bias, accuracy, equity, and fairness (82 %) were the most frequently discussed domains. Transparency (68.4 %), professional accountability (57.8 %), and clinical validation/regulation (52.2 %) were addressed less consistently. Most studies discussed AI in general orthopedics (n = 21), followed by spine (n = 9), and arthroplasty (n = 5). Over 86 % of studies were published between 2024-2025, reflecting recent rapid growth. Despite the growing interest in AI, ethical considerations remain a tertiary focus within the literature. Discourse is concentrated in general orthopedics and spine-focused literature, with limited representation of other subspecialties.

Keywords: Artificial intelligence, AI, Machine learning, ML, Orthopedic surgery, Ortho, Hand surgery, Ethics, Spine surgery, Foot and ankle, Tumor, Arthroplasty.

1. Introduction

Artificial intelligence (AI) refers to computational systems capable of performing tasks that typically require human cognition, including pattern recognition, probabilistic reasoning, and predictive analysis. Within modern medicine, AI, and in particular machine learning (ML) and deep learning, has enabled the extraction of clinically meaningful insights from large, complex datasets that exceed human analytic capacity [1, 2]. Orthopedic surgery has emerged as a leading domain for AI integration due to its use of imaging, standardized procedural workflows, and measurable clinical outcomes. AI applications are increasingly used in diagnostic imaging, where algorithms can assist in fracture detection, osteoarthritis grading, and implant evaluation [3, 4]. In addition, predictive models are being developed to estimate perioperative risk, anticipate complications, and forecast recovery trajectories [5]. AI is also transforming operative planning through tools that support implant selection, alignment optimization, and robotic-assisted procedures, particularly in arthroplasty [6]. Beyond direct clinical applications, AI systems are improving efficiency through workflow optimization, automated documentation, and decision support [2]. Together, these advancements are reshaping precision, efficiency, and scalability in orthopedic practice.

Despite the rapid expansion of AI applications in orthopedics, the current body of literature remains disproportionately focused on technical performance, including model accuracy, predictive validity, and efficiency gains [2, 5]. While these contributions are essential, they often overlook the ethical considerations that determine real-world clinical acceptability and sustainability.

Furthermore, ethical discussions surrounding AI in medicine are typically generalized and not tailored to the specific context of orthopedic surgery. Orthopedics presents unique challenges, including procedural interventions, implant-dependent decisions, and long-term functional outcomes. These factors introduce ethical complexities that are not fully captured by existing frameworks.

The primary aim of this systematic review is to provide a rigorous and comprehensive synthesis of the ethical implications of artificial intelligence in orthopedic surgery. Specifically, this study seeks to identify and categorize the ethical domains addressed in the current literature, evaluate the depth and consistency of ethical engagement across orthopedic subspecialties and AI applications, and highlight critical gaps that may impede safe and equitable clinical implementation.

2. Methods

This systematic review was conducted in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guideline. These guidelines were used to provide an organizational framework for study identification, screening, and eligibility assessment (Fig. 1).

The inclusion criteria include papers published from 2015 to 2025, publications in English, peer reviewed, and discussing orthopedic related subspecialties. Exclusion criteria include general AI ethics papers, studies discussing non-orthopedic specialties as well as studies that do not meet the inclusion criteria.

This review concentrated on orthopedic surgery-related articles that discussed the ethical implications of artificial intelligence in orthopedic surgery. A comprehensive search strategy was developed in collaboration with a research librarian, utilizing both controlled vocabulary and keyword-based terms related to artificial intelligence, orthopedic surgery, and ethics. Search terms related to artificial intelligence included "artificial intelligence", "machine learning", "deep learning", "neural networks", "large language model", and "algorithms". Terms related to orthopedic surgery included "orthopedic", "orthopaedic", "spine surgery", "arthroplasty", "sports medicine", "trauma", and other procedure-specific terminology. Ethical concepts were captured using terms such as "ethics", "ethical" and "bioethics". These terms were combined using Boolean operators and adapted for each database. Article search was conducted on three databases, namely Pubmed (n = 505), Embase (n = 750), and Web of Science (n = 699). All searches were limited to articles published between 2015 and 2025 to reflect contemporary development in artificial intelligence.

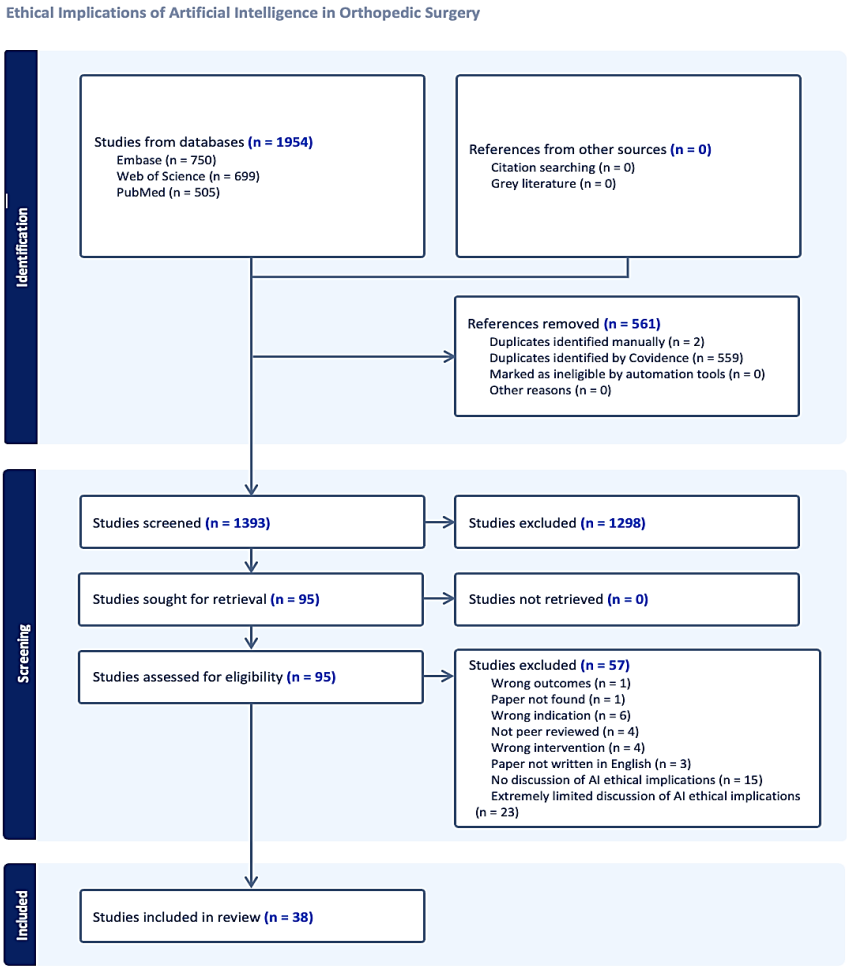


Fig. 1. PRISMA Flowsheet.

3. Results

The literature search identified a total of 1954 records across databases, including Embase (n = 750), Web of Science (n = 699), and PubMed (n = 505) (Fig. 1). No additional records were identified through other sources. Following removal of 561 duplicate records, 1393 articles remained for title and abstract screening. Of these, 1298 articles were excluded based on irrelevance to the study objectives. A total of 95 studies underwent full text-review, 57 articles were excluded, and 38 met the inclusion criteria for final analysis.

There was a modest increase (n = 3) in publications between 2021 to 2023 (Fig. 3), suggesting a gradual emergence of ethical discourse as AI applications became widely explored. However, the most substantial growth occurred in recent years. In 2024,

the number of included studies increased notably to 10 publications, reflecting a growing recognition of the need to address ethical challenges alongside technological advancements. This upward trend continued sharply into 2025, which accounted for the largest proportion of included studies (n = 23, 60.5 %).

The majority of the studies discussed artificial intelligence in general orthopedics (n = 21, 55.3 %), followed by spine surgery (n = 9), and arthroplasty (n = 5) (Fig. 4). Less frequently represented orthopedic subspecialties included trauma (n = 2), hand (n = 2), and pediatrics (n = 1).

Across the included articles, multiple ethical domains were identified, with varying frequency and emphasis. The most frequently discussed ethical concern was privacy, consent, and data security, reported in 92.1 % (n = 35) of articles (Fig. 2). These concerns centered on the protection of patient data, risks of data breaches, and challenges related to informed consent in the context of AI-driven decision-making. The second most commonly cited domain was bias, accuracy, equity, and fairness, represented in 81.6 % (n = 31) of articles.

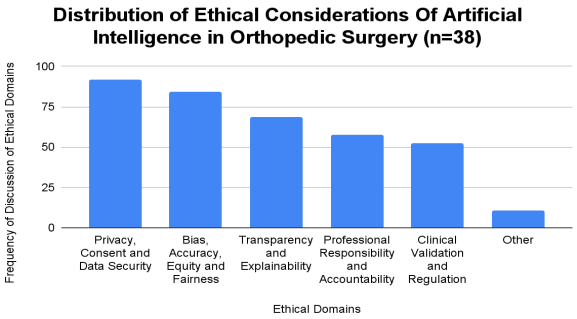


Fig. 2. Distribution of Ethical Considerations in AI Applications in Orthopedic Surgery.

4. Discussion/Conclusions

Despite the growing interest in AI, ethical considerations remain a tertiary focus within the literature. Approximately 75 % of discourse is concentrated in general orthopedics and spine-focused literature, with limited representation of other subspecialties. While privacy and bias are frequently discussed, fewer studies address regulatory and clinical validation challenges. As AI tools transition toward broader clinical integration, limited engagement with regulatory and validation considerations may pose challenges to implementation. Future work should prioritize empirically grounded evaluation and specialty-specific ethical frameworks to guide responsible adoption. As the use of AI tools expands in orthopedic surgery, limited ethical guidance may impact appropriate clinical adoption across subspecialties. Orthopedic surgery presents unique challenges because AI-driven recommendations may directly influence intraoperative decisions, implant selection, and perioperative management. Without clearly defined standards for accountability, informed consent, explainability, and physician oversight, uncertainty arises.

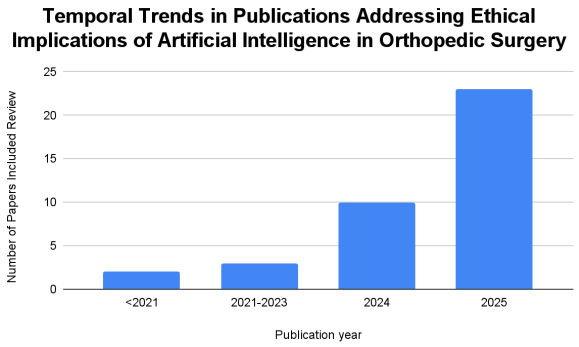


Fig. 3. Temporal Trends in Publications Addressing Ethical Implications of Artificial Intelligence in Orthopedic Surgery.

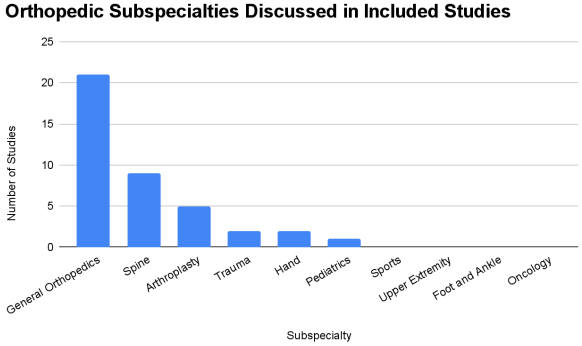


Fig. 4. Orthopedic Subspecialties Discussed in Included Studies.

References

[1]. J. Amann, A. Blasimme, E. Vayena, D. Frey, et al., Explainability for artificial intelligence in healthcare: A multidisciplinary perspective, *BMC Medical Informatics and Decision Making*, Vol. 20, Issue 1, 2020, 310.

[2]. C. Batailler, et al., Artificial intelligence in knee arthroplasty: Current concepts, *Knee Surgery, Sports Traumatology, Arthroscopy*, Vol. 29, Issue 11, 2021, pp. 3413-3422.

[3]. I. Y. Chen, E. Pierson, S. Rose, S. Joshi, et al., Ethical machine learning in healthcare, *Nature Medicine*, Vol. 27, Issue 9, 2021, pp. 1489-1494.

[4]. R. Lindsey, A. Daluiski, S. Chopra, A. Lachapelle, et al., Deep neural network improves fracture detection by clinicians, *Proceedings of the National Academy of Sciences*, Vol. 115, Issue 45, 2018, pp. 11591-11596.

[5]. D. McGraw, Building public trust in uses of Health Insurance Portability and Accountability Act de-identified data, *Journal of the American Medical Informatics Association*, Vol. 20, Issue 1, 2013, pp. 29-34.

[6]. Z. Obermeyer, B. Powers, C. Vogeli, S. Mullainathan, Dissecting racial bias in an algorithm used to manage the health of populations, *Science*, Vol. 366, 2019, pp. 447-453.

Appendix

Table 1. Papers Included in Study.

First Author	Year	Title	Journal
Myers	2020	Artificial Intelligence and Orthopaedics: An Introduction for Clinicians	J Bone Joint Surg Am
Cervone	2024	The Perspectives of Robotic Surgeons	Surg Technol Int
Adida	2024	Machine Learning in Spine Surgery: A Narrative Review	Neurosurgery
Rizk	2024	Machine Learning-Assisted Decision Making in Orthopaedic Oncology	JBJS Rev
Ali	2025	Artificial Intelligence in Planning for Spine Surgery	Curr Rev Musculoskelet Med
Han	2025	Artificial Intelligence in Orthopedic Surgery: Current Applications, Challenges, and Future Directions	MedComm
Yao	2024	Large Language Models in Orthopaedics: Definitions, Uses, and Limitations	J Bone Joint Surg Am
Carroll	2024	Generative Artificial Intelligence and Prompt Engineering: A Primer for Orthopaedic Surgeons	JBJS Rev
Leal	2025	Artificial Intelligence in Orthopaedic and Trauma Surgery Education: Applications, Ethics, and Future Perspectives	J Am Acad Orthop Surg Glob Res Rev
Kaya Bicer	2023	Artificial Intelligence Use in Orthopedics: An Ethical Point of View	EFORT Open Rev
Bartkowski	2025	Artificial Intelligence in Medicine With Emphasis on Orthopedic Practice	Cureus
Vaishya	2025	Integrating Artificial Intelligence into Orthopedics: Opportunities, Challenges, and Future Directions	J Hand Microsurg
Zeitlin	2025	Machine Learning for Hand Surgeons: Emerging Clinical Applications	J Hand Surg Am
Banskota	2025	Artificial Intelligence in Orthopaedic Education, Training and Research: A Systematic Review	BMC Med Educ
Oettl	2025	Artificial Intelligence Agents in Orthopaedics: Concepts, Capabilities and the Road Ahead	Knee Surg Sports Traumatol Arthrosc
Kim	2025	Artificial Intelligence in Total Knee Arthroplasty: Clinical Applications and Implications	Knee Surg Relat Res
Banatwala	2024	A Comprehensive Exploration of Artificial Intelligence in Orthopaedics Within Lower-Middle-Income Countries	J Pak Med Assoc
Koucheki	2025	Integrating Artificial Intelligence and Virtual Reality in Orthopedic Surgery: A Comprehensive Review	HSS J
Dashtbozorg	2024	Emerging Technologies in Hand Orthopedic Surgery: Current Trends and Future Directions	Galen Med J
Kumar	2025	Integrating Artificial Intelligence in Orthopedic Care: Advancements in Bone Care and Future Directions	Bioengineering
Baghbani	2025	The Revolutionary Impact of Artificial Intelligence in Orthopedics: Comprehensive Review of Current Benefits and Challenges	J Robot Surg

First Author	Year	Title	Journal
Sanker	2025	Current Trends and Future Prospects of Language Models and Processing Systems in Spine Surgery	Neurosurg Rev
Giorgino	2023	ChatGPT in Orthopedics: A Narrative Review Exploring the Potential of Artificial Intelligence in Orthopedic Practice	Front Surg
Kiwinda	2025	Bioethical Considerations of Deploying Artificial Intelligence in Clinical Orthopedic Settings	HSS J
Branstetter	2025	Navigating the Intersection of Technology and Surgical Education	Orthop Clin North Am
Gurusamy	2025	Leveraging Data to Transform Surgical Outcomes and Precision in Orthopaedics	J Clin Orthop Trauma
Bhadani	2025	Data-Driven Strategies in Orthopaedics: Optimizing Surgical Precision and Patient Outcomes	Indian J Orthop
Shahid	2025	Effectiveness and Reliability of AI in Diagnosis and Robot-Assisted Spinal and Cranial Surgery	Ann Med Surg
Muelbauer	2025	The Future is Now: How AI is Reshaping Spine Care	N Am Spine Soc J
Kader	2025	The Future of Precision Orthopaedics: Personalized Data-Driven Practice	Bone Jt Open
Yahanda	2025	Current Applications and Future Implications of Artificial Intelligence in Spine Surgery and Research	Global Spine J
Salman	2025	Transforming Orthopedics: A Glimpse Into the Future With Artificial Intelligence	J Musculoskelet Surg Res
Jawaraya	2025	Medicolegal Implications in Orthopaedic Surgery: The Emerging Challenges of Artificial Intelligence and Robotic Integration	Indian Journal of Orthopaedics
Nugraha	2025	AI in Pediatric Spine Care: Clinical, Research, and Ethical Considerations	Journal of Clinical Medicine
Singh Rana	2025	A Bioethical Perspective on Orthopaedic Robot-Assisted Surgery: Consent, Access, and Accountability	Journal of Bone and Joint Surgery
Fields	2024	Implications of Artificial Intelligence	Seminars in Spine Surgery
Prasanna	2024	Intersection of Orthopaedics and Artificial Intelligence: A Review	SSR Institute of International Journal of Life Sciences
Jayakumar	2019	Value-based Healthcare: Can Artificial Intelligence Provide Value in Orthopaedic Surgery?	Clinical Orthopaedics and Related Research

(045)

AI-Concordant Care: A Novel Approach toward Precision Depression Treatment

Yijun Shao [1,4], Yan Cheng [1,4], Hank Wen-Chih Wu [2,3], Tracey H. Taveira [2,5], John E. McGeary [2,3], Kevin McConeghy [2,3], Ying Yin [1,4], Ali Ahmed [1,4], Qing Zeng-Treitler [1,4]

[1] George Washington University, Biomedical Informatics Center, Washington, DC, USA
[2] Providence VA Medical Center, Providence, RI, USA
[3] Brown University, Providence, RI, USA
[4] Washington DC VA Medical Center, Washington, DC, USA
[5] University of Rhode Island, Kingston, RI, USA
Tel.: + 001 2029942987
E-mail: zengq@gwu.edu

Summary: Major depressive disorder (MDD) is a common and serious mental health condition for which individual treatment response varies greatly. We developed a deep neural network (DNN) model to predict patient-level depression symptom outcomes and evaluated AI-concordant care, defined as treatment regimens aligned with AI-predicted symptom reduction. In 230323 patients with MDD, the DNN model was trained and validated to predict follow-up Patient Health Questionnaire-9 (PHQ-9) scores (range: 0–27). In testing, mean and median absolute errors were 4.0 and 3.0 points, respectively; 73 % of patients had prediction errors ≤ 5 points. In a separate cohort of 41309 Veterans with MDD receiving sertraline, PHQ-9 score reduction occurred in 63.9 % of the AI-concordant group compared with 42.9 % of the AI-discordant group. AI-concordant regimens were associated with greater odds of symptom reduction after multivariable adjustment (OR 1.17, 95 % CI 1.10–1.24; p<0.001). These findings support AI-based precision MDD treatment as a promising framework to improve outcomes.

Keywords: Major depressive disorder, PHQ-9, Deep neural network, AI-concordant care, Precision treatment, Clinical decision support.

1. Introduction

Major depressive disorder (MDD) is a common mental health (MH) condition that impacts both physical and mental function [1]. Approximately 21.0 million or 8.3 % of US adults have had at least one MDD episode [2], and a meta-analysis found that 23 % of Veterans have depression [3]. MDD is associated with risks of hospitalization (12 %) [4] and death (14.8 suicide deaths per 100000 population) [5]. The options for treatment range from medications, psychotherapy, to electroconvulsive therapy [6], as well as novel therapies such as transcranial magnetic stimulation (TMS) [7]. The medications include antidepressants, antipsychotics, and anticonvulsants, among others, highlighting the complexity of the pharmacologic treatments for MDD. Comorbidities, non-MDD medications, and genetic and environmental factors further complicate the treatment choices.

Fewer than 60 % of individuals respond to the first antidepressant trial and almost a third have adverse effects that can undermine continued use [8-10]. Many patients must undergo multiple trials (of doses and/or medications) to obtain a meaningful change in depression symptoms [11, 12]. The guidance on MDD treatment to date has been limited by the number of RCTs with small sample sizes to draw on and massive costs to obtain larger sample sizes.

Additionally, traditional statistical approaches in the analysis of data limit the determination of multi-layer, often non-linear interactions. Although prediction models and clinical decision support tools are emerging, their scope remains limited, often focusing narrowly on acute treatment or initiation with monotherapy or stopping short of providing clear guidance informed by individualized outcome prediction [13-16]. The main innovation of the study lies in the prediction of MDD symptoms at any given clinical encounter and across medication regimens extending beyond monotherapy, as well as the evaluation of a clearly defined AI-concordant care treatment strategy.

2. Objectives

To demonstrate the feasibility of training AI models for MDD treatment response prediction and to assess whether AI-concordant treatment regimens are associated with improved MDD symptoms.

3. Methods

We trained a DNN model to predict patient-level Patient Health Questionnaire-9 (PHQ-9) outcomes. We first created a cohort of 230323 patients with MDD, each with a randomly selected PHQ-9 survey date as the index date and a follow-up PHQ-9 survey, as the outcome PHQ-9 score, occurring 30-180 days after the index date. Since every PHQ-9 score is one of

28 integers, 0, 1, 2,..., 27, the DNN model was designed to output a probability distribution over the 28 integers, and the predicted score was the smallest integer k ($0 \leq k \leq 27$) such that the cumulative probability from 0 to k was >0.5. For each patient, the absolute value of the difference between the predicted score and the actual score was the absolute error. The input data included MDD duration, baseline PHQ-9 score, time to outcome PHQ-9 survey, clinical history (MDD diagnosis, comorbidities, and prior healthcare utilization), demographic characteristics (age, race, and sex (as a biological variable)), and treatment history. No post-index treatment information or future clinical data were used when generating predictions intended to support baseline treatment decisions. The model was trained on 80 % of the patients, validated on the next 10 %, and tested on the remaining 10 %. The rationale for including time to the outcome PHQ-9 survey as a feature was that treatment response is inherently time-dependent, and incorporating follow-up time enables the trained model to generate outcome predictions at specified future assessment intervals. Because all features were aggregated at baseline in this pilot study without time-varying components, we employed a feedforward neural network (FFNN) architecture. The network consisted of an input layer, six hidden layers with residual connections, and an output layer. Hidden layers used Rectified Linear Unit (ReLU) activation functions. Model optimization was performed using the Adam optimizer, and model performance was optimized using cross-entropy loss. To reduce overfitting, dropout regularization with a rate of 0.1 was applied. Hyperparameters, including the number of hidden units and training settings, were empirically selected based on pilot performance and validation results. The model was implemented in Python using the PyTorch deep learning framework.

As a pilot test to assess the effectiveness of AI-concordant regimens, we tested the DNN model in a dataset completely separated from the training data, containing Veterans with MDD who were receiving sertraline, a common SSRI. We identified 41309 Veterans with a randomly selected PHQ-9 survey score >4 as the baseline PHQ-9 score. We classified the treatment regimen that an individual patient received within the 30 days following the baseline PHQ-9 survey into two groups: AI-concordant, if the treatment regimen was predicted by the AI model to reduce PHQ-9 scores, and AI-discordant otherwise (Fig. 1). This approach identified an AI-concordant strategy, not necessarily the most optimal AI-concordant strategy, because we did not test all patients and all possible AI-concordant drug combinations.

4. Results

On the testing set, the mean and median absolute errors were 4.0 and 3.0 points, respectively. We also found that on 73 % of patients in the testing set, the absolute errors were less than 5 points (Fig. 2). The distribution of absolute errors showed that prediction errors were concentrated in the lower range, supporting the feasibility of using the DNN model to predict patient level PHQ-9 outcomes.

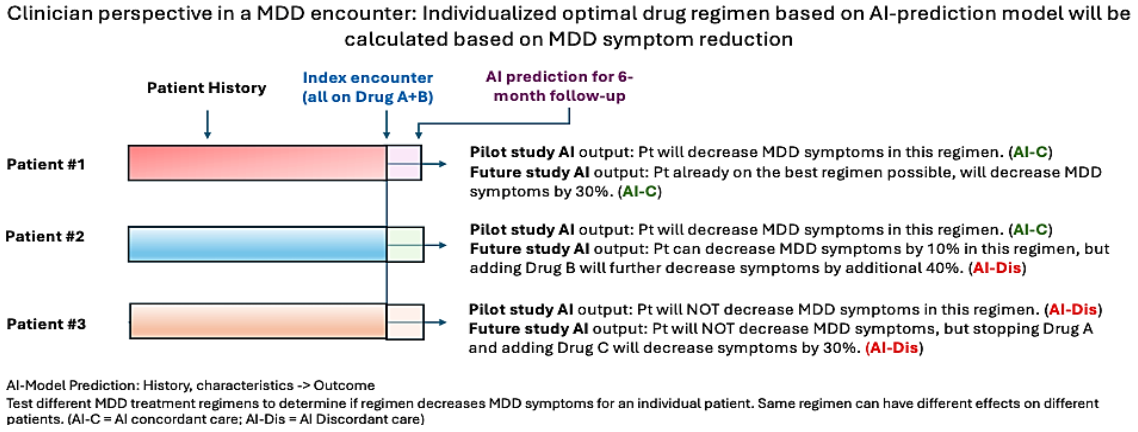


Fig. 1. AI-concordant vs AI-discordant care: Clinician perspective in an MDD encounter where individualized optimal drug regimen will be calculated using AI prediction model based on MDD symptom reduction.

In the pilot test of AI-concordant regimens, 27918 Veterans received AI-concordant regimens and 13391 received AI-discordant regimens. PHQ-9 score reduction was observed in 63.9 % of the AI-concordant group compared with 42.9 % of the AI-discordant group. No change was observed in 7.6 % of the AI-concordant group and 10.0 % of the AI-discordant group, while PHQ-9 score increase was observed in 28.5 % and 47.1 %, respectively (Fig. 3). AI-concordant regimens were associated with greater odds of PHQ-9 score reduction, with an unadjusted odds ratio of 2.35 (95 % CI: 2.25-2.45; $p<0.001$) and a multivariable adjusted odds ratio of 1.17 (95 % CI: 1.10-1.24; $p<0.001$).

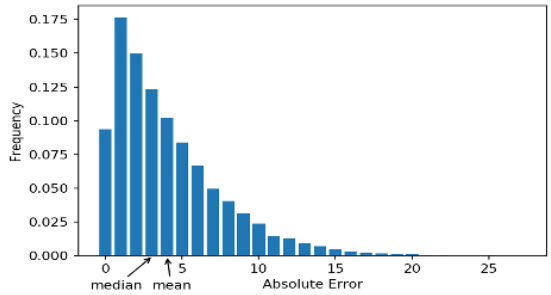


Fig. 2. Absolute prediction errors for follow-up PHQ-9 scores.

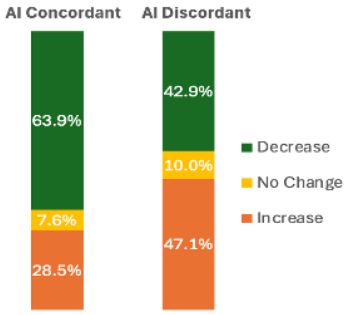


Fig. 3. Outcomes associated with AI-concordant versus AI-discordant regimens.

5. Conclusions

The findings of this study support the feasibility of AI-based MDD treatment response prediction and demonstrate that AI-concordant care is a promising approach toward precision depression treatment. As MDD treatment is inherently complex, clinical decision-making extends beyond treatment initiation or monotherapy selection to include ongoing adjustment of therapeutic strategies over time. Because symptom reduction is the primary indicator of treatment response, we trained an AI model using depressive symptom outcomes as the clinical endpoint. In this study, the model was able to predict treatment response for medication regimens consisting of one or more antidepressants at randomly selected clinical encounters. Furthermore, in an independent sample, we demonstrated that when clinician-prescribed treatments were concordant with model-predicted symptom reduction, patients exhibited significantly greater odds of symptom improvement.

References

[1]. W. Marx, B. Penninx, M. Solmi, T. A. Furukawa, et al., Major depressive disorder, *Nature Reviews Disease Primers*, Vol. 9, Issue 1, 2023, 44.

[2]. National Institute of Mental Health, Major depression 2023, https://www.nimh.nih.gov/health/statistics/major-depression

[3]. Y. Moradi, B. Dowran, M. Sepandi, The global prevalence of depression, suicide ideation, and attempts in the military forces: A systematic review and meta-analysis of cross sectional studies, *BMC Psychiatry*, Vol. 21, Issue 1, 2021, 510.

[4]. L. Citrome, R. Jain, A. Tung, P. B. Landsman-Blumberg, et al., Prevalence, treatment patterns, and stay characteristics associated with hospitalizations for major depressive disorder, *Journal of Affective Disorders*, Vol. 249, 2019, pp. 378-384.

[5]. Centers for Disease Control and Prevention, FastStats – Depression, https://www.cdc.gov/nchs/fastats/depression.htm

[6]. R. Karrouri, Z. Hammani, R. Benjelloun, Y. Otheman, Major depressive disorder: Validated treatments and future challenges, *World Journal of Clinical Cases*, Vol. 9, Issue 31, 2021, pp. 9350-9367.

[7]. S. W. Chung, K. E. Hoy, P. B. Fitzgerald, Theta-burst stimulation: A new form of TMS treatment for depression?, *Depression and Anxiety*, Vol. 32, Issue 3, 2015, pp. 182-192.

[8]. S. H. Kennedy, R. W. Lam, R. S. McIntyre, S. V. Tourjman, et al., Canadian Network for Mood and Anxiety Treatments (CANMAT) 2016 clinical guidelines for the management of adults with major depressive disorder: Section 3. Pharmacological treatments, *The Canadian Journal of Psychiatry*, Vol. 61, Issue 9, 2016, pp. 540-560.

[9]. T. Gkesoglou, S. I. Bargiota, E. Iordanidou, M. Vasiliadis, et al., Prognostic significance of blood-based baseline biomarkers in treatment-resistant depression: A literature review of available studies on treatment response, *Brain Sciences*, Vol. 12, Issue 7, 2022, 940.

[10]. S. Mishra, T. R. Swain, M. Mohanty, Adverse drug reaction monitoring of antidepressants in the psychiatry outpatients department of a tertiary care teaching hospital, *Journal of Clinical and Diagnostic Research*, Vol. 7, Issue 6, 2013, pp. 1131-1134.

[11]. Y. E. Rybak, K. S. P. Lai, R. Ramasubbu, F. Vila-Rodriguez, et al., Treatment-resistant major depressive disorder: Canadian expert consensus on definition and assessment, *Depression and Anxiety*, Vol. 38, Issue 4, 2021, pp. 456-467.

[12]. M. Zhdanava, D. Pilon, I. Ghelerter, W. Chow, et al., The prevalence and national burden of treatment-resistant depression and major depressive disorder in the United States, *The Journal of Clinical Psychiatry*, Vol. 82, Issue 2, 2021, 20m13699.

[13]. E. G. Ostinelli, M. Jaquiery, Q. Liu, R. Elgarf, et al., Personalising antidepressant treatment for unipolar depression combining individual choices, risks and big data: The PETRUSHKA tool, *The Canadian Journal of Psychiatry*, Vol. 70, Issue 10, 2025, pp. 768-781.

[14]. A. Tomlinson, T. A. Furukawa, O. Efthimiou, G. Salanti, et al., Personalise antidepressant treatment for unipolar depression combining individual choices, risks and big data (PETRUSHKA): Rationale and protocol, *Evidence-Based Mental Health*, Vol. 23, Issue 2, 2020, pp. 52-56.

[15]. Y. H. Sheu, C. Magdamo, M. Miller, S. Das, et al., AI-assisted prediction of differential response to antidepressant classes using electronic health records, *npj Digital Medicine*, Vol. 6, Issue 1, 2023, 73.

[16]. F. Alemi, J. Wojtusiak, A. Ursani, K. P. Eklou, et al., Artificial intelligence for management of major depression: Initial design, progress, and research plans, *Alpha Psychiatry*, Vol. 26, Issue 4, 2025, 44608.

(051)

Hierarchical Contrastive Alignment with a Triplet–Focal Objective for Cervical Cytology Classification

Bekhzod Olimov and Sung Wook Ahn
AI Team, Rinorbit, 06082 Seoul, South Korea
E-mail: bekhzod@rinorbit.com

Summary: Two stubborn problems show up whenever one trains a modern deep network on cervical cytology slides. First, the data is heavily skewed toward benign cells, and a plain cross-entropy classifier learns to lean on the majority class. Second, Vision Transformers, which dominate large-scale recognition, have no built-in locality prior and tend to underperform a small CNN on the few hundred slides one usually has access to. We address both with a single training-time recipe, HCA–TFS. The first half (HCA, hierarchical contrastive alignment) attaches a margin-based cosine penalty to the global-pooled feature at each of four backbone stages. The second half (TFS, triplet–focal synergy) replaces cross-entropy at the head with focal loss computed on all three branches of a query/positive/negative triplet. On four datasets (CRIC, PBC, SIPaKMeD, Mendeley) and three architecturally different backbones (ResNet18, ViT-Tiny, MobileNetV3-Large), and reported as mean ± std over 10 random seeds per cell, HCA–TFS improves macro-F1 in 10 of 12 (dataset, backbone) settings, with all-three-backbone wins on SIPaKMeD (+7.4 to +11.3 pp) and PBC (+1.3 to +3.1 pp, std ≤ 0.6), and a +6.1 pp jump on Mendeley for the locality-poor ViT-Tiny. Across all twelve settings HCA–TFS does not significantly degrade any cell. The deployed model is byte-for-byte identical to the cross-entropy baseline – same parameter count, FLOPs, and activation memory at inference – so the gains arrive at zero deployment cost. Hierarchical contrastive supervision behaves, in this regime, like a stand-in for the inductive bias the transformer is missing.

Keywords: Cervical cytology, Contrastive learning, Triplet loss, Focal loss, Vision transformer, Medical image classification.

1. Introduction

Cervical cancer is one of the few cancers with a clear, well-validated screening pathway; the gap between what is medically possible and what happens at the population level is mostly a labour problem. The Pap smear works; reading it does not scale. WHO figures still place cervical cancer as the fourth most common malignancy in women, with most fatal cases in regions short on cytologists [1], and even where staffing is adequate manual review is fatiguing and prone to inter-observer variability [2]. Computer-aided cytology has therefore moved from a curiosity to a real clinical aid in the past decade [3].

The standard recipe – train a CNN with cross-entropy on whatever annotated slides is available – works, but only up to a point. Two failure modes concern us. Class imbalance: clinical corpora are dominated by benign cells, the dysplastic categories that matter (ASC-H, HSIL) are a small minority, and cross-entropy converges to a representation in which the minority margin is narrow [4, 5]. And the awkward fit between Vision Transformers (ViTs) [6] and small medical datasets: with a few hundred to a few thousand images they tend to lose to a CNN of comparable parameter count, because they lack the convolutional priors of locality and translation equivariance and must learn them from scratch – recent papers graft those priors back in through shifted-window attention or deformable token mixing [7, 8].

Most existing remedies attack only one problem at a time. Re-balancing losses such as focal loss [4] fix the gradient at the head but leave the intermediate features alone; contrastive objectives [9-11] reshape the embedding but are usually applied at a single terminal layer [12]. We combine them. HCA–TFS is a per-stage cosine-margin penalty applied at four hierarchical depths plus a focal head that supervises every triplet branch – no extra encoder, no auxiliary classifier, no learnt loss-weight, and it works on top of any backbone with four natural stages.

We make three contributions, each backed by experiments. (i) HCA: a triplet-based cosine-margin penalty at each of four backbone stages, regularising shallow texture, mid-level morphology, and deep semantics together rather than only at the terminal embedding. (ii) TFS: a focal head supervising all three triplet branches, concentrating the gradient on hard cytological outliers without resampling, class weights, or synthetic augmentation. (iii) Multi-paradigm, multi-seed evidence: across four datasets and three architecturally different backbones, reported as mean ± std over 10 random seeds per cell, the recipe improves macro-F1 in 10 of 12 (dataset, backbone) settings – up to +11.3 pp on SIPaKMeD with ResNet18 and +10.1 pp on Mendeley with ResNet18 – and does not significantly degrade any cell. The deployed network is byte-for-byte identical to a plain cross-entropy classifier on the same backbone: same parameter count, FLOPs, and activation memory.

2. Related Work

2.1. Cervical Cytology Pipelines

Automated cytology began with hand-crafted descriptors feeding shallow classifiers; the convolutional turn [13, 14] was a clear step up, but the

residual error on dysplastic-versus-benign classification has shrunk slowly because the diagnostic boundary (ASC-US / LSIL / HSIL / SCC) varies continuously rather than partitioning into visually disjoint classes. Wang et al. [3] pushed back with a generative augmentation pipeline; we share the diagnosis (imbalance is the underlying issue) but prefer to leave the data alone and rework the loss, since synthetic samples can drift from the clinical distribution. We report macro-F1 throughout – the right yardstick for a screening tool whose value lies in the dysplastic minority.

2.2. Triplet and Contrastive Learning

The triplet objective in its modern form goes back to FaceNet [9], and the contrastive family that followed (SimCLR [10], supervised contrastive [11]) is now the default tool for shaping representation geometry. Chen et al. [15] showed that a progressive class-center triplet loss reduces imbalance bias in pathology, and Fuller et al. [12] reported that hierarchical contrastive supervision pays off in few-shot regimes. HCA generalises that observation to a four-stage backbone-agnostic alignment; the new ingredient is the focal classification head, which makes the imbalance and geometry concerns interact rather than just coexist.

2.3. Vision Transformers in Medical Imaging

ViTs [6] are strong on long-range dependencies and weak on small-data inductive bias. Hybrid attempts to inject locality back into ViTs typically rewrite the attention mechanism, through shifted windows or deformable patterns [7, 8]; data-side alternatives such as DeiT distillation or MAE pretraining demand a teacher or unlabeled in-distribution corpus that a clinical lab rarely has. Our angle is the opposite: leave the attention untouched, leave the data alone, and supply the missing prior at training time as a hierarchical contrastive constraint that holds at every depth – a surrogate inductive bias delivered through gradients rather than architecture, so the deployed network is structurally indistinguishable from a plain ViT.

3. Method

We process images in triplets, applying a contrastive penalty at every backbone stage and a focal classifier at the head; the pipeline is summarised in Fig. 1. A query/positive/negative triplet (x_q, x_p, x_n) is passed through one shared backbone, and the four hierarchical stages of that backbone are tapped to produce four global-pooled embeddings per branch. A cosine-margin contrastive loss runs at each of those four depths and a focal classifier supervises all three branches at the terminal embedding. The two arms address adjacent training pathologies in tandem: the contrastive arm shapes embedding geometry from shallow texture through to deep semantics, while the focal arm shifts the classifier gradient onto the hard cytological outliers – the dysplastic minority categories on which macro-F1 lives. Both arms are training-only: the triplet sampler, the per-stage taps, and the cosine penalty are not invoked at inference, so the deployed network has the same parameter count, FLOPs, and activation memory as a plain cross-entropy classifier on the same backbone.

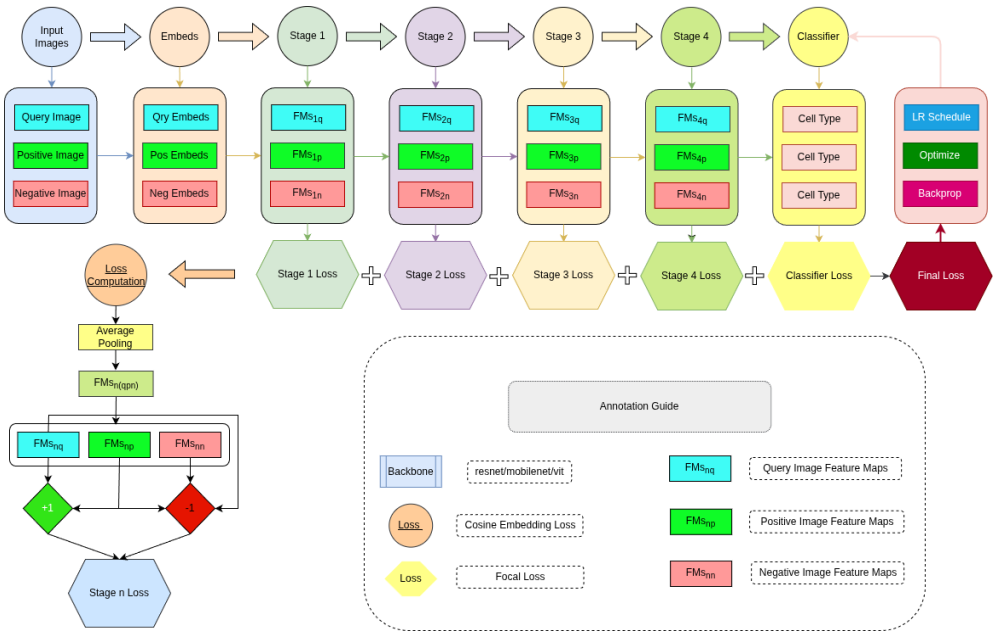


Fig. 1. HCA–TFS pipeline. A triplet (query, positive, negative) runs through a shared backbone whose four hierarchical stages are tapped to produce global-pooled embeddings. A cosine-margin loss aligns same-class pairs and repels cross-class pairs at every stage; the terminal embedding feeds a linear classifier supervised by focal loss on all three branches.

3.1. Triplet Sampling

Let D = {(x_i, y_i)} be the cytology corpus with C classes. For each query x_q with label y_q we draw a positive x_p uniformly from the same class (excluding x_q) and a negative x_n uniformly from the other classes; on a rare singleton-class minibatch we fall back to using the query as its own positive, collapsing that branch's gradient to zero.

3.2. Hierarchical Feature Extraction

Every backbone we use admits a four-stage decomposition – the four ResNet residual stages, the four blocks of the ViT, or the four inverted-bottleneck stages of MobileNetV3. We collapse each stage l by global average pooling to obtain f^(l), implemented via the timm forward_intermediates interface so the same tap code works unchanged across CNNs and transformers.

3.3. Hierarchical Contrastive Alignment

For two vectors u, v write cos(u,v) for their cosine similarity. The per-pair penalty is the margin-based cosine embedding loss:

$$L\_cos(u,v;\, y) = 1 - cos(u,v), \text{ if } y = +1; \quad max(0,\, cos(u,v) - m), \text{ if } y = -1, \tag{1}$$

with margin m controlling inter-class repulsion. HCA applies (1) at every depth, not only at the terminal embedding:

$$L\_HCA = \Sigma_{l=1..4} [\, L\_cos(f\_q^{(l)},\, f\_p^{(l)};\, +1) + L\_cos(f\_q^{(l)},\, f\_n^{(l)};\, -1) \,] \tag{2}$$

Shallow taps push the network toward stain- and acquisition-invariant texture, deep taps toward class-discriminative semantics. Unlike classical deep supervision, HCA needs no auxiliary classification heads.

3.4. Focal Classification Head

A linear classifier maps the terminal embedding to logits with predicted ground-truth probability p_t; the focal loss [4] is:

$$L\_focal = -\alpha\_t\, (1 - p\_t)^{\gamma} \log p\_t \tag{3}$$

We use $\gamma = 2.0$ and $\alpha_t = 1$ throughout (original Lin et al. recipe, not retuned). The classifier supervises all three triplet branches to avoid the asymmetry of query-only supervision.

3.5. Total Objective

The full objective sums (2) and (3) across the triplet branches:

$$L\_TFS = L\_HCA + \Sigma_{i \in \{q,p,n\}} L\_focal(i) \tag{4}$$

We do not introduce a scalar weight between the two terms: both losses are bounded in [0, 1] at the per-sample level, and an early sweep showed the result is insensitive to weights in [0.5, 2]. At inference the contrastive taps are not invoked and the model behaves exactly like a single-branch classifier.

4. Experimental Setup

4.1. Datasets and Backbones

We use four publicly available datasets: CRIC [21] (N = 11,534, cervical cytology with multi-class Bethesda labels and heavy benign skew – the hardest of the four); PBC [17] (N = 17,092, peripheral blood cell classification, a balanced large-scale cellular benchmark); SIPaKMeD [18] (N = 966 cluster-cell images, five-class cervical cytology with high intra-class similarity); and Mendeley [22] (N = 963 liquid-based Pap-smear images, cleanly separable categorical morphology). Each is split 90/5/5 for train/validation/test, with the same per-seed split shared between baseline and HCA–TFS. We exercise three architecturally different backbones to make the cross-paradigm point: ResNet18 [14] (locality-rich CNN), ViT-Tiny/16 [6] (pure transformer, no locality prior), and MobileNetV3-Large [16] (depthwise convolutions with squeeze-and-excitation). All three are initialised from the timm ImageNet weights, with the classifier head reinitialised from scratch.

4.2. Optimisation and Metrics

AdamW [19] (weight decay 1e-4, peak LR 3e-4) under a OneCycleLR schedule [20] (10 % warm-up, cosine annealing). 100 epochs max with early stopping on validation macro-F1 (patience 5); gradients clipped at 1.0; cosine margin m = 0.3; batch size 16; input 224×224; one NVIDIA RTX 4090. We report accuracy, macro-F1, weighted-F1, macro sensitivity, and macro specificity via torchmetrics stateful accumulators. Every reported number is mean ± std over 10 random seeds with the split redrawn independently per seed.

5. Results and Discussion

Table 1 reports the test-set comparison on ViT-Tiny across the four datasets, as mean ± std over 10 random seeds. HCA–TFS lifts accuracy and sensitivity on every dataset; the largest gains land where the baseline is weakest. SIPaKMeD picks up +9.8 pp of accuracy and +9.6 pp of macro-F1 (every one of the 10 seeds favouring HCA–TFS); Mendeley's macro-F1 rises from 92.96 to 99.02 (+6.06 pp), with HCA–TFS reaching the 100 % ceiling on most seeds

(std 3.10) while the baseline is split-sensitive (std 9.95). PBC, the largest and most balanced, sees uniform tight gains (+2.9 accuracy, +3.1 macro-F1) with sub-pp std. CRIC is the exception: macro-F1 nudges up +1.0 pp on average but with ±5.3 std, statistically indistinguishable from zero.

Table 1. ViT-Tiny test-set results (%), mean ± std over 10 random seeds (each seed an independent 90/5/5 split). Acc. = accuracy, F1 (mac/wt) = macro/weighted F1, Sens./Spec. = macro sensitivity/specificity. HCA-TFS rows in bold.

Dataset	Method	Acc.	F1 (mac)	F1 (wt)	Sens.	Spec.
Mendeley	Baseline	97.55 ± 3.16	92.96 ± 9.95	97.03 ± 4.17	92.91 ± 9.36	99.18 ± 1.14
Mendeley	HCA-TFS	**99.80 ± 0.65**	**99.02 ± 3.10**	**99.77 ± 0.73**	**98.75 ± 3.95**	**99.94 ± 0.19**
SIPaKMeD	Baseline	87.55 ± 5.48	87.72 ± 6.03	87.39 ± 5.67	88.05 ± 6.77	96.78 ± 1.44
SIPaKMeD	HCA-TFS	**97.35 ± 2.55**	**97.29 ± 2.69**	**97.37 ± 2.50**	**97.39 ± 2.58**	**99.33 ± 0.63**
CRIC	Baseline	80.87 ± 1.85	61.90 ± 3.43	78.81 ± 2.45	64.59 ± 4.66	94.12 ± 0.57
CRIC	HCA-TFS	**80.78 ± 3.00**	**62.91 ± 4.86**	**79.24 ± 3.18**	**66.11 ± 6.35**	**94.41 ± 0.77**
PBC	Baseline	95.79 ± 0.80	95.52 ± 0.82	95.80 ± 0.79	95.65 ± 0.74	99.39 ± 0.11
PBC	HCA-TFS	**98.69 ± 0.52**	**98.61 ± 0.57**	**98.69 ± 0.52**	**98.56 ± 0.60**	**99.81 ± 0.08**

Table 2 summarises the cross-architecture ablation in macro-F1, mean ± std over 10 seeds. HCA–TFS leads the cross-entropy baseline in 10 of 12 (dataset, backbone) cells; the two remaining cells – ResNet18 + CRIC ($\Delta = -1.17 \pm 9.33$) and MobileNetV3-Large + Mendeley ($\Delta = -0.08 \pm 8.67$) – both have differences well inside their seed-to-seed variance and are no-effect rather than regressions. No cell shows a statistically significant loss. The most consistent gains land on SIPaKMeD (clean win every backbone, +7.4 to +11.3 pp, HCA–TFS std ≤ 2.7) and PBC (+1.3 to +3.1 pp, std ≤ 0.6); the ViT-Tiny + Mendeley jump (+6.06 pp) is where a surrogate inductive-bias prior should help most.

Table 2. Cross-architecture macro-F1 (%) on the held-out test split, mean ± std over 10 random seeds. HCA-TFS rows in bold.

Backbone	Method	Mendeley	SIPaK	CRIC	PBC
ResNet18	Baseline	88.43 ± 9.84	87.52 ± 5.52	61.96 ± 4.69	98.14 ± 0.41
ResNet18	HCA-TFS	**98.50 ± 2.55**	**98.85 ± 1.51**	**60.79 ± 8.22**	**99.46 ± 0.24**
ViT-Tiny	Baseline	92.96 ± 9.95	87.72 ± 6.03	61.90 ± 3.43	95.52 ± 0.82
ViT-Tiny	HCA-TFS	**99.02 ± 3.10**	**97.29 ± 2.69**	**62.91 ± 4.86**	**98.61 ± 0.57**
MobileNetV3-L	Baseline	94.25 ± 6.60	92.49 ± 4.01	59.05 ± 4.71	97.99 ± 0.57
MobileNetV3-L	HCA-TFS	**94.17 ± 6.40**	**99.85 ± 0.47**	**65.05 ± 4.44**	**99.78 ± 0.27**

Two patterns emerge. First, on every backbone HCA–TFS compresses the seed-to-seed std (ResNet18 SIPaKMeD 5.52 → 1.51, MobileNetV3-Large SIPaKMeD 4.01 → 0.47, ViT-Tiny Mendeley 9.95 → 3.10) – the method converts a high-variance baseline into a near-deterministic one, the property that matters most for clinical deployment. Second, CRIC's flat result is variance-dominated rather than directional: two of its six classes are ultra-rare (SCC: 2–8 test images per seed), so on some seeds HCA–TFS drives SCC recall to 0 and on others to between 78 and 100, and the macro statistic picks up that toggling as ±9 pp of noise. A class-aware focal α_t is the natural follow-up. Fig. 2 shows the validation trajectories: HCA–TFS sits on or above the baseline in nearly every panel, and the visible gap widens on the small-data columns (Mendeley, SIPaKMeD).

A practical point: everything HCA–TFS introduces lives at training time. Once training finishes, the deployed network is byte-for-byte identical to a plain cross-entropy classifier on the same backbone: same parameter count, FLOPs, activation memory, checkpoint format, and ONNX export. A clinical pipeline that already runs the baseline in real time will run HCA–TFS in real time too.

6. Limitations and Project Context

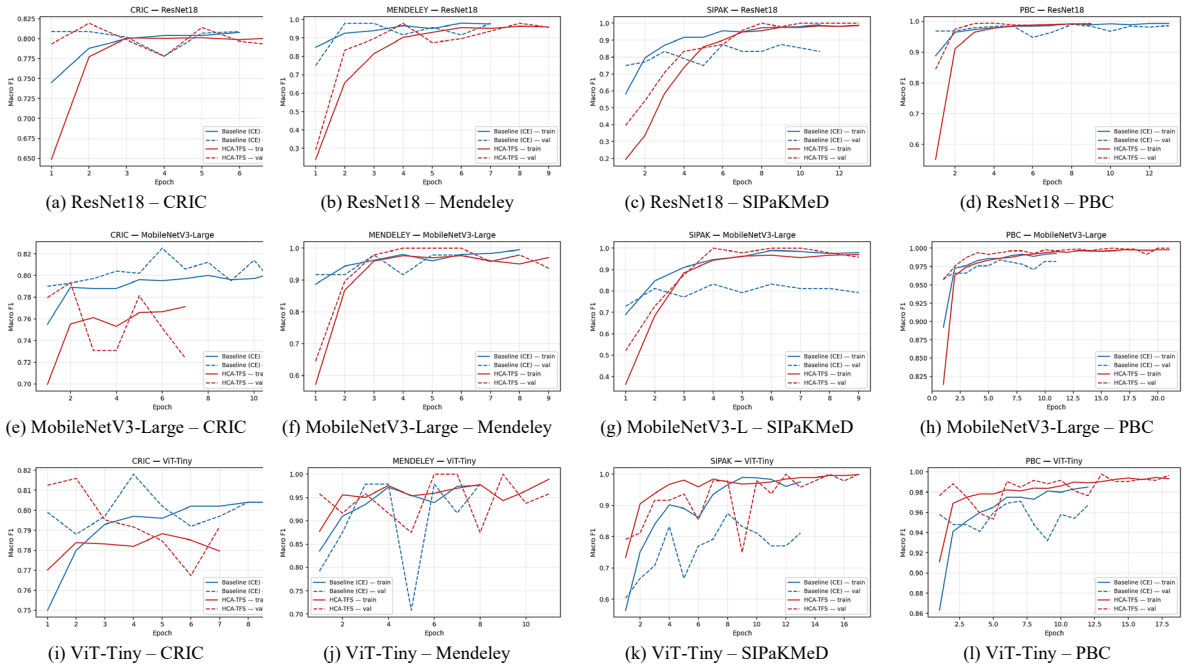
Several limitations should be flagged directly. The held-out test partitions of Mendeley and SIPaKMeD contain only 49 examples each per seed, so the precise Mendeley gap (+10.07 ± 11.03 ResNet18, +6.06 ± 11.13 ViT-Tiny) should be read as a robust positive direction rather than a precisely measured pp value; replication on larger splits is the natural next step. The comparison is against a tuned cross-entropy baseline on the same backbone rather than against published cytology-specific architectures, no experiment tests cross-cohort distribution shift, and only ViT-Tiny is exercised on the transformer side.

This paper is one component of a larger project. The authors' affiliated company is jointly developing a custom cervical-screening imaging device and a

three-stage AI pipeline running on top of it: cell-level detection, instance-level segmentation, and the multi-class classifier described here. Each stage is a self-contained contribution and is being published separately so that each is reviewable on its own merits. As part of the same project, we are preparing the public release of a new cervical-cytology corpus collected with our in-house device; at the April 2026 curation cut-off it already exceeds, in slide count and dysplastic-class coverage, every publicly available cervical-cytology dataset we are aware of, and we expect it to be the largest such resource at release.

7. Conclusion

HCA–TFS combines two ideas – hierarchical contrastive alignment and a focal classification head – at the right point in the network. Applied at four hierarchical taps, the combination behaves as a stand-in for the inductive bias a Vision Transformer lacks on small medical datasets, large enough to move ViT-Tiny from a clear underperformer to a clear winner in the SIPaKMeD regime. Across 10 random seeds on each of 12 (dataset, backbone) cells the recipe wins on 10, lands within seed-to-seed variance on the remaining 2 (no-effect rather than regression), and shows a statistically significant loss on zero – all at zero inference-time cost. Three follow-ups are clear: a class-aware focal α_t for the ultra-rare-class regime of CRIC, stage-wise hard-negative mining in place of uniform triplet sampling, and a cross-cohort distribution-shift study before clinical use.



(a) ResNet18 – CRIC (b) ResNet18 – Mendeley (c) ResNet18 – SIPaKMeD (d) ResNet18 – PBC

(e) MobileNetV3-Large – CRIC (f) MobileNetV3-Large – Mendeley (g) MobileNetV3-L – SIPaKMeD (h) MobileNetV3-Large – PBC

(i) ViT-Tiny – CRIC (j) ViT-Tiny – Mendeley (k) ViT-Tiny – SIPaKMeD (l) ViT-Tiny – PBC

Fig. 2. Training (solid) and validation (dashed) macro-F1 across all twelve (dataset, backbone) combinations, shown for one representative seed per panel (the reported test-set numbers in Tables 1 and 2 are the mean ± std over 10 such seeds). Rows are backbones – ResNet18 (a–d), MobileNetV3-Large (e–h), ViT-Tiny (i–l); columns are datasets – CRIC, Mendeley, SIPaKMeD, PBC, left to right. Blue is the cross-entropy baseline, red is HCA–TFS; curves terminate at different epochs because of early stopping on validation macro-F1 (patience 5). HCA–TFS reaches a higher validation plateau on every panel, and the margin is widest where the backbone's inductive bias is weakest – the ViT-Tiny row and the small-data Mendeley and SIPaKMeD columns.

Acknowledgements

This work was supported by Rinorbit (Seoul, South Korea) as part of an internal R&D programme on AI-assisted cervical screening.

Data Availability and Ethics

All four datasets are publicly available [21, 17, 18, 22] and contain no personally identifiable information; the 90/5/5 split protocol and per-seed test partitions are deterministic.

Code Release

The training and inference code and trained checkpoints are not released with this submission, as the implementation forms part of a product pipeline in active development. Conditional on publication, the authors commit to releasing the HCA–TFS pipeline under a permissive open-source license at the camera-ready stage; reviewers may request a private package through the programme chairs under NDA.

References

[1]. World Health Organization, Cervical cancer, 2024, https://www.who.int/news-room/fact-sheets/detail/cervical-cancer

[2]. E. Bengtsson, P. Malm, Screening for cervical cancer using automated analysis of PAP-smears, *Computational and Mathematical Methods in Medicine*, Vol. 2014, 2014, 842037.

[3]. X. Wang, C. Li, C. Li, M. Wang, Improving cervical cancer classification with imbalanced datasets via generative adversarial networks, *BMC Medical Imaging*, Vol. 22, Issue 1, 2022, 53.

[4]. T.-Y. Lin, P. Goyal, R. Girshick, K. He, et al., Focal loss for dense object detection, in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2980-2988.

[5]. K. Cao, C. Wei, A. Gaidon, N. Arechiga, et al., Learning imbalanced datasets with label-distribution-aware margin loss, in Advances in Neural Information Processing Systems 32 (H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, et al., Eds.), *Curran Associates, Inc*., Red Hook, 2019, pp. 1567-1578.

[6]. A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, et al., An image is worth 16x16 words: Transformers for image recognition at scale, in *Proceedings of the 9th International Conference on Learning Representations (ICLR)*, 2021.

[7]. J. Ha, E. Yoon, S. Kim, J. Kim, et al., Leveraging inductive bias in ViT for medical image diagnosis, in *Proceedings of the 35th British Machine Vision Conference (BMVC)*, 2024.

[8]. H. Alquran, A. M. Alqudah, I. Abu-Qasmieh, A. Al-Badarneh, et al., Enhancing cervical pre-cancerous classification using advanced Vision Transformer, *Diagnostics*, Vol. 13, Issue 18, 2023, 2884.

[9]. F. Schroff, D. Kalenichenko, J. Philbin, FaceNet: A unified embedding for face recognition and clustering, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 815-823.

[10]. T. Chen, S. Kornblith, M. Norouzi, G. E. Hinton, A simple framework for contrastive learning of visual representations, in *Proceedings of the 37th International Conference on Machine Learning (ICML)*, 2020, pp. 1597-1607.

[11]. P. Khosla, P. Teterwak, C. Wang, A. Sarna, et al., Supervised contrastive learning, in Advances in Neural Information Processing Systems 33 (H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, et al., Eds.), *Curran Associates, Inc*., Red Hook, 2020, pp. 18661-18673.

[12]. H. Fuller, F. G. Garcia, V. Flores, Efficient few-shot medical image analysis via hierarchical contrastive vision-language learning, *arXiv*, 2025, arXiv:2501.09294.

[13]. A. Krizhevsky, I. Sutskever, G. E. Hinton, ImageNet classification with deep convolutional neural networks, in Advances in Neural Information Processing Systems 25 (F. Pereira, C. J. Burges, L. Bottou, K. Q. Weinberger, Eds.), *Curran Associates, Inc*., Red Hook, 2012, pp. 1097-1105.

[14]. K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770-778.

[15]. K. Chen, M. Xu, J. Li, Y. Zhang, et al., PCCT: Progressive class-center triplet loss for imbalanced medical image classification, *IEEE Journal of Biomedical and Health Informatics*, Vol. 27, Issue 8, 2023, pp. 3864-3875.

[16]. A. Howard, M. Sandler, G. Chu, L.-C. Chen, et al., Searching for MobileNetV3, in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 1314-1324.

[17]. A. Acevedo, A. Merino, S. Alférez, A. Molina, et al., A dataset of microscopic peripheral blood cell images for development of automatic recognition systems, *Data in Brief*, Vol. 30, 2020, 105474.

[18]. M. E. Plissiti, P. Dimitrakopoulos, G. Sfikas, C. Nikou, et al., SIPaKMeD: A new dataset for feature and image based classification of normal and pathological cervical cells in Pap smear images, in *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, 2018, pp. 3144-3148.

[19]. I. Loshchilov, F. Hutter, Decoupled weight decay regularization, in *Proceedings of the 7th International Conference on Learning Representations (ICLR)*, 2019.

[20]. L. N. Smith, A disciplined approach to neural network hyper-parameters: Part 1 - learning rate, batch size, momentum, and weight decay, Technical Report 5510-026, *US Naval Research Laboratory*, 2018.

[21]. Berkeley Institute for Data Science, CRIC cervix collection, https://bids.berkeley.edu/cric-cervix-collection

[22]. E. Hussain, Liquid based cytology Pap smear images for multi-class diagnosis of cervical cancer, *Mendeley Data*, V4, 2019, 10.17632/zddtpgzv63.4.

(055)

CAPF: A Clinical Agent Permission Framework for HIPAA-Aligned Least-Privilege Authorization in Multi-Agent Healthcare AI Systems

R. Khanna and J. Nandal
Amazon.com, Inc., Seattle, WA, USA
E-mail: rajatkhanna1999@gmail.com, jatinnandal77@gmail.com

Summary: When a clinical AI agent can read a patient's oncology record, place a medication order, and call an external API—all within a single reasoning pipeline—per-permission scoping becomes a practical safety and governance requirement that no existing framework provides at the clinical application layer. Current deployments routinely provision agents with blanket access to entire EHR databases, unrestricted API calls, and unconstrained inter-agent messaging—far exceeding what any single clinical task requires. While the HIPAA minimum-necessary standard (45 CFR §164.514(d)) does not apply to treatment-purpose uses and disclosures, it governs the many non-treatment functions agents perform, and the §164.312 technical safeguards apply to all protected health information (PHI) processing regardless of purpose. We present CAPF (Clinical Agent Permission Framework), among the first clinically grounded frameworks to map healthcare AI-agent tool permissions to HIPAA-aligned least-privilege controls. CAPF defines eight permission classes for medical AI agents, maps each to the relevant HIPAA requirements, and assigns runtime enforcement tiers based on task risk and clinical context. Using three published adversarial attack scenarios, we show analytically that CAPF's permission scoping would have structurally constrained the attack surface in each case.

Keywords: Agentic AI, Healthcare security, HIPAA, Least-privilege authorization, Multi-agent systems, Clinical decision support, AI governance.

1. Introduction

Agentic AI systems—large language models (LLMs) equipped with tool-use capabilities, memory, and inter-agent communication—are rapidly entering clinical healthcare environments. A 2026 systematic survey covering 49 studies identified deployment of LLM-based agents for EHR analysis, differential diagnosis, treatment planning, medication management, and clinical documentation [1]. Unlike traditional software, these agents operate with broad action spaces: they can read and write patient records, execute code, call external services, and dispatch subtasks to other agents—all with minimal human oversight.

This operational breadth creates a fundamental security and governance problem. The HIPAA Privacy Rule's minimum-necessary standard (45 CFR §164.514(d)) requires covered entities to make reasonable efforts to limit PHI use, disclosure, and requests to the minimum needed for the intended purpose. The standard does not apply to disclosures to, or requests by, a health care provider for treatment purposes; however, it does apply to the many non-treatment functions clinical agents perform (documentation for operations or payment, analytics, and secondary use), and covered entities must still implement role-based access policies that limit internal access to the minimum necessary [§164.514(d)(2)] [8]. Independently, the §164.312 technical safeguards—access control, audit controls, integrity, and transmission security—apply to all PHI processing regardless of purpose [9]. Yet current clinical AI agent deployments routinely provision agents with blanket access to entire EHR databases, unrestricted API-calling capabilities, and unconstrained inter-agent messaging—permissions far beyond what any single clinical task requires and in tension with these access-minimization and technical-safeguard expectations.

This over-provisioning creates an exploitable attack surface. Qiu et al. demonstrated that adversarial prompts embedded in web content can redirect medical agents to exfiltrate patient conversations, manipulate recommendations, and hijack host systems [2]. Bashir et al. showed that coordinated collusion among adversarial assistant agents can drive a trusted clinical decision agent toward harmful medication recommendations, achieving a 100 % attack success rate under maximum adversarial coordination and without any defense, across 50 clinical questions [3]. Maiti documented six distinct attack surfaces in a production deployment of nine healthcare AI agents and proposed infrastructure-level defenses [4]. Critically, Cartagena and Teixeira demonstrated empirically that text-level safety alignment does not transfer to tool-call safety: across six frontier models and six regulated domains, agents whose text output refused a harmful request simultaneously executed the forbidden action through tool calls in 219 persistent cases under safety-reinforced system prompts [22]. This text–action divergence makes architectural enforcement at the tool-call boundary—not model alignment alone—a necessary design requirement. OWASP ranks prompt injection, which scales in impact proportionally with agent permissions, as the top vulnerability in LLM applications [10].

Existing frameworks fall short in three ways. Infrastructure-level approaches [4] address kernel isolation and network egress but do not define what application-layer actions each agent should be

permitted to perform relative to clinical context. HIPAA-focused frameworks [5] address PHI sanitization and attribute-based data access but do not scope tool-use permissions in multi-agent pipelines. General-purpose agent authorization specifications [6], [7] are agnostic to clinical workflows and HIPAA requirements. None of these works provides a clinical-specific permission taxonomy covering the full action space of healthcare agents, mapped to regulatory requirements, with a runtime enforcement protocol.

We address this gap with CAPF (Clinical Agent Permission Framework), which contributes: (1) an eight-class clinical permission taxonomy covering the complete tool-use action space of medical AI agents; (2) a HIPAA mapping that aligns each permission class with 45 CFR §164.514(d) requirements and §164.312 technical safeguards [9]; (3) a three-tier runtime enforcement protocol that assigns permission enforcement levels dynamically based on task category and clinical risk; and (4) an analytical security analysis demonstrating how CAPF's permission scoping would have structurally limited three published attack scenarios against medical AI agents. CAPF is aligned with the NIST AI Agent Standards Initiative (announced February 2026), which identifies agent identity, authorization, and security as priority standardization areas [11], and with the FDA's AI/ML Action Plan for Software as a Medical Device [12].

2. Background and Related Work

2.1. Agentic AI in Clinical Environments

Clinical AI agents extend LLMs with tool-use, multi-step reasoning, and inter-agent communication. Vatsal et al. identified a critical asymmetry in the deployment landscape: information-retrieval tasks (medical QA, EHR summarization) are well studied, while action-oriented tasks (treatment planning, prescription, order placement) remain underserved—precisely because the field lacks governance frameworks for safe autonomous action [1]. Frameworks such as MedAgents [15] and AgentClinic [16] are beginning to address these higher-stakes action-oriented tasks, increasing the urgency of principled authorization controls before wider clinical adoption.

2.2. Known Attacks on Medical AI Agents

Three classes of attacks have been documented against clinical AI agents. (1) Prompt injection and web-content hijacking: Qiu et al. demonstrated indirect prompt injection via web browsing that can redirect agent behavior, exfiltrate PHI, and execute malicious URLs without user awareness [2]. (2) Multi-agent collusion: Bashir et al. showed that adversarial assistant agents can build false consensus to manipulate a trusted decision-making agent, achieving attack success rates up to 100 % under maximum adversarial coordination across 50 clinical questions when no defense is present [3]. (3) Identity spoofing and privilege escalation: Maiti documented that agents can be deceived into accepting instructions from unauthorized principals, executing operations on clinical data systems that violate access-control intent [4].

2.3. Existing Authorization and Security Frameworks

Maiti proposed a zero-trust architecture using kernel-level container isolation, credential-proxy sidecars, and network-egress filtering for nine production healthcare agents [4]. This addresses infrastructure security but does not define application-layer permission classes aligned with HIPAA. Neupane et al. presented a HIPAA-compliant agentic AI system using Attribute-Based Access Control (ABAC) for PHI field-level governance, BERT-based PHI sanitization, and immutable audit trails for clinical LLM tasks [5]. This governs data access but not the full tool-use action space of autonomous agents (code execution, agent delegation, external network calls). Zhu et al. defined the OpenPort Protocol (OPP), a governance-first specification for AI agent tool access with least-privilege authorization, ABAC-style policy constraints, and write-operation risk gating [6]. OPP is domain-agnostic and does not address clinical context, HIPAA compliance, or multi-agent delegation in healthcare. Fleming et al. introduced an LLM-judged Task-Based Access Control (TBAC) model that synthesizes just-in-time policies for novel agent tasks and escalates high-risk requests to human review [7]; like OPP, this is not adapted to clinical workflows or regulatory requirements.

Agostino and D'Souza introduced the ALARA (As Low As Reasonably Achievable) principle as a least-privilege context-engineering strategy for composable multi-agent teams [17]. ALARA addresses context engineering (what information is passed to the agent's reasoning layer), not permission authorization (what tool-use actions the agent may execute); the two are orthogonal and complementary. Uchibeke introduced the Open Agent Passport (OAP), a domain-agnostic pre-action authorization layer that intercepts tool calls, evaluates them against a declarative policy, and produces cryptographically signed audit records, reporting 0 % social-engineering success in a live adversarial testbed [21]. CAPF extends this pre-action paradigm to the clinical domain by grounding permission classes in HIPAA requirements and clinical workflow context, which OAP does not address. Kaptein, Khan, and Podstavnychy formalize runtime governance as deterministic policies over agent execution paths [23]; that framework operates at the sequence level, whereas CAPF operates at the individual tool-call permission level with HIPAA regulatory mapping.

Gap: Pre-action authorization, runtime governance, agent tool-call safety, HIPAA-compliant agentic AI, and least-privilege agent control have each

been addressed in recent work [5], [6], [7], [21], [23]. However, to our knowledge no prior work combines a clinically grounded, HIPAA-mapped permission taxonomy covering the complete tool-use action space of medical AI agents with a runtime enforcement protocol. CAPF is among the first clinically grounded frameworks to do so.

3. The CAPF Framework

3.1. Threat Model

We model an attacker as any entity—external adversary, compromised agent, or malicious data source—that can influence an agent's input context (prompt, retrieved document, inter-agent message, or web content) with the goal of inducing the agent to perform an action exceeding its legitimate authorization scope. *Assets:* PHI in EHR systems; clinical orders (medications, procedures); audit trails; inter-agent trust relationships; external network connectivity. *Attacker capabilities:* (1) inject adversarial content into the agent's reasoning context; (2) impersonate authorized principals in inter-agent communication; (3) coordinate with multiple agents to create false consensus. *Attacker goals:* PHI exfiltration; unauthorized clinical order placement; lateral movement across the agent network; audit evasion. *Security objective:* even if an agent's reasoning is compromised, the runtime permission layer must structurally prevent tool-use actions exceeding the agent's assigned permission scope for its current task context.

Insider-threat note: clinical staff with legitimate credentials who over-provision agent access is a parallel concern not modeled by external-adversary frameworks. CAPF addresses this through its per-task permission-scoping model: permission sets are assigned at task instantiation time rather than as blanket role grants, so even a privileged administrator cannot provision an agent with more permissions than the declared task type requires without modifying the task definition itself—an auditable and policy-controlled operation.

3.2. Clinical Permission Taxonomy

We define eight permission classes covering the complete action space of healthcare AI agents, following the Attribute-Based Access Control (ABAC) model [18] in which permissions are scoped to subject (agent), action (permission class), and environment context (task type and patient scope). Each class identifies the actions it grants, the minimal clinical justification required, and the associated attack surface if over-provisioned. Table 1 summarizes the taxonomy; the tier column (T1–T3) refers to the enforcement level defined in Section 3.4. WRITE-PHI-ORD and EXEC-CODE are designated Tier 3 (mandatory human review) because unauthorized execution of these permissions creates life-safety risks that cannot be recovered post-facto.

3.3. HIPAA Minimum-Necessary Mapping

45 CFR §164.514(d) requires covered entities to (1) identify personnel and systems that need PHI access; (2) define the category of PHI required for each role; and (3) make reasonable efforts to limit access consistently [8]. CAPF operationalizes this requirement at the permission-class level for AI agents. We note that the minimum-necessary standard does not govern treatment-purpose uses and disclosures; CAPF therefore treats the minimum-necessary mapping as (i) binding for the non-treatment functions agents perform and for internal role-based access limitation under §164.514(d)(2), and (ii) a recommended best practice for treatment-purpose tasks. The §164.312 technical safeguards in Table 2 apply to all PHI processing regardless of purpose. Table 2 maps each permission class to the specific HIPAA provision it touches, enabling clinical AI deployment teams to use CAPF directly as a compliance checklist.

Table 1. CAPF clinical permission taxonomy.

Permission	Actions granted	Over-priv. risk	Tier
`READ-PHI-DEMO`	Read demographic PHI: name, DOB, address, insurance ID	Identity theft; social engineering	T1
`READ-PHI-CLIN`	Read clinical PHI: diagnoses, labs, medications, imaging reports	Targeted exfiltration; inference attacks	T2
`WRITE-PHI-NOTE`	Write / update clinical notes and documentation	Record falsification; liability	T2
`WRITE-PHI-ORD`	Place clinical orders: medications, procedures, referrals	Unauthorized prescription; patient harm	T3
`EXEC-CODE`	Execute code, scripts, or system commands	System compromise; data destruction	T3
`AGENT-DELEGATE`	Send / receive task delegations to / from other agents	Collusion; privilege escalation	T2
`EXT-NETWORK`	Make HTTP requests to external endpoints	Data exfiltration; injection via web content	T3
`AUDIT-READ`	Read access and audit logs	Attack reconnaissance; log-evasion planning	T2

Table 2. CAPF permissions mapped to HIPAA requirements.

Permission	HIPAA ref.	Compliance requirement
READ-PHI-DEMO	§164.514(d)(2)	Demographic only; clinical data excluded unless separately authorized
READ-PHI-CLIN	§164.514(d)(2),(3)	Scoped to fields required for the specific diagnostic / treatment task
WRITE-PHI-NOTE	§164.514(d)(4); §164.312(b)	Audit trail required; authorship attribution mandatory
WRITE-PHI-ORD	§164.514(d)(4); §164.312(a)(1)	Clinician identity verification; unique identifier logged
EXEC-CODE	§164.312(a)(1),(c)(1)	Integrity risk (not PHI disclosure); audit logging mandatory
AGENT-DELEGATE	§164.514(d)(3)	BAA scope governs agent-to-agent PHI transfer
EXT-NETWORK	§164.312(e)(1),(e)(2)	Transmission security required; endpoints must be allowlisted
AUDIT-READ	§164.312(b)	Log access must be audited; separation of privilege enforced

3.4. Three-Tier Runtime Enforcement Protocol

CAPF assigns each agent a permission set scoped to its current task context (task type, patient record in scope, and session role). At runtime, every tool call is intercepted by the CAPF authorization layer, which applies one of three enforcement tiers.

Tier 1 – Full autonomy: the action falls within the agent's pre-approved permission set for the current task context; no additional check is required. Applied to READ-PHI-DEMO for agents performing patient-identification tasks.

Tier 2 – Supervised autonomy: the action requires the agent to log a structured rationale (task ID, permission class, justification) to the immutable audit trail before execution. The audit event is available for real-time monitoring but does not block execution. Applied to READ-PHI-CLIN, WRITE-PHI-NOTE, AGENT-DELEGATE, and AUDIT-READ.

Tier 3 – Mandatory review: the action requires explicit clinician or system-administrator approval before execution. The agent generates a structured approval request including the specific permission sought, the clinical justification, the data or endpoint affected, and the expected outcome. Applied to WRITE-PHI-ORD, EXEC-CODE, and EXT-NETWORK.

Tier assignment is static per-permission by default (Table 1) but can be escalated dynamically based on three deterministically observable structural conditions: (1) an agent requests access to an endpoint not present in its pre-provisioned allowlist; (2) a delegation-chain depth exceeds the maximum configured hops for the session; or (3) a permission class not in the agent's task-scoped set is requested. When any of these conditions is detected, any Tier 1 or Tier 2 action for that session is escalated to Tier 3 pending human review. These conditions are evaluated structurally at the action boundary without relying on content-based prompt analysis, so the escalation mechanism does not depend on solving the open problem of adversarial-prompt detection. This structural escalation addresses the indirect prompt-injection and privilege-escalation attack classes documented in [2], [4].

For AGENT-DELEGATE actions specifically, CAPF enforces privilege non-escalation at each delegation boundary: upon receiving a delegation request, the CAPF layer validates that the receiving agent's permission set is a strict subset of the delegating agent's current task-scoped set. Any delegation request that attempts to confer out-of-scope permissions is rejected and logged as a Tier 3 escalation event, preventing colluding agents from bootstrapping elevated permissions through chained delegation.

4. CAPF Architecture

Fig. 1 illustrates the CAPF architecture. The CAPF Authorization Layer sits between the Task Orchestrator Agent and all downstream tool-calling agents. It intercepts every tool invocation, consults the permission taxonomy and HIPAA policy engine, assigns an enforcement tier, and either (a) allows the action with logging (Tier 1/2) or (b) creates an approval request and suspends execution pending human confirmation (Tier 3). The immutable audit trail records all authorization decisions with task context, permission class, tier applied, and outcome.

The architecture integrates with existing agentic frameworks. In LangGraph [13], the CAPF layer maps to a conditional routing node that intercepts ToolNode invocations. In Microsoft AutoGen [14], it maps to a middleware function registered in the agent's tool-calling pipeline. In AWS Bedrock Agents, it maps to a Lambda function in the action-group routing layer. No modification to the underlying LLM or agent reasoning is required; CAPF operates orthogonally to the agent's reasoning process, enforcing constraints at the action boundary.

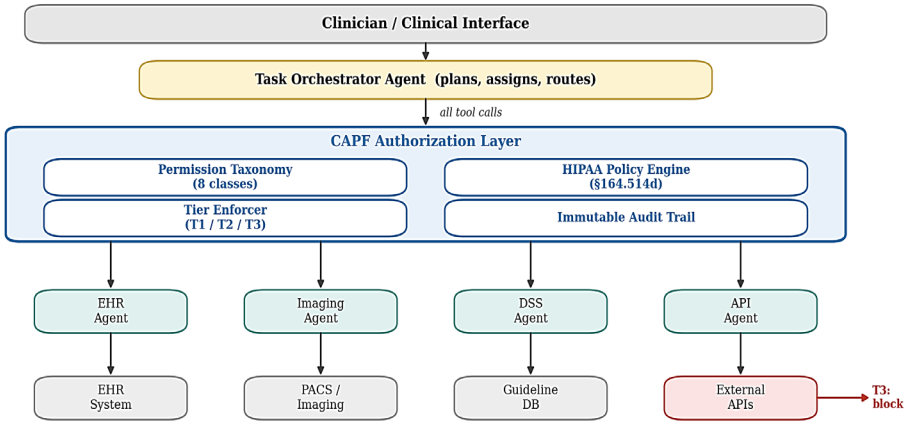


Fig. 1. CAPF system architecture. All tool-use actions from the Task Orchestrator pass through the CAPF Authorization Layer, which enforces the permission taxonomy (8 classes), applies the HIPAA policy engine, assigns the enforcement tier (T1–T3), and logs to an immutable audit trail. High-risk actions (EXT-NETWORK, WRITE-PHI-ORD, EXEC-CODE) require Tier 3 approval before execution.

5. Analytical Security Analysis

We evaluate CAPF analytically against three published attack scenarios drawn from the peer-reviewed and preprint literature on medical AI agent security. We do not fabricate experimental results; rather, we trace each attack's mechanism through the CAPF permission model to identify the specific structural constraint that would limit its impact at the action boundary, independently of the agent's reasoning layer.

5.1. Attack A1: PHI Exfiltration via Web-Content Injection

Attack. Qiu et al. demonstrated that adversarial prompts embedded in web pages retrieved by a medical agent's browsing tool can redirect the agent to (a) return PHI from prior conversation turns to an external URL and (b) execute malicious system commands on behalf of the attacker [2].

CAPF response. Under CAPF, access to external URLs requires the EXT-NETWORK permission (Tier 3). A medical agent performing an EHR-summarization task would hold only READ-PHI-CLIN (Tier 2) and WRITE-PHI-NOTE (Tier 2); it would hold no EXT-NETWORK permission. Any tool call targeting an external URL would be intercepted and suspended pending Tier 3 human approval. Because the attack requires the agent to autonomously access adversary-controlled URLs, CAPF's permission scoping breaks the attack chain at the network-egress step without requiring detection of the adversarial prompt content itself.

5.2. Attack A2: Multi-Agent Collusion toward Harmful Prescription

Attack. Bashir et al. showed that adversarial assistant agents can collude to build false consensus, driving a trusted "AI doctor" agent to recommend harmful medications across 50 clinical questions, with attack success rates up to 100 % under maximum adversarial coordination when no defense is present [3].

CAPF response. Under CAPF, each agent in a multi-agent clinical network holds a scoped AGENT-DELEGATE permission (Tier 2) that restricts (a) which agents it may receive delegations from and (b) which permission classes can be conferred via delegation. CAPF enforces privilege non-escalation at each delegation boundary: the receiving agent's permission set must be a strict subset of the delegating agent's task-scoped set, with any out-of-scope delegation rejected and logged as Tier 3 (Section 3.4). Crucially, the final recommendation step—WRITE-PHI-ORD for placing a medication order—requires Tier 3 clinician approval regardless of the agent's reasoning content. The collusion attack succeeds by corrupting reasoning; CAPF does not prevent corrupted reasoning but prevents it from being acted upon without human review at the critical write-order step.

5.3. Attack A3: Identity Spoofing for Unauthorized EHR Write

Attack. Maiti documented that healthcare agents can be deceived into accepting instructions from spoofed principals, causing the agent to execute write operations on clinical data systems for unauthorized parties [4].

CAPF response. Under CAPF, every Tier 2 and Tier 3 action requires a structured rationale logged to the immutable audit trail, including the task ID and session role of the authorizing principal. WRITE-PHI-NOTE and WRITE-PHI-ORD are Tier 2 and Tier 3 respectively. Any write executed under a spoofed identity produces an audit event with a mismatched session role, enabling post-hoc detection. More importantly, WRITE-PHI-ORD (Tier 3) suspends

execution pending approval from a verified clinician identity, breaking the attack at the action boundary for the highest-risk writes.

5.4. Summary

Table 3 summarizes the attack evaluations. CAPF's structural constraint operates at the permission level, not the reasoning level, making it complementary to content-based defenses (prompt sanitization, adversarial-prompt detection) rather than competitive with them.

6. Implementation Considerations

Permission provisioning. Clinical administrators define each agent's permission set as a JSON policy document mapping to the CAPF taxonomy classes in Table 1. Below is a discharge-planning agent policy: it holds read and note-write access but is denied orders, code execution, external network, and delegation.

```
{"agent_id":"discharge-planner-v1",
 "task_type":"discharge_planning",
 "patient_scope":"ward_B_admitted",
 "permissions":["READ-PHI-DEMO",
   "READ-PHI-CLIN","WRITE-PHI-NOTE"],
 "denied":["WRITE-PHI-ORD","EXEC-CODE",
   "EXT-NETWORK","AGENT-DELEGATE"],
 "tier_overrides":{}}
```

If the agent attempts EXT-NETWORK (e.g., via an injected prompt), CAPF rejects the call, logs it, and escalates the session to Tier 3. The `patient_scope` field resolves against a SMART on FHIR [19] `patient/$` compartment endpoint, layering CAPF atop an existing FHIR authorization server.

Immutable audit trail. All decisions are written to an append-only log containing timestamp, session ID, agent ID, task type, permission class, tier applied, rationale (Tier 2), and approval status with approver ID (Tier 3). Tamper-evidence uses SHA-256 hash chaining per NIST SP 800-92 [20], satisfying HIPAA §164.312(b) [9].

Latency and standards alignment. Tier 1 decisions add negligible overhead (permission lookup only); Tier 2 adds a log write; Tier 3 latency is an operational choice with configurable timeouts, consistent with FDA high-risk AI/ML guidance [12]. By default, CAPF fails closed on Tier 3 timeout: the pending action is denied, the session is flagged for review, and the denial is written to the immutable audit trail—consistent with the principle that patient safety takes precedence over operational continuity. CAPF addresses the authorization and identity axes of NIST's AI Agent Standards Initiative [11] via per-task permission sets and principal attribution in every Tier 2/3 audit event.

7. Discussion and Limitations

Scope and evaluation. CAPF is a framework paper without empirical evaluation on live deployments. The analysis in Section 5 traces published attack mechanisms through the CAPF model analytically; it is not an experimental benchmark. Empirical evaluation on real or simulated multi-agent clinical pipelines is the primary future-work direction. The eight permission classes cover the current action space; novel capabilities (e.g., surgical robotics) may require additional classes, added without changing the tier-enforcement mechanism.

Complementary defenses and adoption. CAPF constrains what an agent can do, not what it reasons. A corrupted agent holding Tier 1 permissions can still return false information within those permissions; CAPF must be deployed alongside reasoning-layer defenses (output verification, retrieval-provenance tracking). It also requires organizational discipline: granting all eight permissions to every agent effectively disables the framework, so least-privilege provisioning is a deployment prerequisite.

Table 3. CAPF analytical security analysis summary.

Atk	Mechanism	CAPF constraint	Effect
A1	Web-injection exfiltration [2]	EXT-NETWORK T3	Network egress blocked; human approval required
A2	Collusion → harmful Rx [3]	WRITE-PHI-ORD T3 + delegation scope check	Medication order gated on clinician review; privilege escalation via delegation blocked
A3	Spoofed-identity EHR write [4]	WRITE-PHI-ORD T3 + audit log	High-risk write blocked; spoofed identity detected in audit trail

8. Conclusions

We presented CAPF, a Clinical Agent Permission Framework that provides a HIPAA-mapped, clinically grounded permission taxonomy for multi-agent healthcare AI systems—among the first to map clinical agent tool permissions to HIPAA-aligned least-privilege controls. CAPF defines eight permission classes, maps each to 45 CFR §164.514(d) and the §164.312 technical safeguards [8], [9], and enforces them through a three-tier runtime protocol. Analytical evaluation against three published attack scenarios shows that CAPF's permission scoping structurally constrains the attack surface at the action boundary,

independently of the agent's reasoning layer. As NIST's AI Agent Standards Initiative [11] and the FDA's AI/ML action plan [12] drive toward mandatory governance for autonomous AI in high-stakes domains, CAPF provides a reference design for the authorization dimension. Future work includes empirical evaluation using the MIMIC-IV dataset, formal verification of the taxonomy's completeness, and integration with emerging agent-identity standards.

Acknowledgements

The authors used AI-assisted writing tools for drafting and editing portions of this manuscript. All technical claims, framework design, permission taxonomy, HIPAA mappings, and bibliographic citations are the authors' own original work and have been independently verified against primary sources. This work was conducted independently in the authors' personal capacity and does not represent the views, positions, or endorsement of any employer.

References

[1]. S. Vatsal, H. Dubey, A. Singh, Agentic AI in healthcare and medicine: A seven-dimensional taxonomy for empirical evaluation of LLM-based agents, *IEEE Access*, Vol. 14, 2026, pp. 4840–4863.

[2]. J. Qiu, L. Li, J. Sun, H. Wei, et al., Emerging cyber attack risks of medical AI agents, arXiv preprint, arXiv:2504.03759, 2025.

[3]. A. Bashir, T. A. Han, Z. U. Shamszaman, Many-to-one adversarial consensus: Exposing multi-agent collusion risks in AI-based healthcare, arXiv preprint, arXiv:2512.03097, 2025.

[4]. S. Maiti, Caging the agents: A zero-trust security architecture for autonomous AI in healthcare, arXiv preprint, arXiv:2603.17419, 2026.

[5]. S. Neupane, S. Mittal, S. Rahimi, Towards a HIPAA compliant agentic AI system in healthcare, arXiv preprint, arXiv:2504.17669, 2025.

[6]. G. Zhu, C. Wang, Z. Wang, Z. Li, et al., OpenPort Protocol: A security governance specification for AI agent tool access, arXiv preprint, arXiv:2602.20196, 2026.

[7]. C. Fleming, A. Kundu, R. Kompella, Uncertainty-aware, risk-adaptive access control for agentic systems using an LLM-judged TBAC model, arXiv preprint, arXiv:2510.11414, 2025.

[8]. U.S. Department of Health and Human Services, Minimum Necessary Requirement, 45 CFR §164.514(d), 2000, https://www.hhs.gov/hipaa/for-professionals/privacy/guidance/minimum-necessary-requirement/index.html

[9]. U.S. Department of Health and Human Services, Centers for Medicare & Medicaid Services, HIPAA Security Series: Security Standards: Technical Safeguards, Vol. 2, Paper 4, May 2005, revised March 2007. Available at: https://www.hhs.gov/sites/default/files/ocr/privacy/hipaa/administrative/securityrule/techsafeguards.pdf

[10]. OWASP Foundation, OWASP Top 10 for Large Language Model Applications 2025, *OWASP Foundation*, 2025.
Available at: https://owasp.org/www-project-top-10-for-large-language-model-applications/ (accessed May 2026).

[11]. National Institute of Standards and Technology, Announcing the "AI Agent Standards Initiative" for Interoperable and Secure Innovation, News Release, 17 February 2026.
Available at: https://www.nist.gov/news-events/news/2026/02/announcing-ai-agent-standards-initiative-interoperable-and-secure

[12]. U.S. Food and Drug Administration, Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) Action Plan, *Center for Devices and Radiological Health*, January 2021. Available at: https://www.fda.gov/media/145022/download

[13]. LangChain, Inc., LangGraph, Software repository, GitHub, 2026.
Available at: https://github.com/langchain-ai/langgraph (accessed May 2026).

[14]. Q. Wu, G. Bansal, J. Zhang, Y. Wu, et al., AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation, arXiv preprint, arXiv:2308.08155, 2023.

[15]. X. Tang, A. Zou, Z. Zhang, Z. Li, et al., MedAgents: Large language models as collaborators for zero-shot medical reasoning, in *Findings of the Association for Computational Linguistics: ACL 2024*, *Association for Computational Linguistics*, 2024, pp. 599–621.

[16]. S. Schmidgall, R. Ziaei, C. Harris, E. Reis, et al., AgentClinic: A multimodal agent benchmark to evaluate AI in simulated clinical environments, arXiv preprint, arXiv:2405.07960, 2024.

[17]. C. J. Agostino, N. D'Souza, Herding CATs: ALARA for agent harness engineering in portable composable multi-agent teams, arXiv preprint, arXiv:2603.20380, 2026.

[18]. V. C. Hu, D. Ferraiolo, R. Kuhn, A. Schnitzer, et al., Guide to Attribute Based Access Control (ABAC) Definition and Considerations, NIST Special Publication 800-162, *National Institute of Standards and Technology*, 2014, 47 p.

[19]. J. C. Mandel, D. A. Kreda, K. D. Mandl, I. S. Kohane, et al., SMART on FHIR: A standards-based, interoperable apps platform for electronic health records, *Journal of the American Medical Informatics Association*, Vol. 23, Issue 5, 2016, pp. 899–908.

[20]. K. Kent, M. Souppaya, Guide to Computer Security Log Management, NIST Special Publication 800-92, *National Institute of Standards and Technology*, 2006, 72 p.

[21]. U. Uchibeke, Before the tool call: Deterministic pre-action authorization for autonomous AI agents, arXiv preprint, arXiv:2603.20953, 2026.

[22]. A. Cartagena, A. Teixeira, Mind the GAP: Text safety does not transfer to tool-call safety in LLM agents, arXiv preprint, arXiv:2602.16943, 2026.

[23]. M. Kaptein, V.-J. Khan, A. Podstavnychy, Runtime governance for AI agents: Policies on paths, arXiv preprint, arXiv:2603.16586, 2026.

The views expressed in this work are solely those of the authors in their personal capacity and do not represent the position, opinion, or endorsement of Amazon.com, Inc. or its affiliates.

(057)

ANESTHOS: A Human-in-the-Loop Perioperative Workflow Architecture for Clinical Decision Support, Safety and Continuous Operational Intelligence

A. Binagui Buitureira
North West Regional Hospital, Burnie 7320, Tasmania, Australia
Tel.: + 61 497324745
E-mail: abinagui@hotmail.com, alessandra.binagui@ths.tas.gov.au

Summary: ANESTHOS is a human-in-the-loop perioperative workflow platform designed to improve safety, traceability, standardisation and clinical governance across the surgical journey. Unlike isolated predictive AI tools, ANESTHOS integrates assessment, risk stratification, safety checklists, documentation, handover, audit and continuous professional development into a unified workflow.

Keywords: Artificial intelligence, Perioperative medicine, Anesthesia workflow, Clinical decision support, Human-in-the-loop AI, Patient safety, Auditability, Healthcare operations, Digital health, Workflow integration.

1. Introduction/Background

Perioperative care remains fragmented across multiple stages of the patient journey. Current AI solutions frequently focus on isolated predictive models rather than continuous workflow integration into real-world clinical practice [1, 2]. While many digital capabilities already exist, they often remain disconnected across platforms, limiting continuity and operational learning [2, 5].

2. Objective/Goals

To present ANESTHOS, a human-in-the-loop AI-supported perioperative workflow architecture designed to improve safety, traceability, auditability, standardization, and clinician support across the entire perioperative pathway.

3. Methods

ANESTHOS integrates preoperative assessment, risk evaluation, safety checklists, documentation support, handover systems, audit capabilities and CPD. The system follows a clinician-augmentation model rather than autonomous decision-making [3, 6] The novelty resides in the orchestration and implementation of these components within a single workflow architecture [2, 5].
The proposed architecture includes interconnected modules for:

- Preoperative assessment;
- Dynamic risk evaluation;
- Voice-assisted surgical safety checklists;
- Intraoperative documentation support;
- Structured handover systems;
- Acute pain service integration;
- Audit and benchmarking capabilities;
- Education and continuous professional development (CPD).

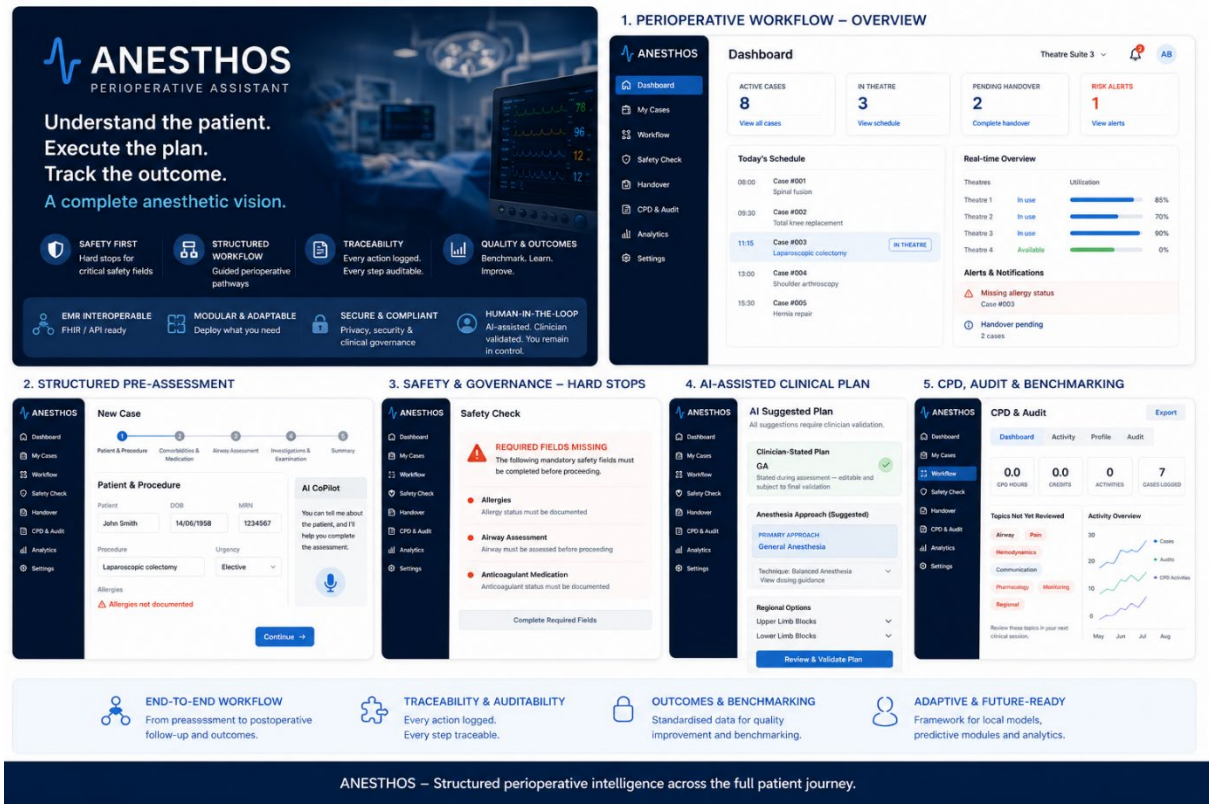
The system is designed under a clinician-augmentation model rather than autonomous decision-making. Human validation remains mandatory at all critical decision points. Safety-oriented hard stops are integrated for high-risk variables such as allergies, airway assessment, and critical perioperative omissions.
The architecture incorporates structured and multimodal clinical data including demographics, comorbidities, medications, laboratory values, airway evaluation, physiological monitoring, and perioperative outcomes. AI-generated outputs include structured summaries, suggested anesthetic plans, risk alerts, checklist support, and evidence-linked recommendations.

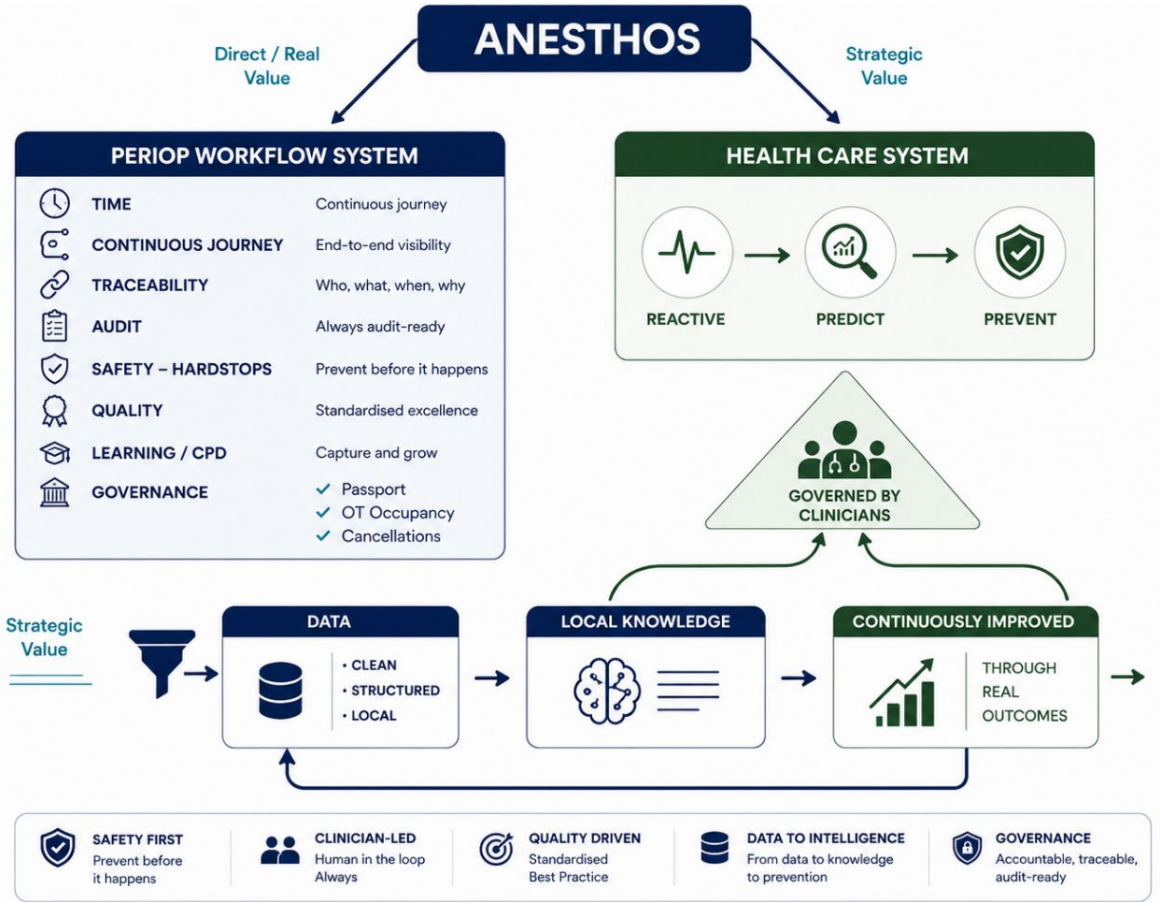
4. Architectural Contributions

The proposed framework introduces several operational innovations:

1. Continuous perioperative traceability across the full patient journey;
2. AI-supported standardization while preserving clinician oversight;
3. Integration of safety systems directly into workflow processes;
4. Structured auditability for intra- and inter-operator benchmarking;
5. Potential generation of high-quality real-world perioperative datasets for quality improvement and future machine learning development;
6. Reduction of administrative burden and duplication of documentation;
7. Integration of education, CPD, and evidence-based clinical support into routine workflow.

Unlike isolated predictive AI tools, ANESTHOS focuses on implementation realism, workflow continuity, and operational integration within existing hospital environments.


ANESTHOS
PERIOPERATIVE ASSISTANT
Understand the patient.
Execute the plan.
Track the outcome.
A complete anesthetic vision.
SAFETY FIRST
Hard stops for critical safety fields
STRUCTURED WORKFLOW
Guided perioperative pathways
TRACEABILITY
Every action logged. Every step auditable.
QUALITY & OUTCOMES
Benchmark. Learn. Improve.
EMR INTEROPERABLE
FHIR / API ready
MODULAR & ADAPTABLE
Deploy what you need
SECURE & COMPLIANT
Privacy, security & clinical governance
HUMAN-IN-THE-LOOP
AI-assisted. Clinician validated. You remain in control.
1. PERIOPERATIVE WORKFLOW – OVERVIEW
Dashboard
2. STRUCTURED PRE-ASSESSMENT
3. SAFETY & GOVERNANCE – HARD STOPS
4. AI-ASSISTED CLINICAL PLAN
5. CPD, AUDIT & BENCHMARKING
END-TO-END WORKFLOW
From preassessment to postoperative follow-up and outcomes.
TRACEABILITY & AUDITABILITY
Every action logged. Every step traceable.
OUTCOMES & BENCHMARKING
Standardised data for quality improvement and benchmarking.
ADAPTIVE & FUTURE-READY
Framework for local models, predictive modules and analytics.
ANESTHOS – Structured perioperative intelligence across the full patient journey.


ANESTHOS
Direct / Real Value
Strategic Value
PERIOP WORKFLOW SYSTEM
TIME
Continuous journey
CONTINUOUS JOURNEY
End-to-end visibility
TRACEABILITY
Who, what, when, why
AUDIT
Always audit-ready
SAFETY – HARDSTOPS
Prevent before it happens
QUALITY
Standardised excellence
LEARNING / CPD
Capture and grow
GOVERNANCE
Passport
OT Occupancy
Cancellations
HEALTH CARE SYSTEM
REACTIVE
PREDICT
PREVENT
GOVERNED BY CLINICIANS
Strategic Value
DATA
• CLEAN
• STRUCTURED
• LOCAL
LOCAL KNOWLEDGE
CONTINUOUSLY IMPROVED
THROUGH REAL OUTCOMES
SAFETY FIRST
Prevent before it happens
CLINICIAN-LED
Human in the loop Always
QUALITY DRIVEN
Standardised Best Practice
DATA TO INTELLIGENCE
From data to knowledge to prevention
GOVERNANCE
Accountable, traceable, audit-ready

5. Discussion

Current healthcare AI development often faces a translational gap between laboratory performance and real-world clinical implementation [5, 6]. ANESTHOS addresses this gap by prioritizing usability, governance, workflow integration, and clinician-centered design. The proposed framework may facilitate safer perioperative practice, stronger quality assurance systems, and improved institutional learning cultures. [3]

Importantly, the system is not designed to replace clinician judgment. Instead, it aims to enhance situational awareness, reduce variability, improve documentation quality, and support structured perioperative decision-making.

Structured perioperative datasets may enable future predictive models and continuous quality improvement.

6. Conclusion

ANESTHOS represents an integrated perioperative workflow architecture combining clinical decision support, safety systems and future predictive intelligence within a unified human-in-the-loop framework. [3, 6]

References

[1]. V. Bellini, E. Rafano Carnà, M. Russo, F. Di Vincenzo, et al., Artificial intelligence and anesthesia: A narrative review, *Annals of Translational Medicine*, Vol. 10, Issue 9, 2022, Article 528.

[2]. H.-K. Yoon, H.-L. Yang, C.-W. Jung, H.-C. Lee, Artificial intelligence in perioperative medicine: A narrative review, *Korean Journal of Anesthesiology*, Vol. 75, Issue 3, 2022, pp. 202–215.

[3]. A. E. Davidson, J. M. Ray, Y. Levites Strekalova, P. Rashidi, et al., Human-centered development of an explainable AI framework for real-time surgical risk surveillance, arXiv preprint, arXiv:2504.02551, 2025.

[4]. V. Bellini, M. Valente, G. Bertorelli, B. Pifferi, et al., Machine learning in perioperative medicine: A systematic review, *Journal of Anesthesia, Analgesia and Critical Care*, Vol. 2, 2022, Article 2.

[5]. D. A. Hashimoto, E. R. Witkowski, L. Gao, O. R. Meireles, et al., Artificial intelligence in anesthesiology: Current techniques, clinical applications, and limitations, *Anesthesiology*, Vol. 132, Issue 2, 2020, pp. 379–394.

[6]. E. J. Topol, High-performance medicine: The convergence of human and artificial intelligence, *Nature Medicine*, Vol. 25, Issue 1, 2019, pp. 44–56.

ISBN 978-84-09-86146-0

9788409861460