AIHealth 2026

The Third International Conference on AI-Health

ISBN: 978-1-68558-362-0

March 8th –12th, 2026

Valencia, Spain

AIHealth 2026 Editors

Elena Mancuso, Engineering Ingegneria Informatica SpA, Italy

Michael Massoth, Hochschule Darmstadt - University of Applied Sciences, Germany

AIHealth 2026

Foreword

The Third International Conference on AI-Health (AIHealth 2026), held between March 8 - 12, 2026, covered topics blending Artificial Intelligence and health sciences.

Quality healthcare should be extended to all communities. Independent of how big and complex the healthcare systems are, physicians are under increasing time and workload pressures and spending less time with patients. The challenge to deliver high-quality healthcare against administrative burdens is big and increasing.

Healthcare facilities also produce great amounts of data and record high volumes of patient records information. This information is valuable and necessary to quality patient care. This information requires an enormousness effort (time, personnel) to be timely processed for prediction, evaluation and monitoring patients' health.

Artificial Intelligence (AI) comes to rescue in terms of accuracy, precision, rapidity and processing a large volume of data. AI-based health systems benefit for recent advances in sophisticated AI mechanisms for predicting patient health conditions (personalized, at large scale), producing useful analytics on varii patient health aspects, as well as monitoring and controlling patient under scrutiny.

We take here the opportunity to warmly thank all the members of the AIHealth 2026 Technical Program Committee, as well as the numerous reviewers. The creation of such a high quality conference program would not have been possible without their involvement. We also kindly thank all the authors who dedicated much of their time and efforts to contribute to AIHealth 2026.

Also, this event could not have been a reality without the support of many individuals, organizations, and sponsors. We are grateful to the members of the AIHealth 2026 organizing committee for their help in handling the logistics and for their work to make this professional meeting a success.

We hope that AIHealth 2026 was a successful international forum for the exchange of ideas and results between academia and industry and for the promotion of progress in the fields of AI and health sciences.

We are convinced that the participants found the event useful and communications very open. We also hope that Valencia provided a pleasant environment during the conference and everyone saved some time for exploring this beautiful city

AIHealth 2026 Chairs:

AIHealth 2026 Steering Committee
Les Sztandera, Thomas Jefferson University, USA
Hesham H. Ali, University of Nebraska at Omaha, USA
Maura Mengoni, Università Politecnica delle Marche, Italy
Vitaly Herasevich, Mayo Clinic, USA
Amina Souag, Canterbury Christ Church University, UK

AIHealth 2026 Publicity Chairs
Francisco Javier Díaz Blasco, Universitat Politècnica de València, Spain
Ali Ahmad, Universitat Politècnica de València, Spain

Sandra Viciano Tudela, Universitat Politecnica de Valencia, Spain
José Miguel Jiménez, Universitat Politecnica de Valencia, Spain

AIHealth 2026

Committee

AIHealth 2026 Steering Committee
Les Sztandera, Thomas Jefferson University, USA
Hesham H. Ali, University of Nebraska at Omaha, USA
Maura Mengoni, Università Politecnica delle Marche, Italy
Vitaly Herasevich, Mayo Clinic, USA
Amina Souag, Canterbury Christ Church University, UK

AIHealth 2026 Publicity Chairs
Francisco Javier Díaz Blasco, Universitat Politècnica de València, Spain
Ali Ahmad, Universitat Politècnica de València, Spain
Sandra Viciano Tudela, Universitat Politecnica de Valencia, Spain
José Miguel Jiménez, Universitat Politecnica de Valencia, Spain

AIHealth 2026 Technical Program Committee

Alaa Abd-Alrazaq, Weill Cornell Medicine-Qatar, Doha, Qatar
Hesham H. Ali, University of Nebraska at Omaha, USA
Alireza Atashi, Tehran University of Medical Sciences, Iran
Michael Beigl, Karlsruhe Institute of Technology (KIT), Germany
Sid-Ahmed Berrani, Ecole Nationale Polytechnique, Algiers, Algeria
Elizabeth Borycki, University of Victoria, Canada
An Braeken, Vrije Universiteit Brussel, Belgium
Philippe Cinquin, CHUGrenoble Alpes, France
Marcos Cordeiro d'Ornellas, Universidade Federal de Santa Maria (UFSM) | Hospital Universitário (HUSM), Brazil
Manuel Domínguez-Morales, University of Sevilla, Spain
Sai Anvesh Durvasula, Parabole.ai, USA
Duarte Duque, 2Ai - School of Technology | IPCA, Portugal
Vitaly Herasevich, Mayo Clinic, USA
Hamza Hussain, Oakland University William Beaumont School of Medicine, USA
Haralampos Karanikas, University of Thessaly, Greece
Sarfraz Khokhar, Rasimo Systems, USA
Baihua Li, Loughborough University, UK
Sushil K. Meher, All India Institute of Medical Sciences, New Delhi, India
James Meng, Norwich Medical School | University of East Anglia, UK
Maura Mengoni, Polytechnic University of Marche, Italy
Daniela Micucci, University of Milano - Bicocca, Italy
George Mihalas, Victor Babes Univ. Med.&Pharm, Timisoara | Academy of Medical Sciences, Com. Medical Informatics & Data Protection, Romania
Kartik Palani, iManage / University of Illinois Urbana-Champaign, USA
Nadav Rappoport, Ben-Gurion University of the Negev, Israel
Stefano Rinaldi, University of Brescia, Italy

Copyright Information

For your reference, this is the text governing the copyright release for material published by IARIA.

The copyright release is a transfer of publication rights, which allows IARIA and its partners to drive the dissemination of the published material. This allows IARIA to give articles increased visibility via distribution, inclusion in libraries, and arrangements for submission to indexes.

I, the undersigned, declare that the article is original, and that I represent the authors of this article in the copyright release matters. If this work has been done as work-for-hire, I have obtained all necessary clearances to execute a copyright release. I hereby irrevocably transfer exclusive copyright for this material to IARIA. I give IARIA permission or reproduce the work in any media format such as, but not limited to, print, digital, or electronic. I give IARIA permission to distribute the materials without restriction to any institutions or individuals. I give IARIA permission to submit the work for inclusion in article repositories as IARIA sees fit.

I, the undersigned, declare that to the best of my knowledge, the article is does not contain libelous or otherwise unlawful contents or invading the right of privacy or infringing on a proprietary right.

Following the copyright release, any circulated version of the article must bear the copyright notice and any header and footer information that IARIA applies to the published article.

IARIA grants royalty-free permission to the authors to disseminate the work, under the above provisions, for any academic, commercial, or industrial use. IARIA grants royalty-free permission to any individuals or institutions to make the article available electronically, online, or in print.

IARIA acknowledges that rights to any algorithm, process, procedure, apparatus, or articles of manufacture remain with the authors and their employers.

I, the undersigned, understand that IARIA will not be liable, in contract, tort (including, without limitation, negligence), pre-contract or other representations (other than fraudulent misrepresentations) or otherwise in connection with the publication of my work.

Exception to the above is made for work-for-hire performed while employed by the government. In that case, copyright to the material remains with the said government. The rightful owners (authors and government entity) grant unlimited and unrestricted permission to IARIA, IARIA's contractors, and IARIA's partners to further distribute the work.

Table of Contents

Real-World Validation of Arkangel AI: A Conversational Agent for Real-Time, Evidence-Based Medical Question-Answering

Natalia Castano Villegas, Isabella Llano
Evidence Department
Arkangel AI
Bogotá, Colombia
natalia@arkangel.ai, isabella.llano@arkangel.ai

Maria Camila Villa, Jose Zea
Product Department
Arkangel AI
Bogotá, Colombia
camila.villa@arkangel.ai, jose@arkangel.ai

***Abstract*—The paper presents the first external validation of Arkangel AI (formerly MedSearch), a retrieval-augmented large language model-based conversational agent for evidence-based medical question-answering. *Problem:* Large language model-powered conversational agents are evaluated mainly on medical question-answering datasets with strong benchmark performance; however, multiple-choice formats do not assess complex, open-ended clinical reasoning or real-world search behavior. *Why it matters:* In healthcare, validity, safety, and currency of answers matter as much as speed; prior work has not fully addressed these in ecologically valid settings with mixed healthcare personnel. *Gap:* Existing evaluations rarely combine time-to-answer, number of searches, and expert-rated validity across multiple domains in a single blinded trial. *Solution:* The authors conducted a randomized, double-blind trial comparing Arkangel AI versus traditional non-AI search in healthcare personnel answering four clinical cases (four questions each). Validity was assessed with six domains (accuracy, consensus, bias, currency, safety) on a 3-point scale; specialists were blinded to group. *Main outcome:* Arkangel AI users achieved higher validity scores across all domains ($p < 0.01$), arrived at a final answer in less than half the time (three minutes faster) than the control group, with approximately 50% fewer searches per case; total average acceptability score 2.86 on a scale from 1 to 3. Most users found Arkangel AI helpful for daily practice and would recommend it. The main conclusion is that large language model-supported methods can improve clinical search efficiency without sacrificing, and even augmenting, answer quality in this setting; broader validations are needed.**

***Keywords*—Large language model assessment; human evaluation; healthcare; real-world validation.**

I. INTRODUCTION

Application of large language models (LLMs) as conversational agents (CAs) in healthcare has been evaluated using medical question-answering (QA) datasets, with excellent performance on international licensing-style exams [1, 2, 3]. Multiple-choice formats fall short when the goal is to assess open-ended clinical reasoning, source traceability, and real-world search behavior. To address this, we developed Arkangel AI (formerly MedSearch), an LLM-powered CA that performs real-time, evidence-based searches and provides curated references [4]. Our first manuscript reported internal validation (90.26% on MedQA). Here we report the first external validation using real-world healthcare personnel.

Research questions. (RQ1) Does Arkangel AI improve response validity versus traditional search? (RQ2) Does it reduce time and number of searches? (RQ3) How acceptable is it to users? (RQ4) How much does evaluator variability affect scores?

Limitations of this work. The sample was predominantly clinical-year medical students (>70%) and all cases were non-urgent, outpatient; one specialist per specialty scored answers, so classical inter-rater agreement is not reported for this dataset.

Paper structure. Section II describes methodology; Section III reports results; Section IV discusses findings; Section V concludes.

II. METHODOLOGY

Design and participants. Randomized, double-blind trial; over 100 healthcare personnel recruited in Colombia via social media, professional networks, and masterclasses. After informed consent, participants were randomly assigned to Group A (Arkangel AI) or Group B (traditional search: Google, PubMed, guidelines; no AI). Four outpatient clinical cases (orthopedics, psychiatry, pediatrics, gynecology), four questions per case (diagnostic, management, research, general), were designed by external specialists. Quizizz recorded time per question; Airtable captured specialist ratings. Researchers and evaluators were blinded to participant identity and group assignment.

Outcomes. (1) Validity: six domains (correctness, consensus alignment, demographic bias, treatment bias, currency, patient risk) on a 3-point ordinal scale; one specialist per

 ISBN: 978-1-68558-362-0

specialty scored all QA pairs. Total Average Validity Score = mean of the six domain scores. (2) Efficiency: time per case and number of searches per case (self-reported). (3) Acceptability: four items on a 3-point scale, assessed in Group A only.

Analysis. Distributions were non-normal (Shapiro–Wilk); Mann–Whitney U test for group comparisons; medians and IQR as primary estimates; means (SD) and 95% CI also reported. Subgroup analyses by specialty and question type are exploratory.

III. RESULTS

106 participants answered 1600 questions (406 case-units). After exclusions (platform error $n = 2$, protocol violation $n = 1$), 55 remained in Group A and 48 in Group B; more than 70% were medical students.

Validity. Group A (Arkangel AI) scored higher than Group B on all six validity domains (all $p < 0.01$). Largest relative improvement: response accuracy (13.12%); smallest: medical consensus alignment (3.25%). Table I summarizes the key outcomes across both groups.

TABLE I. KEY OUTCOMES: ARKANGEL AI (GROUP A) VS. TRADITIONAL SEARCH (GROUP B)

OUTCOME	GROUP A	GROUP B	P-VALUE
Validity (1–3)	Higher in all 6 domains	Lower	< 0.01
Time per case	~3 min faster (69%)	—	< 0.001
Searches per case	~50% fewer	—	< 0.001

Efficiency. Group A reached a final answer in less than half the time (approximately three minutes faster per case) than Group B, with approximately 50% fewer searches per case. Mann–Whitney U, $p < 0.001$.

Acceptability. In Group A, total average acceptability score was 2.86 on a scale from 1 to 3. Dimension scores: utility 2.98, daily use 2.87, recommendation 2.93, truthfulness confidence 2.65 (lowest).

Evaluator variability. Linear mixed models (group as fixed effect, evaluator as random effect) showed significant group effects across all six validity domains (all $p < 0.01$). ICC values indicated that most variance was within-evaluator (intra-evaluator). The single-rater-per-specialty design is a stated limitation.

IV. DISCUSSION

Results support that Arkangel AI improves efficiency (time and searches) and validity scores versus traditional search in this predominantly student sample. Strengths include the randomized, double-blinded design, expert-designed cases, and multidimensional validity framework. Key limitations are the overrepresentation of medical students (>70%), single specialist per specialty, and acceptability assessed only in the intervention group. All authors are affiliated with Arkangel AI; evaluators and case designers were external. Findings align with recent work showing that retrieval-augmented LLM workflows can reduce clinical search burden [5], while persistent physician skepticism toward AI-driven tools remains a barrier [6].

V. CONCLUSION AND FUTURE WORK

This extended abstract presents the first external validation of Arkangel AI, demonstrating superior validity, efficiency, and acceptability versus traditional search in an elective clinical setting. Large language model-supported, evidence-based search can enhance physician workflows in this context. Future work includes multicenter recruitment, two independent raters per answer with adjudication and inter-rater reliability, pre-registered protocols, and validation in general practitioners, specialists, and higher-acuity scenarios.

ACKNOWLEDGMENT

The authors thank the healthcare personnel and medical students who participated, and the specialist physicians who designed the cases and evaluated responses.

REFERENCES

[1] A. Gilson et al., “How Does ChatGPT Perform on the USMLE?” JMIR Med Educ., vol. 9, p. e45312, 2023.

[2] A. J. Thirunavukarasu et al., “Large language models in medicine,” Nat. Med., vol. 29, pp. 1930–1940, 2023.

[3] K. Singhal et al., “Towards expert-level medical question answering with large language models,” arXiv:2305.09617, 2023.

[4] I. Llano et al., “MedSearch: A conversational agent for real-time, evidence-based medical question-answering,” SSRN 5092702, 2025.

[5] S. Shool et al., “A systematic review of LLM evaluations in clinical medicine,” BMC Med. Inform. Decis. Mak., vol. 25, p. 117, 2025.

[6] T. Sakamoto, Y. Harada, and T. Shimizu, “Facilitating trust calibration in AI-driven diagnostic support,” JMIR Form. Res., vol. 8, p. e58666, 2024.

AI for Referable Knee Radiograph Detection in Primary Care: A Pathway-Specific Taxonomy for Dataset Generation from Reports

[1,2,*]Imanol Pinto, [2]Álvaro Olazarán, [2]David Jurio, [2]Miguel Sainz, [4]Natalia Álvarez, [1,3]Mikel Galar

[1]Institute of Smart Cities, Public University of Navarre, Campus Arrosadia, 31006 Pamplona, España

[2] General Directorate of Telecommunications and Digitalization, Government of Navarre, C/ Cabarceno 6, Sarriguren, 31621, Spain

[3]IdiSNA, Navarre Institute of Health Research, C/ Irunlarrea 3, Pamplona, 31008, Spain

[4]University Hospital of Navarre, Musculoskeletal Radiology Service, C/ Irunlarrea 3, Pamplona 31008, Spain

[*]Corresponding author. E-mail: imanolpinto@proton.me

***Abstract*—Radiologist shortages often delay musculoskeletal radiograph interpretation in primary care, with the knee among the most frequently imaged regions. To our knowledge, we developed the first Artificial Intelligence (AI) system for knee radiograph assessment in this setting, using a pathway-specific taxonomy to automatically label reports, thereby advancing the field of medical disease extraction. A total of 57,460 knee radiology reports (2010–2024) from the Public Healthcare Service of Navarre, Spain, were retrospectively processed with a Natural Language Processing (NLP) pipeline guided by this taxonomy. The pipeline extracted thousands of frequent and relevant findings, reliably linked them to images through a projection and laterality classifier, and organized them into 43 hierarchical categories for dataset generation. To assess feasibility, we trained a ConvNeXt-Small model to classify radiographs as referable (requiring specialist review) or non-referable, and validated it on an independent test set of 494 studies (39.7% referable). Ground truth was defined by consensus of three radiologists from a panel of five experts. On this test set, the model achieved an Area Under the ROC Curve (AUC) of 0.880 (95% Confidence Interval [CI]: 0.843–0.915), with 81.9% sensitivity and 83.1% specificity, significantly outperforming routine reports (AUC 0.798; p=0.0002). Compared with the individual radiologists, the model achieved comparable sensitivity but lower specificity (3.6–11% below radiologists). These results support the potential of our deep learning algorithm as a primary care decision-support tool, helping reduce unnecessary referrals and radiologist workload, while showing how pathway-specific taxonomies enable scalable, efficient AI in data-limited settings.**

Keywords-Deep Learning for radiology; Medical disease extraction; AI-based health systems and applications; Knee Radiography; Primary Care

I. INTRODUCTION

Musculoskeletal (MSK) conditions are one of the most frequent reasons for patient visits in primary care, where plain radiography constitutes the primary modality for initial imaging assessment [1][2]. Nevertheless, delays in expert radiological interpretation represent an obstacle in many healthcare systems, often postponing diagnosis and subsequent treatment. Therefore, addressing these delays is essential for improving administrative efficiency in primary care. In this context, general practitioners frequently assume responsibility for preliminary image review despite lacking specialized training to identify the full spectrum of MSK conditions.

Recent advances in Artificial Intelligence (AI), particularly in Deep Learning (DL), have demonstrated strong performance in the analysis of MSK radiographs [3]. Building on this progress, we present an AI-based framework designed specifically for the evaluation of knee radiographs in primary care —one of the most frequently imaged anatomical regions in this clinical context, second only to the spine [1][2]. Knee radiographs offer an ideal starting point for AI-based decision support due to their relatively simple anatomical structure and standardized diagnostic criteria, in contrast to the spine, which presents greater complexity both anatomically and diagnostically [4][5].

Despite this opportunity, there are no studies focusing on AI tools for preliminary interpretation of knee radiographs in primary care settings. Most existing research concentrates on specific MSK disorders, such as osteoarthritis [6], fractures [7], bone age [8] and tumors [9], using manually annotated datasets and specialist-oriented labels [10].

To address this gap, we implemented a data pipeline that takes advantage of routinely collected clinical data from the Public Healthcare Service of Navarre (Servicio Navarro de Salud–Osasunbidea, SNS-O) to create a large-scale training set. The pipeline uses an image classifier to reliably identify target knee radiographs, and applies Natural Language Processing (NLP) to radiology reports to extract clinical findings, which are then organized into categories using a custom taxonomy that reflects primary care decision-making needs.

We assessed the effectiveness of this approach by training a DL model on the resulting dataset and evaluating it on a radiologist-annotated test set, achieving promising results.

To our knowledge, this is the first AI system specifically designed for preliminary knee radiograph assessment in primary care. In this context, emulating the diagnostic approach of a radiologist, the proposed AI system analyzes a broad range of radiological findings. Our methodological framework advances the field of medical disease extraction and illustrates how NLP-based labeling, guided by a taxonomy aligned with the clinical pathway, can enable the development of AI-based health systems and applications in settings with limited annotated data.

This paper is organized as follows: Section II reviews related work, Section III describes the methodology and dataset creation, Section IV presents validation results, and Section V discusses conclusions, limitations, and future work.

II. RELATED WORK

The application of AI to MSK radiology has advanced rapidly, particularly for trauma detection [7], bone-age assessment [8], osteoarthritis grading [6] and implant evaluation [10]. Meta-analyses and large studies report high diagnostic accuracy for fracture detection on plain radiographs (pooled sensitivities/specificities often >91%) and demonstrate that AI assistance can improve clinician performance in routine practice [7][11][12]. For the knee, DL models have achieved expert-level performance in Kellgren–Lawrence (KL) osteoarthritis grading, with reported AUCs up to ≈ 0.96 and Cohen's kappa around 0.86 [6][13]. While large MSK repositories exist, such as MURA for upper limbs [14] and the Osteoarthritis Initiative for knee osteoarthritis [15], to our knowledge, no comparable large-scale, multi-label (multiple findings per image) knee datasets are currently available.

To address the scarcity of annotated data, prior thoracic imaging studies have applied NLP to radiology reports to generate large-scale, weakly supervised labels (e.g., PadChest [16], CheXpert [17]). Since chest imaging taxonomies cannot be directly applied to knee radiographs, we developed a novel taxonomy for primary care, grounded in real-world finding prevalence and clinical relevance. This taxonomy enables the construction of large-scale, multi-label knee datasets and the development of AI models tailored to general practitioners' needs.

III. METHODS

This study focuses on developing and evaluating an AI framework for knee radiograph assessment in primary care. Rather than aiming for maximal model performance, our primary goal was to demonstrate the feasibility of creating a clinically meaningful decision-support system from automatically labeled retrospective data.

A. Dataset Creation

The SNS-O comprises 63 primary care centers and 3 hospitals, serving a population of over 600,000 patients. From its radiology information system, we retrieved 263,763 knee radiograph studies requested by general practitioners (in primary care) between 2010 and 2024, of which 57,460 (21%) had an associated formal radiology report and were available for analysis.

Each study typically contained multiple radiograph images of one or both knees, often captured in multiple projections (anteroposterior, lateral, axial and others) and of varying laterality (right, left, or bilateral), and interpreted by a radiologist in a single report. To transform this unstructured material into an AI-ready dataset, three tasks were required: (1) mapping radiological findings from the free-text reports into structured categories, (2) reliably identifying projection and laterality for each image, and (3) linking each image with its corresponding findings.

Accordingly, as illustrated in Figure 1, we developed a data pipeline with two main branches: label generation and image selection. Label generation involved the use of NLP techniques to extract radiological findings from reports, which were selected and organized using our primary care–oriented taxonomy, detailed in Section III-B. As image metadata was usually missing, we developed a neural network to perform image selection, specifically to determine projection and laterality.

The resulting dataset comprised 63,976 single-knee radiograph studies from 28,719 patients (54% female; mean age 58.2 years). For each knee, findings were mapped to a category within our taxonomy and assigned a global grade of *referable* or *non-referable*, indicating whether expert radiologist evaluation was recommended. The pipeline yielded a dataset of 56,152 non-referable studies (87.77%) and 7,824 referable studies (12.23%), each consisting of a single knee report linked to its corresponding radiographs.

Figure 2 illustrates an example study alongside its radiology report (translated to English), extracted finding categories, and identified projections and laterality. In this case, since the study involved only a single knee, no laterality-based splitting was required. The two core components of the pipeline (label generation and image selection) are described below.

Label generation. Our label generation pipeline processed each report in four main steps: (1) laterality analysis, splitting reports that described both knees into separate reports, one per knee; (2) biomedical named entity recognition, extracting all mentioned medical entities; (3) negation detection, discarding entities appearing in a negated context; and (4) taxonomy mapping, filtering entities, assigning them to relevant finding categories, and generating a final binary label (referable/non-referable).

Reports were split by laterality using regular expressions, since most included distinct sections for each knee. Single-laterality reports were left intact. Entity recognition used a Spanish RoBERTa-base biomedical model [18], while negation detection relied on a fine-tuned BERT-base multilingual model trained on the NUBES Spanish clinical dataset [19][20].

When applied to our retrospective dataset, the NLP pipeline extracted 102,127 entity mentions (6,676 unique entities). Of these, 87,189 (85%) were successfully mapped using the taxonomy detailed in Section III-B. A report was labeled as referable if at least one referable finding category was present.

Image selection. Most radiographs lacked reliable DICOM metadata to accurately determine laterality and projection. To address this, we manually annotated 2,214 knee radiographs, assigning one of the following nine labels: anteroposterior (right/left), lateral (right/left), axial (right/left), bilateral, biaxial, or other. We then trained a ConvNeXt-Tiny model on this task using Fast.ai and Timm libraries [21][22], with ImageNet pretraining and input images resized to 350×350px. The network was trained for 22 epochs with a learning rate of 0.006, using standard augmentations, such as resizing, brightness/contrast adjustments, random erasing, and geometric transformations. Horizontal and vertical flips were deliberately excluded to preserve anatomical laterality. The model achieved an AUC of 0.9984 and an accuracy of 97.4%,

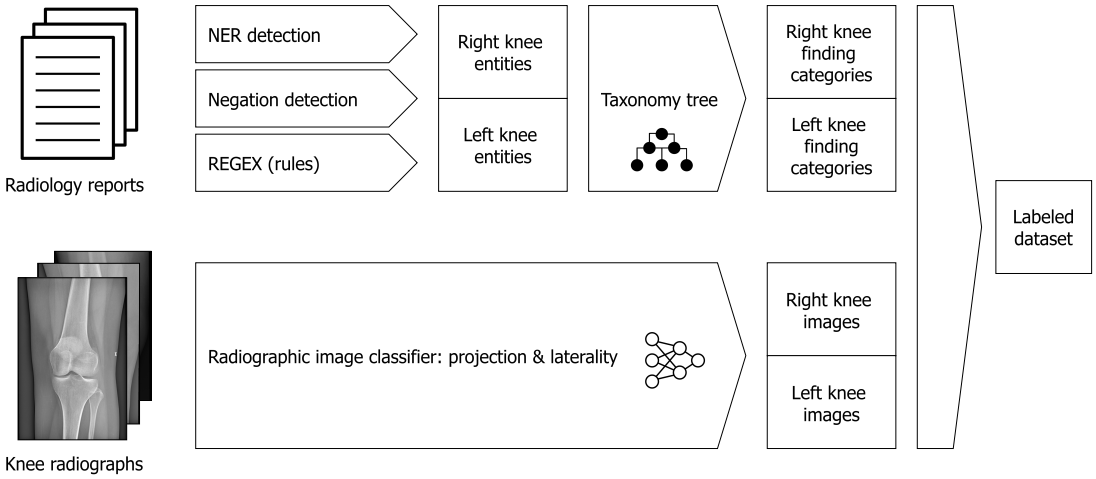


Figure 1. Data pipeline for dataset creation. The NLP pipeline (top) and the image selection (bottom) branches are merged to produce our dataset.

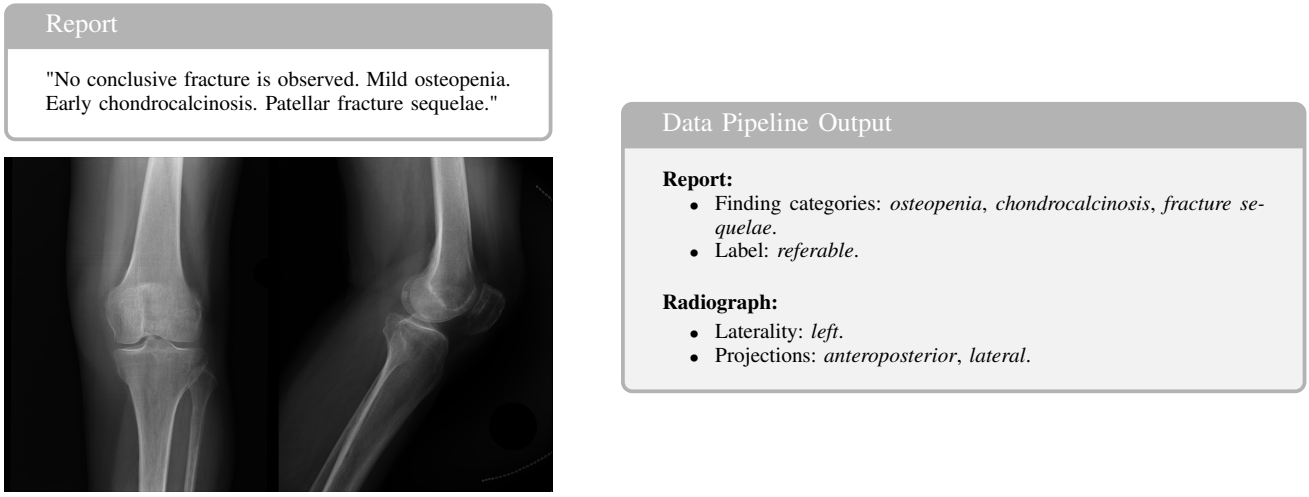


Figure 2. Example knee study with corresponding radiology report and two projections (left). The outputs of the data pipeline are also displayed (right).

enabling reliable assignment of radiographic laterality and projection across the dataset. For the purposes of this study, only anteroposterior and lateral projections were retained, as they are the standard views for clinical knee assessment. When a radiograph was classified as bilateral, the image was divided in two standard projections.

B. A Taxonomy for Primary Care Knee Radiography

A radiographic finding taxonomy was developed to categorize the reports. The complete taxonomy is presented in Figure 3, which shows the resulting finding categories and their relative frequencies, selected based on entity frequency and clinical relevance in primary care.

Finding categories were initially defined by a senior MSK radiologist based on SNS-O clinical pathways and subsequently reviewed and refined within a multidisciplinary workgroup comprising five MSK radiologists and one senior general practitioner. Categories were retained if they appeared in at least 0.9% of the dataset or posed significant clinical risk. Findings were classified as referable when they typically required radiologist review, advanced imaging, or management changes beyond primary care. In the absence of formal guidelines, decisions were guided by routine SNS-O practice to safely reduce radiologist workload.

Using this taxonomy, extracted clinical entities were consolidated into broader categories and classified as referable or non-referable, providing practical guidance for general practitioners (e.g., “chondrocalcinosis” and “calcification” were mapped to non-specific calcification, classified as non-referable). Detailed entity mapping is provided in the Supplementary Material.

C. Model Training

To evaluate the utility of the generated dataset, we trained a baseline ConvNeXt-Small model to classify knee radiographs as referable or non-referable. The dataset included 62,309 anteroposterior and lateral radiographs, each labeled according to the overall referability grade of its corresponding knee. A

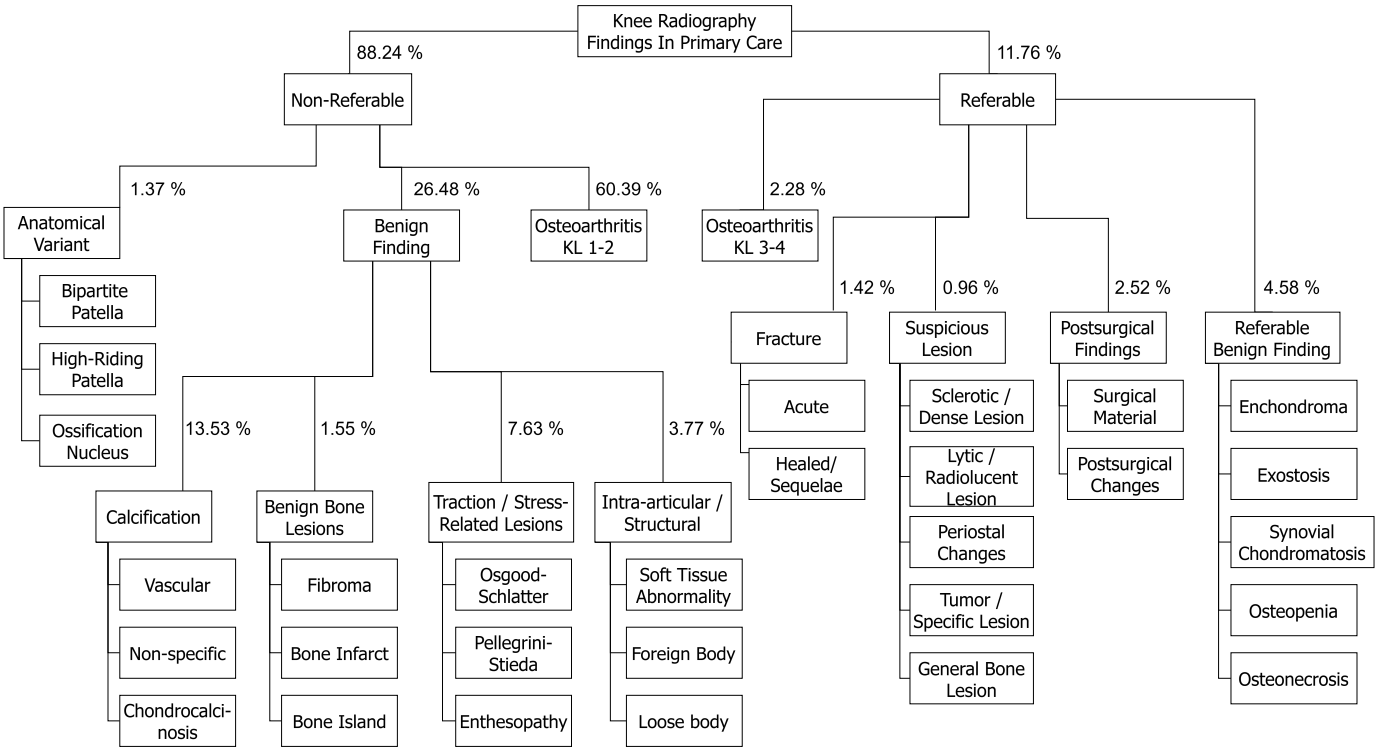


Figure 3. Taxonomy tree of knee radiological findings in primary care, with relative frequencies calculated from the 87,189 mapped entities.

validation set of 10,711 images was reserved, ensuring patient-level separation. During DICOM-to-JPG conversion, we applied the windowing configuration stored in each DICOM study, to preserve clinically optimal brightness and contrast.

Leveraging the Fast.ai library [21], the model was initialized with ImageNet weights and trained on 900×900 padded images using cross-entropy loss, default augmentations, and a learning rate of 0.01. The best performance was achieved at epoch 6, with an AUC of 0.8413 and an accuracy of 89.85% on the validation set.

We tried several variations in architecture, image size, learning rate, and training duration, but none yielded meaningful improvements. We chose a ConvNeXt-Small architecture for its strong ImageNet results, fast inference, and modern design, serving as a robust baseline. Further gains are expected with targeted model refinement, improved preprocessing and enhanced training strategies tailored to our dataset.

D. Reference Standard Test Set

We created a random test set of 494 radiographic single-knee studies (anteroposterior and lateral images), enriched to increase the proportion of referable knees from 11.60% to 39.70%, ensuring one knee per patient and strict patient-level separation to prevent data leakage. Each study was independently reviewed by three expert radiologists drawn from a pool of five, with each radiologist reviewing a different but overlapping random test set partition (around 250 studies). Reviewers identified all relevant findings according to the predefined categories of the taxonomy (Section III-B), and a study was deemed referable if any finding mapped to a referable category.

The image model was evaluated against the three-radiologist majority vote, which served as the clinical reference standard, and was also compared with each individual radiologist, using the majority vote of the other two as ground truth (excluding cases where the two remaining radiologists disagreed). In parallel, all test set reports were manually annotated by engineers based solely on their clinical text, providing the reference standard for evaluating the NLP pipeline.

IV. RESULTS

A. NLP pipeline validation

The NLP pipeline, when evaluated against the manually annotated reports from the test set, achieved micro- and macro-averaged F1-Scores of 0.8968 and 0.9219, respectively, across all finding categories. The lowest performance was observed for osteoarthritis KL 3–4 category (a referable category), with an F1-Score of 0.1714, while all other findings achieved F1-Scores above 0.7368. Detailed per-category results are provided in the Supplementary Material.

Error analysis revealed that misclassifications mainly arose from complex degenerative cases where the KL grade was not explicitly indicated in the report. For instance, reports often document "gonarthrosis" without explicit grading, despite providing a descriptive context that implies advanced osteoarthritis. In such cases, the pipeline tended to assign a KL 1–2 label, although these reports would likely be interpreted by a radiologist as KL 3–4. This systematic bias was more

frequent in the enriched test set, where KL grades were less often documented explicitly, than in the training set. Nevertheless, many KL 3–4 cases were correctly captured when the grade was explicitly stated, allowing the trained image model to successfully classify most KL 3–4 radiographs as referable in the test set (see Section IV-B).

B. Model validation

Image classification performance was evaluated against the majority vote of the radiologists (Table I). The model achieved a higher AUC than the manually annotated reports (0.8800 vs. 0.7983), with significance confirmed by 100,000 paired bootstrap iterations (p = 0.00019). At a matched sensitivity of 81.94%, the model achieved higher specificity than the reports, reaching 83.14% compared to 77.71%.

TABLE I. PERFORMANCE ON THE TEST SET FOR CLASSIFYING KNEE RADIOGRAPHS AS REFERABLE OR NON-REFERABLE.

Source	AUC	Kappa	Sensitivity[2]	Specificity
Model	0.8800 0.8434–0.9147	0.6098 0.5428–0.6778	81.94% 75.00–87.52	83.14% 79.14–86.86
Reports[1]	0.7983 0.7560–0.8361	0.5393 0.4669–0.6185	81.94% 75.69–88.19	77.71% 73.43–82.00

The ground truth was obtained with the majority vote of three expert radiologists. All confidence intervals are 95%, estimated via 1,000-iteration bootstrap resampling with replacement.

[1] Radiology reports, produced during routine clinical practice, were manually labeled here without image review.

[2] The model's operating point was adjusted to match the sensitivity of the radiology reports.

At a more granular level, Table II breaks down detection performance on the test set by grouped finding categories, where recalls reflect the proportion of correctly identified studies. Both the model and reports show minor differences identifying referable findings, and notably, they both detected all six suspicious lesions (osteosarcoma and periosteal reactions). However, the model was more effective at identifying normal cases, achieving higher recall for these than the reports (95% vs. 85%).

TABLE II. COMPARISON OF MODEL AND MANUALLY ANNOTATED REPORTS IN THE TEST SET, STRATIFIED BY FINDING CATEGORY.

Finding Category[1]	**Total**	**Model Recall**	**Report Recall**
Non-referable			
Benign finding	272	55.15%	52.57%
Osteoarthritis (KL 1–2)	208	66.83%	65.39%
Anatomical variant	145	66.21%	64.83%
No finding / Normal	80	95.00%	85.00%
Referable			
Osteoarthritis (KL 3–4)	59	83.05%	76.27%
Acute fracture	20	80.00%	90.00%
Suspicious lesion	6	100.00%	100.00%
Referable benign finding	24	70.83%	70.83%
Postsurgical findings	47	100.00%	97.87%
Healed / Sequelae fracture	9	100.00%	88.89%

[1] Please refer to the taxonomy tree (Figure 3) for a detailed decomposition.

Then, the model's performance was compared with individual radiologists and the manually annotated reports (Table III). Across all partitions (i.e., the subsets of studies reviewed by each radiologist), it consistently outperformed the reports in AUC, sensitivity, and specificity. Compared with individual radiologists, the model reached similar or slightly higher sensitivity (up to +14%), but generally lower specificity (3.6–11% below) and kappa (0.66–0.72 vs. 0.70–0.86). As shown in Figure 4, its ROC curves indicated strong performance, though still below 3 of 5 radiologists.

These results prompted a detailed error analysis, in which we reviewed each discordant case by reading the original reports and comparing them with the model outputs. We found that model errors were usually straightforward misclassifications—such as false negatives of fractures or positives for benign calcifications. In contrast, report errors often stemmed from textual ambiguity, incomplete descriptions or lack of sufficient clinical context, which either prevented the reader from extracting findings or rendered them too ambiguous to be reliably mapped into the taxonomy. Therefore, despite the measured metrics, the nature of the errors between the model and the reports differed.

V. CONCLUSION AND FUTURE WORK

We introduced a novel system for AI-assisted general interpretation of knee radiographs in primary care, leveraging automatically generated labels from real-world radiology reports through NLP. To support this, we developed a pathway-specific taxonomy that organizes findings into 43 categories and evaluated the resulting model against a reference test set annotated by expert radiologists. Importantly, the taxonomy was grounded in retrospective finding prevalence, ensuring the meaningfulness of both the categories and the resulting labels.

The model, trained to classify a knee radiograph as referable or non-referable, outperformed manually annotated routine radiology reports (AUC 0.880 vs. 0.798) and successfully referred all suspicious lesions—the most critical referable cases—while maintaining high specificity. Although its overall performance was generally below that of individual expert radiologists, it achieved comparable agreement in certain cases, suggesting its potential as a decision-support tool for general practitioners in primary care.

Error analysis showed that most report-related errors stemmed from the inherent challenges of clinical text interpretation (ambiguity, incomplete descriptions, or lack of context), while model errors reflected more straightforward misclassifications. In fact, our NLP pipeline, based on entity recognition and regular expressions, particularly struggled to assign osteoarthritis grades when they were not explicitly stated—introducing a performance ceiling for the model. These limitations highlight opportunities for improvement through advanced NLP methods, such as recent large language models [23].

While the test set was enriched to 39.70% referable cases to ensure statistical power for rare pathologies, we acknowledge that in a natural primary care prevalence (12.23%), the 3.6–11% specificity gap compared to radiologists might lead to increased false-positive referrals. Future deployment would require calibration and a configurable threshold to prioritize

TABLE III. PERFORMANCE COMPARISON PER INDIVIDUAL RADIOLOGIST ON TEST SET PARTITIONS.

	Radiologist 1 partition 252 knees, 27.8% referable			**Radiologist 2 partition** 246 knees, 24.4% referable			**Radiologist 3 partition** 259 knees, 23.9% referable			**Radiologist 4 partition** 253 knees, 30.8% referable			**Radiologist 5 partition** 250 knees, 22.4% referable		
	AUC	Kappa	Sens / Spec	AUC	Kappa	Sens / Spec	AUC	Kappa	Sens / Spec	AUC	Kappa	Sens / Spec	AUC	Kappa	Sens / Spec
Radiologist	0.9220	0.8591	87.14 / 97.25	0.9204	0.8178	90.00 / 94.09	0.8826	0.6995	88.71 / 87.82	0.8248	0.7073	66.67 / 98.29	0.9167	0.7836	91.07 / 92.27
Model	0.9175	0.6963	88.57 / 86.26	0.9429	0.6883	88.33 / 87.10	0.9231	0.6553	90.32 / 84.26	0.9049	0.7220	80.77 / 91.43	0.9292	0.6902	89.29 / 87.63
Reports	0.8313	0.5877	87.14 / 79.12	0.8312	0.5664	86.67 / 79.57	0.8450	0.5755	90.32 / 78.68	0.8367	0.6393	83.33 / 84.00	0.8536	0.5941	89.29 / 81.44

Each radiologist annotated a different partition of the test set, resulting in six overlapping partitions whose union constitutes the entire test set.
For each partition, the ground truth was defined by the majority vote of the other two radiologists; samples without consensus were excluded.
Sens / Spec denote sensitivity and specificity, respectively; values are reported as percentages.

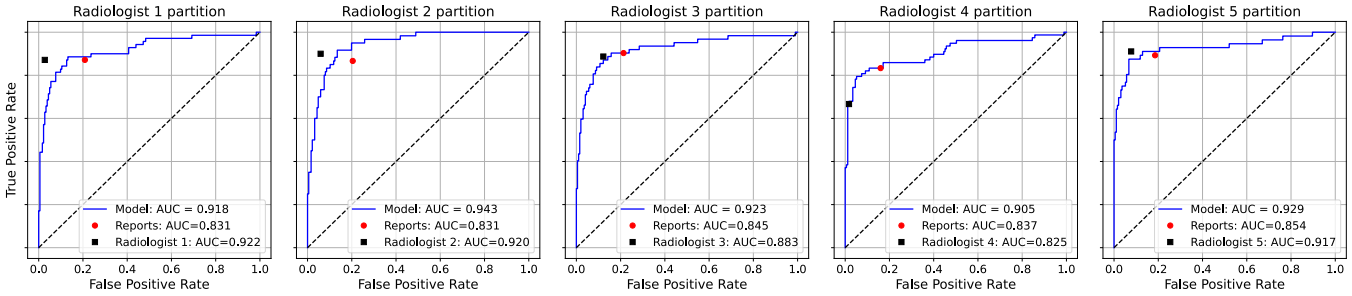


Figure 4. ROC curve of the model on the test set partitions. The operating points of radiologists and manually labeled radiology reports are overlaid.

either high-sensitivity screening or high-specificity workload reduction.

As a first step toward clinical use, we aimed to provide general practitioners with a tool that could help avoid unnecessary referrals, reduce radiologist workload, and shorten diagnostic times. We acknowledge that this approach simplifies decision-making in primary care by grouping important findings together, regardless of urgency or pathway—for example, advanced osteoarthritis and suspicious lesions are both classified as referable. Nevertheless, our results illustrate the feasibility of developing clinically meaningful AI systems based on NLP pipelines guided by pathway-specific taxonomies, offering a scalable strategy to extend AI to other body regions or imaging modalities, especially in contexts with limited labeled data but abundant radiology reports.

This study has several limitations: the small test set size and rarity of many findings limited per-finding evaluation; no external dataset was available to assess generalizability; weak labels derived from radiology reports may miss visible but unreported findings; plain radiographs have inherent diagnostic limitations; and potential selection bias arises from including only studies requested by general practitioners.

Future work will focus on developing multilabel models to stratify findings and enhance explainability, further improving the NLP pipeline, expanding the test set, and using pseudo-labeling to incorporate images without radiology reports. Prospective, multicenter, and multimodal studies will be essential to evaluate real-world impact and ensure safe integration into clinical workflows.

ETHICS STATEMENT

This work was approved by the SNS-O. No patient consent was required, as all clinical data were fully anonymized.

ACKNOWLEDGMENTS

This work was made possible thanks to the contributions of Borja de la Osa, Aritz Oscoz, Jimena Angarita, Marta Tirapu, Ana Andueza and Ainhoa Ovelar.

REFERENCES

[1] R. Haas *et al.*, "Prevalence and characteristics of musculoskeletal complaints in primary care: An analysis from the population level and analysis reporting (POLAR) database," *BMC Primary Care*, vol. 24, no. 1, pp. 1–10, 2023.

[2] K. P. Jordan *et al.*, "Annual consultation prevalence of regional musculoskeletal problems in primary care: An observational study," *BMC Musculoskeletal Disorders*, vol. 11, no. 1, pp. 144–152, 2010.

[3] F. C. Oettl *et al.*, "Artificial intelligence-assisted analysis of musculoskeletal imaging—a narrative review of the current state of machine learning models," *Knee Surgery, Sports Traumatology, Arthroscopy*, vol. 33, no. 1, pp. 24–38, 2025.

[4] G. U. Kim, M. C. Chang, D. H. Sung, J. B. Song, and H. J. Park, "Diagnostic modality in spine disease: A review," *Asian Spine Journal*, vol. 14, no. 6, pp. 910–920, 2020.

[5] S. Newman, H. Ahmed, and N. Rehmatullah, "Radiographic vs. MRI vs. arthroscopic assessment and grading of knee osteoarthritis—are we using appropriate imaging?" *Journal of Experimental Orthopaedics*, vol. 9, no. 1, pp. 2–11, 2022.

[6] K. A. Thomas *et al.*, "Automated classification of radiographic knee osteoarthritis severity using deep neural networks," *Radiology: Artificial Intelligence*, vol. 2, no. 2, pp. 2638–6100, 2020.

[7] A. Nowroozi *et al.*, "Artificial intelligence diagnostic accuracy in fracture detection from plain radiographs and comparing it with clinicians: A systematic review and meta-analysis," *Clinical Radiology*, vol. 79, no. 8, pp. 579–588, 2024.

[8] A. A. Bajjad, F. S. Al-Shehri, and S. M. Al-Malki, "Artificial intelligence in bone age assessment of healthy individuals: A scoping review," *Journal of the World Federation of Orthodontists*, vol. 13, no. 2, pp. 95–102, 2024.

[9] C. E. von Schacky *et al.*, "Multitask deep learning for segmentation and classification of primary bone tumors on radiographs," *Radiology*, vol. 301, no. 2, pp. 398–406, 2021.

[10] S. Gitto *et al.*, "AI applications in musculoskeletal imaging: A narrative review," *European Radiology Experimental*, vol. 8, no. 1, pp. 22–35, 2024.

[11] R. Lindsey *et al.*, "Deep neural network improves fracture detection by clinicians," *Proceedings of the National Academy of Sciences*, vol. 115, no. 45, pp. 11 591–11 596, 2018.

[12] P. G. Anderson *et al.*, "Deep learning assistance closes the accuracy gap in fracture detection across clinician types," *Clinical Orthopaedics and Related Research*, vol. 481, no. 3, pp. 580–588, 2023.

[13] A. Tiulpin and S. Saarakkala, "Automatic grading of individual knee osteoarthritis features in plain radiographs using deep convolutional neural networks," *Diagnostics*, vol. 10, no. 11, pp. 932–945, 2020.

[14] P. Rajpurkar *et al.*, "MURA: Large dataset for abnormality detection in musculoskeletal radiographs," *arXiv preprint arXiv:1712.06957*, pp. 1–11, 2018.

[15] F. Eckstein, W. Wirth, and M. C. Nevitt, "Recent advances in osteoarthritis imaging: The osteoarthritis initiative," *Nature Reviews Rheumatology*, vol. 8, no. 10, pp. 622–630, 2012.

[16] A. Bustos, A. Pertusa, J. M. Salinas, and M. de la Iglesia-Vayá, "PadChest: A large chest x-ray image dataset with multi-label annotated reports," *Medical Image Analysis*, vol. 66, p. 101 797, 2020.

[17] J. Irvin *et al.*, "CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, pp. 590–597, 2019.

[18] C. P. Carrino *et al.*, "Pretrained biomedical language models for clinical NLP in Spanish," in *Proceedings of the 21st Workshop on Biomedical Language Processing*, Association for Computational Linguistics, 2022, pp. 193–199.

[19] A. J. Tamayo, *Negation scope detection in spanish clinical texts using mBERT fine-tuned on the NUBEs dataset*, https://github.com/ajt/NegScope, 2025.

[20] S. Lima López, N. Perez, M. Cuadros, and G. Rigau, "NUBes: A corpus of negation and uncertainty in spanish clinical texts," in *Proceedings of the 12th Language Resources and Evaluation Conference*, European Language Resources Association, 2020, pp. 5772–5781.

[21] J. Howard and S. Gugger, "Fastai: A layered API for deep learning," *Information*, vol. 11, no. 2, pp. 108–122, 2020.

[22] R. Wightman, *Pytorch image models*, https://github.com/huggingface/pytorch-image-models, version 1.2.2, 2019.

[23] S. H. Kim *et al.*, "Benchmarking the diagnostic performance of open-source LLMs in 1,933 Eurorad case reports," *npj Digital Medicine*, vol. 8, no. 1, pp. 1–9, 2025.

Implementation of AI Characters for Simulation of Root Cause Analysis in the ICU Setting

Yuqi Hu[1], Qiwen Xiong[1], Zhenzhen Qin[1], Brandon Watanabe[2], Yujing Wang[1], Ilmi Yoon[1*]

[1]Northeastern University [2]San Francisco State University
[*]Corresponding to: i.yoon@northeastern.edu

***Abstract*—Root Cause Analysis (RCA) is widely used to investigate adverse events in healthcare settings, yet hands-on RCA training is often constrained by instructor time, standardized patient availability, and logistical burden. This paper documents the implementation of an AI-based RCA simulation for an ICU adverse event case, with an emphasis on prompt engineering for large language model (LLM) components. Specifically, we present necessary details for implementation, including: (1) a fully specified ICU case narrative, (2) role schemas and system prompts for virtual avatars, (3) states-of-mind prompts that modulate interviewee behavior, (4) voice design specifications for virtual avatar, and (5) analytic rubrics and assessment prompts for formative and summative feedback. We also include examples of LLM-generated dialogue and assessment reports to illustrate expected outputs and formats. Lastly, we discuss limitations and future directions for the presented system.**

Keywords-root cause analysis; AI; virtual avatars; healthcare education; ICU; formative assessment

I. Introduction

Adverse events remain a major patient safety challenge. Estimates suggest that more than 250,000 patients in the United States experience adverse events each year [1], in which medical treatment itself contributes to patient harm [2]. Root Cause Analysis (RCA) is widely used to investigate such events and reduce recurrence by focusing on system-level contributors rather than individual blame [3][4]. RCA is broadly adopted across healthcare organizations to examine failures involving workflows, communication, and policy gaps [5][6]. In large systems, RCAs occur routinely; for example, Veteran Affairs medical centers report multiple RCAs per year on average [7]. Although procedures vary, most RCAs follow a common structure: describing the event, identifying contributing factors, determining root causes, and proposing corrective actions [8][9].

Despite its value, RCA is often inconsistently executed, in part because the skills it requires are difficult to develop without repeated practice. Effective investigations depend on eliciting detailed accounts, reconciling conflicting perspectives, distinguishing proximate from systemic contributors, and translating findings into feasible corrective actions [10][11]. Yet hands-on RCA instruction is frequently constrained by instructor time and logistical burden [12]. Although RCA content appears in medical education, nursing programs, and continuing education, training depth and consistency vary widely [13]–[17]. As a result, learners may receive limited opportunities to practice the interpersonal and investigative components of RCA, particularly interviewing stakeholders who may be defensive, uncertain, or emotionally affected by the event.

Recent advances in AI, especially large language models (LLMs), create opportunities for scalable, interactive practice. LLMs can generate coherent, context-sensitive dialogue and can be shaped through prompt engineering to produce role-consistent responses and controlled behavioral variability [18]. In healthcare education, LLM-enabled simulations have been explored to support clinical reasoning and communication skills through interactive virtual agents [19]–[21]. LLMs can also generate structured formative feedback and evaluate open-ended learner work when paired with explicit rubrics, improving the actionability of feedback while reducing educator workload [22]–[24].

In this paper, we document the implementation of an AI-based RCA simulation for an ICU adverse event case. The system builds on *Nurse Town*, a Unity-based 3D simulation game for nursing education [25], and adapts a textbook ICU failure scenario [26]. We focus on replication-oriented artifacts and design decisions, including: (1) a fully specified case narrative, (2) role schemas for five virtual ICU team member avatars, (3) states-of-mind prompts that modulate disclosure style, (4) voice design targets that support role-appropriate affect, (5) a structured RCA report template, and (6) analytic rubrics for formative interview feedback and summative report assessment.

II. Related Work

A. Root Cause Analysis and Patient Safety Training

RCA gained prominence in the 1990s quality improvement movement as healthcare organizations increasingly recognized that harm often stems from system failures rather than isolated individual mistakes [27]. In the United States, RCA was promoted as a key method for investigating adverse events and preventing recurrence [28], and it became institutionalized through accreditation and safety oversight (e.g., sentinel event analyses) [29][30]. Guidance from major safety organizations has further supported broad adoption in both clinical and administrative domains [31], and structured RCA programs have been used in large healthcare systems to operationalize patient safety initiatives [32]. RCA's emphasis on learning-oriented, system-focused improvement can support a culture of safety, shifting attention away from blame and toward actionable redesign [33][34].

However, RCA effectiveness depends heavily on execution quality. Teams must gather accurate accounts, navigate interpersonal dynamics during interviews, and synthesize evidence into root causes and measurable corrective actions [3][9]. Training gaps are common, and inconsistent preparation can limit the rigor and impact of RCA findings [10][11]. While RCA concepts are taught in professional programs, scalable hands-on practice remains difficult; simulation-based or case-based interventions can improve realism but are often constrained by staffing and logistics [12][13][15].

B. LLM-Enabled Simulation and Automated Feedback

LLMs can sustain multi-turn, context-sensitive dialogue, enabling conversational agents that simulate realistic interviews for educational purposes [18]. With prompt engineering and role constraints, LLM-driven agents can maintain persona consistency, selectively disclose information, and express interview stances that mirror real-world dynamics (e.g., defensiveness, uncertainty) that are central to RCA practice [35]. In healthcare education, such agents have been explored to support clinical reasoning and communication skill development through interactive practice [19]–[21]. LLMs can also support assessment and feedback on open-ended learner work; prior studies suggest they can generate formative coaching and evaluate text outputs, especially when guided by analytic rubrics that clarify expectations and improve scoring consistency [22]–[24]. These findings motivate our use of rubric-guided assessment and prompt-controlled behavioral variability within an RCA simulation context.

III. Simulation System Implementation

A. System Overview and Session Flow

The RCA training experience is delivered as a Unity-based 3D simulation in which learners investigate an ICU adverse event by interviewing five virtual ICU team members. Figure 1 illustrates the overall architecture. A typical session begins with the learner reviewing the case background and RCA instructions in an in-game help panel, then selecting an interviewee (e.g., primary nurse, ICU nurse, physician, respiratory therapist, or medical student). Interviews are conducted as speech-to-speech conversations: the learner speaks a question, the system captures audio and transcribes it via speech-to-text (STT), and the transcript is appended to the ongoing interview record. The transcript is then sent to an LLM-driven dialogue component that generates a role-consistent response conditioned on (i) the case facts, (ii) the character schema, and (iii) a sampled state-of-mind prompt that controls the interviewee's disclosure style (e.g., defensive vs. self-reflective). Next, the generated response is rendered through emotional text-to-speech (TTS), while the avatar's facial expression, lip sync, and body gestures are animated to align with the intended affect.

After completing all interviews, the learner transitions to a synthesis phase in which they complete a structured RCA report using the provided template. The report is uploaded within the system and evaluated by a separate LLM-based assessment module guided by analytic rubrics. The module produces two

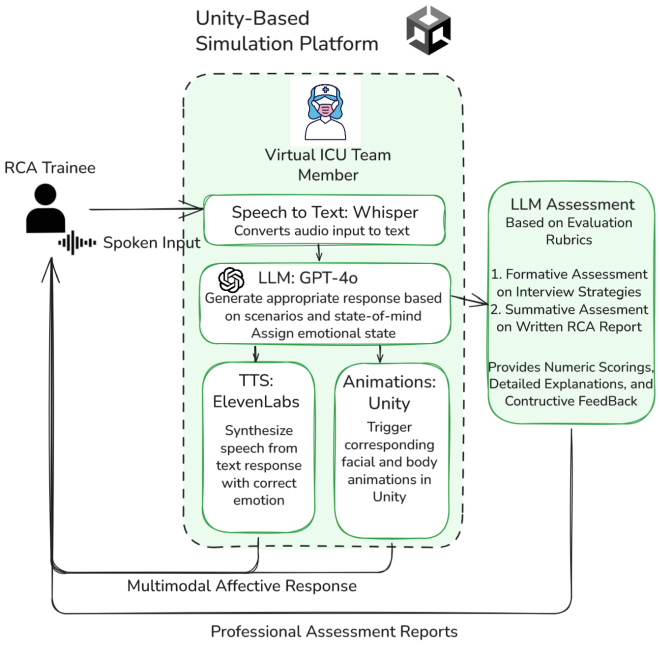


Figure 1. Architecture of the virtual RCA simulation system.

deliverables: (1) formative feedback on interviewing behaviors (e.g., depth of inquiry, follow-up quality, and theme coverage) based on the interview transcripts, and (2) summative scoring and narrative feedback on the written RCA report (e.g., clarity of problem statement, identification of direct and root causes, feasibility of corrective actions, and use of interview evidence). Feedback is returned to the learner in a structured report view that can be revisited to support iterative practice and improvement. An example of learner-virtual character dialogue and feedback report can be found in the Appendix.

B. Simulation Scenario Overview

The simulation centers on an ICU adverse event in which inconsistent wristband conventions and communication breakdowns during an emergency contribute to a fatal outcome. A patient admitted for IV antibiotics for pneumonia was initially stable; during intake, the primary nurse identified a penicillin allergy and applied a red wristband for "no blood draw" and a blue wristband mistakenly believed to indicate "allergy" based on another hospital's convention, even though in this hospital blue signifies "Do Not Resuscitate" (DNR). Shortly after the infusion began, the patient developed signs of anaphylaxis; the nurse stopped the infusion, administered oxygen, and called a code when the patient became unresponsive. During resuscitation, an ICU nurse noticed the blue wristband and raised concern that the patient might be DNR, prompting the team to pause and verify code status while the primary nurse searched for the chart; during this delay, the rhythm deteriorated from ventricular tachycardia to asystole, and despite resuming efforts after confirming full-code status, the patient was pronounced dead. The case surfaces multiple

system contributors relevant to RCA training, including non-standardized wristband color systems, unclear ownership of wristband application, disorganized supplies, communication breakdowns under time pressure, and fatigue-related risk from excessive work hours.

C. Virtual Characters in the Simulation

The simulation includes five interviewees representing key ICU stakeholders involved in the event. Each character is instantiated as a unique 3D avatar with an LLM-driven persona, specified with (i) professional responsibilities and perspective, (ii) the subset of case facts the character is expected to know, (iii) likely concerns or biases during an RCA interview, and (iv) communication tendencies (e.g., willingness to elaborate, formality, and stance toward accountability). In addition to language, each character is assigned a voice profile to support immersion and to cue typical interpersonal dynamics encountered in real RCA interviews (e.g., fatigue, hesitancy, urgency). Table I summarizes each character's role and key attributes/actions, along with the corresponding voice design targets used to realize their speech.

Pedagogically, the *Primary Nurse* serves as the central witness and is positioned to reveal frontline workflow pressures (e.g., workload and fatigue) as well as local conditions that enabled the wristband mix-up. The *Code Team Medical Student* represents a junior member who can surface hierarchy and speaking-up dynamics during emergencies, creating opportunities for learners to practice eliciting information from hesitant informants. The *Code Team ICU Nurse* is designed as a more assertive safety advocate who identifies the wristband ambiguity during the code, helping learners probe system-level safeguards and workaround behaviors. The *Code Team Doctor* reflects leadership priorities and time-critical decision-making under pressure, allowing learners to explore how urgency and institutional metrics can shape actions during a code. Finally, the *Respiratory Therapist/Anesthesiologist* focuses on airway management and visual status indicators, prompting learners to examine whether safety signals were sufficiently clear and redundant for high-acuity settings.

To ensure that each LLM-driven avatar reliably maintains its assigned persona, we use a layered prompt structure with explicit constraints. First, a *role system prompt* encodes the character schema (profession, responsibilities, viewpoint, and communication style) and establishes hard boundaries on what the avatar can claim to have seen or known. Second, a *case grounding block* provides the incident narrative and a character-specific "known facts" list, reducing hallucination by constraining answers to an approved knowledge set. Third, a short *behavioral policy* enforces interaction norms (e.g., answer strictly as the character; do not provide omniscient summaries; if uncertain, state uncertainty and ask for clarification; avoid inventing documentation that was not reviewed; keep answers within a reasonable length; and remain consistent with prior turns). In practice, these constraints are complemented by lightweight runtime checks (e.g., maintaining a rolling memory of prior claims and re-injecting a compact persona reminder) to reduce drift over longer interviews.

Beyond textual persona control, we assign each character a voice target that is consistent with their role and typical affect during an RCA interview. These voice profiles function as an additional modality-level constraint that helps preserve characterization (e.g., a fatigued nurse sounds strained rather than exuberant; a junior learner sounds hesitant; a physician leader sounds steady and commanding). Operationally, we treat voice design as part of the persona specification: each role schema includes a short voice descriptor (tone, pace, intensity, and emotional range), which is used to select or configure the TTS voice and to maintain consistent delivery across turns.

D. Assigning States-of-Mind to Virtual Avatars

To reduce "scripted" interviews and train adaptive questioning, each avatar adopts one of five states-of-mind at the start of an interview. The state-of-mind does not change the underlying case facts the character may disclose; instead, it modulates *how* information is presented (e.g., level of detail, willingness to accept responsibility, emotional tone, and tendency to elaborate). This design encourages learners to practice follow-up strategies, clarification requests, and rapport-building techniques when encountering reluctance, uncertainty, or frustration.

We implement states-of-mind using a dedicated *style controller prompt* that is injected alongside the fixed character schema. The prompt specifies (i) the target interview stance (e.g., defensive vs. self-reflective), (ii) the preferred linguistic markers (e.g., hedging, minimization, apology, formality), and (iii) disclosure guidelines (e.g., "do not volunteer extra details unless asked" for detached; "redirect blame unless pressed" for defensive). Importantly, we separate *content constraints* from *style constraints*: the character schema and case-grounding block define what the avatar can know and say, while the state-of-mind prompt determines how forthcoming, emotional, or detailed the delivery should be. This separation reduces the risk that style modulation changes factual content.

To keep behavior consistent across long interviews, the state-of-mind is reiterated in a short reminder prefix for each turn (e.g., "Maintain a frustrated tone; emphasize system issues; avoid admitting fault unless directly asked."). We also include guardrails that prevent the avatar from breaking character (e.g., "do not speak as an AI"; "do not narrate the simulation") and from introducing facts outside the approved "known facts" list. When the learner asks ambiguous questions, the prompt instructs the avatar to request clarification rather than guessing, which helps preserve both persona fidelity and factual consistency.

E. Automated Evaluation of Simulation Sessions

The simulation includes an automated evaluation component that assesses learner performance across two complementary products of an RCA investigation: (i) the learner's interview process and (ii) the learner's final written RCA report. At a high level, the system treats each session as a structured record composed of interview transcripts (time-stamped learner

TABLE I. SUMMARY OF VIRTUAL CHARACTERS IN THE ICU FAILURE SIMULATION, INCLUDING ROLE ATTRIBUTES/ACTIONS AND VOICE DESIGN TARGETS.

Character	Key attributes and actions	Voice design target
Primary Nurse	Fatigued and overworked; confused wristband codes due to experience at another hospital; delayed resuscitation while searching for the chart.	Weary and slightly breathy tone reflecting fatigue; professional but strained; may convey frustration or defensiveness.
Code Team Medical Student	Noticed the blue wristband but unsure of its meaning; hesitant to speak up; relied on others for direction.	Soft-spoken and hesitant with pauses indicating uncertainty; youthful and tentative.
Code Team ICU Nurse	Immediately raised concern about the wristband; frustrated by lack of standard DNR indicators; suggested DNR signage above beds.	Direct and experienced; clear articulation; slightly impatient at times but professional.
Code Team Doctor	Led the code team; focused on rapid defibrillation; frustrated by the delay; emphasized teamwork and clear protocols.	Calm and commanding; steady pacing; urgency with control.
Respiratory Therapist / Anesthesiologist	Managed airway; questioned reliance on wristbands alone; recommended better visual indicators for patient status.	Pragmatic and task-focused; concise delivery; occasional disbelief/concern when noting system flaws.

questions and avatar responses) and a completed RCA report template submitted at the end of the scenario. These artifacts are then analyzed by an LLM-based assessment module guided by analytic rubrics, producing both formative feedback (to improve interviewing technique) and summative feedback (to evaluate the quality of RCA synthesis and recommendations).

After completing interviews with the five virtual team members, learners transition to a synthesis phase and complete an RCA report using a structured template. The template first captures basic event metadata, including the event title, date of the event, dates the RCA was initiated and completed, and the names and roles of the facilitator and team members. It then guides learners through a stepwise RCA workflow. Learners begin by selecting and describing the event, articulating why the incident warrants investigation. They next document the RCA charter and team formation, then build a detailed timeline to establish a shared narrative of what occurred. The template prompts learners to identify contributing factors linked to specific steps in the timeline, then iteratively identify root causes using structured "why" reasoning. Finally, learners propose corrective actions that target identified root causes and specify how success will be measured, encouraging a shift from description to prevention-oriented planning.

The assessment module uses analytic rubrics to ensure feedback is interpretable and actionable. Each rubric dimension is scored and accompanied by a brief rationale, a description of strengths and weaknesses, and concrete suggestions for improvement. For *formative* evaluation of interview performance, the rubric emphasizes the learner's depth of inquiry (e.g., use of open-ended questions and targeted follow-ups), comprehensiveness of investigation (e.g., coverage of actions, decisions, communication breakdowns, and system-level contributors such as fatigue or unclear protocols), and active listening/adaptability (e.g., probing inconsistencies and adjusting questioning strategy in response to new information). The formative rubric also evaluates whether the learner identifies and pursues key themes relevant to the case (e.g., wristband meaning, role ownership, protocol gaps, workload pressures, emergency communication) and whether the learner maintains professionalism and clarity throughout the interview.

For *summative* evaluation of the written RCA report, the rubric focuses on the clarity of the problem statement and event description, the correctness and completeness of identified direct causes and contributing factors, and the depth of systemic analysis (e.g., workflow design, protocol and training gaps, and communication structures). It further evaluates how effectively the report integrates interview evidence—including triangulation across perspectives and acknowledgement of contradictions or uncertainty—and whether proposed corrective actions are specific, feasible, and appropriately targeted to root causes rather than superficial fixes. Finally, the summative rubric considers overall organization, coherence, and writing quality, reflecting the expectation that an RCA report should be clear enough to support implementation and accountability.

IV. DISCUSSION AND CONCLUSION

This paper documents an implementation-oriented RCA simulation package for an ICU patient safety case, emphasizing scenario specification, character and persona design, behavioral variability via states-of-mind, learner deliverables, and rubric-guided assessment scaffolds. A key goal of this work is reusability: the artifacts are structured so they can be adopted in multiple instructional formats and with varying levels of technical infrastructure.

TABLE II. STATES-OF-MIND USED TO MODULATE AVATAR DISCLOSURE STYLE AND TONE.

State-of-mind	Behavioral signature in interviews	Example dialogue snippet
Defensive	Deflects blame; minimizes responsibility; may redirect to policies, other roles, or external constraints; can be vague unless prompted with specifics.	"Placing wristbands isn't really my job. I assumed the ED had handled it. The cabinet was a mess anyway."
Self-Reflective / Honest	Acknowledges mistakes; offers detailed chronology; expresses regret; more likely to volunteer contributing factors and opportunities for improvement.	"I should've double-checked the wristband color. I just assumed, and that was a serious mistake on my part."
Confused / Uncertain	Hesitant recall; inconsistent details; uses hedging language; requires clarification and timeline reconstruction.	"I think... it was the blue band that made everyone pause? Or maybe someone said something about DNR... I'm not totally sure."
Overly Professional / Detached	Formal and emotionally flat; answers narrowly and literally; avoids interpretation or speculation; limited volunteering of context.	"At the time of arrival, the patient was in V-tach. I began airway management. I cannot comment on code status decisions."
Frustrated	Emphasizes systemic problems; expresses dissatisfaction with processes and resources; may downplay personal mistakes; tends to "vent" about organizational issues.	"This wouldn't have happened if we had standardized wristband colors. It's ridiculous how disorganized things are here."

First, the case narrative and character role schemas can be used beyond the Unity simulation itself. Instructors may deploy them as paper-based RCA exercises, as interview-based assessments (e.g., OSCE-style stations), or as guided small-group activities in which learners practice eliciting perspectives and reconstructing a timeline. Second, the states-of-mind mechanism provides a lightweight way to introduce controlled variability in interview dynamics. By shifting disclosure style and emotional tone without altering the underlying facts, the system encourages learners to practice adaptive questioning strategies, clarification, and rapport-building in the face of defensiveness, uncertainty, or frustration. Third, the structured RCA template and analytic rubrics support consistent expectations across instructors and cohorts. They can be used with automated assessment to reduce faculty workload and provide rapid feedback, or they can be applied in instructor-led settings to standardize grading and feedback.

Overall, these materials are intended to lower barriers to adopting simulation-based RCA training and to support replication and adaptation across institutions and curricula. Future extensions include expanding the scenario library to cover a wider range of adverse events, refining persona control to reduce drift over long interviews, and evaluating how different configurations of automated feedback influence learning outcomes and user trust.

APPENDIX

A. Sample Interview Transcript (Excerpt)

Below is an excerpt demonstrating a "Frustrated" primary nurse interview style.

- Learner: Hello, I'm here to talk about the ICU incident earlier.
- Primary Nurse (Frustrated): Hi. I'm here to discuss that, too. It's been... quite overwhelming, to say the least. Where would you like to start?
- Learner: Why did the code team pause resuscitation?
- Primary Nurse (Frustrated): It was the wristband confusion. At my other job, blue means allergy. Here it means DNR. Everyone hesitated because of that.
- Learner: Have you ever seen this confusion before?
- Primary Nurse (Frustrated): Not here, but that's exactly the problem—different hospitals use different systems. It's frustrating.
- Learner: What would prevent this from happening again?
- Primary Nurse (Frustrated): Standardize wristband colors, and organize supplies. Also, make it crystal clear who is responsible for applying bands.
- Learner: Anything else contribute?
- Primary Nurse (Frustrated): Fatigue. I've been working nonstop. That absolutely clouded my judgment.

B. Sample Assessment Output Format (Formative)

- **Depth of Inquiry: 8/10.** Strong open-ended questions and relevant follow-ups; add targeted probes about decisions during the code.
- **Comprehensiveness: 7/10.** Covered communication/protocol issues; further explore staffing and training gaps.
- **Active Listening: 9/10.** Adapted questions to interviewee comments; probe inconsistencies more explicitly.

Summative report feedback (example formatting):

- **Problem Statement: 9/10.** Clear description of wristband misinterpretation and resuscitation delay.
- **Causes:** Immediate 8/10; Contributing 9/10. Strong analysis of fatigue and environment; add detail on hierarchy/speaking-up dynamics.
- **Solutions: 9/10.** Practical and targeted; include accountability and measurement plans.

REFERENCES

[1] J. G. Anderson and K. Abrahamson, "Your health care may kill you: Medical errors", in *Building Capacity for Health Informatics in the Future*, IOS Press, 2017, pp. 13–17.

[2] M. Jason Boulanger and C. Keohane, "Role of patient safety organizations in improving patient safety", *Patient Safety in Obstetrics and Gynecology, An Issue of Obstetrics and Gynecology Clinics*, vol. 46, no. 2, p. 257, 2019.

[3] J. J. Rooney and L. N. Heuvel, "Root cause analysis for beginners", *Quality progress*, vol. 37, pp. 45–56, 2004.

[4] I. Tjia et al., "Wake up safe and root cause analysis: Quality improvement in pediatric anesthesia", *Anesthesia & Analgesia*, vol. 119, no. 1, pp. 122–136, 2014.

[5] E. Flink et al., "Lessons learned from the evolution of mandatory adverse event reporting systems", in *Advances in Patient Safety: From Research to Implementation (Volume 3: Implementation Issues)*. Agency for Healthcare Research and Quality, 2005.

[6] K. M. Kellogg et al., "Our current approach to root cause analysis: Is it contributing to our failure to improve patient safety?", *BMJ Quality & Safety*, vol. 26, no. 5, pp. 381–387, 2017.

[7] K. B. Percarpio, B. V. Watts, and W. B. Weeks, "The effectiveness of root cause analysis: What does the literature tell us?", *The Joint Commission Journal on Quality and Patient Safety*, vol. 34, no. 7, pp. 391–398, 2008.

[8] P. M. Williams, "Techniques for root cause analysis", in *Baylor University Medical Center Proceedings*, vol. 14, Taylor & Francis, 2001, pp. 154–157.

[9] P. Gangidi, "A systematic approach to root cause analysis using 3×5 whys technique", *International Journal of Lean Six Sigma*, vol. 10, no. 1, pp. 295–310, 2019.

[10] A. W. Wu, A. K. Lipshutz, and P. J. Pronovost, "Effectiveness and efficiency of root cause analysis in medicine", *Jama*, vol. 299, no. 6, pp. 685–687, 2008.

[11] R. J. Latino, "How is the effectiveness of root cause analysis measured in healthcare?", *Journal of Healthcare Risk Management*, vol. 35, no. 2, pp. 21–30, 2015.

[12] M. Aboumrad, J. Neily, and B. V. Watts, "Teaching root cause analysis using simulation: Curriculum and outcomes", *Journal of Medical Education and Curricular Development*, vol. 6, p. 2 382 120 519 894 270, 2019.

[13] L. W. Hall et al., "Effectiveness of patient safety training in equipping medical students to recognise safety hazards and propose robust interventions", *BMJ Quality & Safety*, vol. 19, no. 1, pp. 3–8, 2010.

[14] M. A. Dolansky, K. Druschel, M. Helba, and K. Courtney, "Nursing student medication errors: A case study using root cause analysis", *Journal of Professional Nursing*, vol. 29, no. 2, pp. 102–108, 2013.

[15] J. Lambton and L. Mahlmeister, "Conducting root cause analysis with nursing students: Best practice in nursing education", *Journal of Nursing Education*, vol. 49, no. 8, pp. 444–448, 2010.

[16] P. Bowie, J. Skinner, and C. de Wet, "Training health care professionals in root cause analysis: A cross-sectional study of post-training experiences, benefits and attitudes", *BMC Health Services Research*, vol. 13, pp. 1–10, 2013.

[17] M. Murphy et al., "Implementation of a mock root cause analysis to provide simulated patient safety training", *BMJ Open Quality*, vol. 6, no. 2, e000096, 2017.

[18] H. Naveed et al., "A comprehensive overview of large language models", *arXiv preprint arXiv:2307.06435*, 2023.

[19] D. P. Mohapatra et al., "Leveraging large language models (llm) for the plastic surgery resident training: Do they have a role?", *Indian Journal of Plastic Surgery*, vol. 56, no. 5, pp. 413–420, 2023.

[20] C. Levin, M. Suliman, E. Naimi, and M. Saban, "Augmenting intensive care unit nursing practice with generative ai: A formative study of diagnostic synergies using simulation-based clinical cases", *Journal of Clinical Nursing*, 2024.

[21] N. Kapadia et al., "Evaluation of large language model generated dialogues for an ai based vr nurse training simulator", in *International Conference on Human-Computer Interaction*, Springer, 2024, pp. 200–212.

[22] T. Ruwe and E. Mayweg-Paus, "Embracing llm feedback: The role of feedback providers and provider information for feedback effectiveness", in *Frontiers in Education*, Frontiers Media SA, vol. 9, 2024, p. 1 461 362.

[23] J. Schneider, B. Schenk, and C. Niklaus, "Towards llm-based autograding for short textual answers", *arXiv preprint arXiv:2309.11508*, 2023.

[24] W. Xie, J. Niu, C. J. Xue, and N. Guan, "Grade like a human: Rethinking automated assessment with large language models", *arXiv preprint arXiv:2405.19694*, 2024.

[25] Y. Hu, Q. Xiong, L. Yi, and I. Yoon, "Nurse town: An llm-powered simulation game for nursing education", in *2025 IEEE Conference on Artificial Intelligence (CAI)*, IEEE, 2025, pp. 215–222.

[26] S. H. Campbell and K. Daley, Eds., *Simulation scenarios for nursing educators: making it real*. Springer Publishing Company, 2017.

[27] L. Emanuel et al., "What exactly is patient safety?", *Journal of Medical Regulation*, vol. 95, no. 1, pp. 13–24, 2009.

[28] D. H. Havens and L. Boroughs, ""to err is human": A report from the institute of medicine", *Journal of Pediatric Health Care*, vol. 14, no. 2, pp. 77–80, 2000.

[29] J. Sherwin, "Contemporary topics in health care: Root cause analysis", *PT Motion*, vol. 3, pp. 28–31, 2011.

[30] The Joint Commission, *Sentinel event policy and procedures*, https://www.jointcommission.org/resources/sentinel-event/, Accessed: April 3, 2025, 2025.

[31] I. for Healthcare Improvement, *Patient safety 104: Root cause and systems analysis summary sheet*, https://www.ihi.org/sites/default/files/lms/legacy/education/IHIOpenSchool/Courses/Documents/SummaryDocuments/PS%20104%20SummaryFINAL.pdf, Accessed: April 3, 2025, 2015.

[32] J. P. Bagian et al., "Developing and deploying a patient safety program in a large health care delivery system: You can't fix what you don't know about", *The Joint Commission Journal on Quality Improvement*, vol. 27, no. 10, pp. 522–532, 2001.

[33] R. A. M. Iedema et al., "Turning the medical gaze in upon itself: Root cause analysis and the investigation of clinical error", *Social Science & Medicine*, vol. 62, no. 7, pp. 1605–1615, 2006.

[34] C. Vincent, "Understanding and responding to adverse events", *New England Journal of Medicine*, vol. 348, no. 11, pp. 1051–1056, 2003.

[35] C. Grévisse, "Raspatient pi: A low-cost customizable llm-based virtual standardized patient simulator", in *International Conference on Applied Informatics*, 2024, pp. 125–137.

De-Identifying German Clinical Notes Under Small-Corpus Constraints

Transferring State-of-the-Art Approaches from English Benchmarks

Anna-Lena Artmann
Data Science and Artificial Intelligence
Center for Advanced Studies of the Baden-Württemberg Cooperative State University
Heilbronn, Germany
e-mail: cas417266@cas.dhbw.de

***Abstract*-Most de-identification methods are trained on English corpora, limiting cross-lingual transfer to German where annotated data are scarce. We evaluate transferability of state-of-the-art approaches via a systematic review (Jan 2023 - Apr 2025; 55 publications) and a controlled experiment on Graz Synthetic Clinical text Corpus. We compare BiLSTM-CRF, gELECTRA, and an 8B LLaMA-3 variant (SauerkrautLM), reporting entity-level precision, recall, and F1 score. The review shows a continued shift to transformers and hybrids but limited cross-language comparability due to heterogeneous datasets and metrics. On GraSCCo, BiLSTM-CRF is a reliable baseline (entity-level Micro-F1 = 0.96, Macro-F1 = 0.95). gELECTRA performs well on structured identifiers but drops on rare or variable categories (micro-F1 0.66) due to data/label sparsity. One-shot decoding with Llama-3.1-SauerkrautLM was unreliable under on-premises constraints. For German small-corpus settings, compact encoders are the most pragmatic near-term solution, while larger corpora will be necessary to realize full potential of transformer encoders like gELECTRA.**

Keywords-De-identification; GSA region; gELECTRA; BiLSTM; GraSCCo.

I. INTRODUCTION

Clinical narratives are central to data-driven medical research, as they capture rich contextual information that structured health data cannot provide [1]. To enable secondary use, however, all Personally Identifiable Information (PII) must first be removed [2]. This process, known as de-identification, poses particular challenges for free-text records, which are less predictable than structured fields [3].

Over the past decade, de-identification research has evolved considerably. Rule-based and hybrid systems remain relevant for well-structured entities, yet recent surveys (e.g., [1]) document a clear methodological shift toward data-driven approaches. Especially since the introduction of transformer architectures, such as BERT (Bidirectional Encoder Representations from Transformers) [4], deep learning methods have consistently achieved state-of-the-art (SoTA) performance, with F1-scores surpassing 95% on English benchmarks like i2b2 (Informatics for Integrating Biology and the Bedside) [5] and MIMIC-III (Medical Information Mart for Intensive Care) [1][6]. These corpora and results have effectively set the reference standard for de-identification. While these advances are impressive, they reflect an English-centric research landscape.

The dominance of English corpora means that most methods are trained and validated in data-rich settings, while other languages remain underexplored [7]. For German, the limited availability of annotated corpora presents a critical bottleneck. To our knowledge, as of June 2025, the only publicly available resource suitable for training and evaluation is the Graz Synthetic Clinical text Corpus (GraSCCo), which comprises 64 synthetic examples [8]. This scarcity not only limits model development but also hinders meaningful benchmarking and comparison [9].

Moreover, linguistic and institutional differences reduce the straightforward transferability of English-trained methods [7][9]. High performance on English datasets does not necessarily generalize to German clinical contexts [10]. Consequently, developing effective methods for German de-identification requires both methodological adaptation and careful evaluation under data-scarce conditions.

This study addresses this gap by systematically analyzing SoTA approaches for clinical de-identification and evaluating their transferability to German data. To this end, promising methods are identified in international literature and applied to a small example dataset (GraSCCo). The study thereby explores the opportunities and limitations of modern approaches in a realistic data-limited scenario reflective of the current situation in clinical practice across the German-speaking GSA region (Germany, Austria, Switzerland). The guiding research question is to what extent state-of-the-art methods can be adapted to support de-identification of German clinical free text under small-corpus constraints.

This paper is structured as follows. Section 2 describes the methodology, combining a systematic literature review with a controlled laboratory experiment. Section 3 presents the results of both the literature review and the empirical evaluation on a small German clinical corpus. Section 4 discusses the findings, focusing on methodological reliability, comparability, and cross-lingual transferability. Finally, Section 5 concludes the paper and outlines implications for practice as well as directions for future research.

II. METHODS

This study combines a systematic literature review with a laboratory experiment to assess the transferability of SoTA de-identification methods from English benchmarks to German data-scarce settings.

A. Literature Review

As a starting point, insights into the current research landscape were obtained by analyzing the two major de-identification challenges [5] and [11] and reviewing the recent comparative studies [1] and [2]. Building on these, a systematic literature review of international publications from January 2023 to April 2025 was conducted to capture the latest developments in clinical text de-identification.

The review was guided by three central questions: (1) which methodological approaches have achieved the most reliable performance, (2) what challenges limit comparability across studies, and (3) to what extent can findings from English benchmarks be transferred to the German clinical context?

The review followed Cooper's [12] taxonomy across six dimensions: focus on both methods and outcomes, critical integration of findings, conceptual and methodological organization, neutral perspective, selective yet representative coverage, and orientation toward a scientific audience in natural language processing. Search was conducted primarily in IEEE Xplore and PubMed, supplemented by Google Scholar. Eligible studies focused on the de-identification of clinical free text, reported transparent methods and results, and included performance metrics, such as F1-scores (or, for generative models, accuracy).

In total, 55 publications were reviewed, including more than 30 based on English data as well as the ten top-ranked methods summarized by [1]. For an in-depth analysis, the 25 best-performing studies (measured by F1-score) were compared in detail. Generative language models, which often report accuracy instead of F1-scores, were analyzed separately to account for methodological differences.

To capture the specific developments in the German research landscape, studies on German-language de-identification were treated separately. In this case, all available publications up to May 2025 were included, regardless of their publication year or reported performance. Cooper's taxonomy was therefore applied in a slightly adapted form, complemented by a chronological organization to trace the methodological development of German approaches alongside international trends.

The insights from this review provided the basis for selecting candidate methods for the subsequent laboratory experiment, which evaluates their applicability under data-scarce conditions.

B. Laboratory work

To complement the literature review and evaluate the most promising approaches under realistic data-poor conditions, we conducted a controlled experiment. Controlled experimentation allows for a systematic comparison of model architectures under constant conditions [13].

The independent variable was model architecture; the following systems were compared:

- BiLSTM-CRF (Bidirectional Long Short-term Memory with Conditional Random Field): A bidirectional recurrent neural network with a conditional random field output layer, widely used for sequence labeling in data-limited natural language processing tasks [14][15].
- gELECTRA (German Efficiently Learning an Encoder that Classifies Token Replacements): A German-adapted ELECTRA transformer, fine-tuned for token-level de-identification [16][17].
- LlaMA-3.1-SauerkrautLM-8B-Instruct (Large Language Model Meta AI): An instruction-tuned large language model based on Llama-3.1-8B-Instruct [18], fine-tuned on German-English data [19].

BiLSTM-CRF and gELECTRA were compared under identical conditions (dataset splits, runtime, and sentence-aware chunking into 256-token windows with a 64-token stride, matching gELECTRA's context). LLaMA-3.1 was evaluated separately in one-shot in-context mode on the same test set and metrics. Performance was measured by precision, recall, and F1-score (entity-level, micro- and macro-averaged) on GraSCCo (64 synthetic German clinical notes annotated in accordance with the HIPAA/Safe Harbor identifiers [20]), split 80/20 into training and test sets. BiLSTM-CRF (SpaCy-based [21]) and gELECTRA were trained in Google Colab (T4; 20 epochs; weight_decay 0.01), while LLaMA-3.1-SauerkrautLM was run locally via Ollama [22] to simulate data-sensitive deployment without cloud access. The inclusion of LLaMA-3.1-SauerkrautLM was exploratory, intended to highlight the limitations of large language models (LLM) under extreme data scarcity.

The working hypothesis was that with very small training sets (<100 texts), recurrent architectures, such as BiLSTM yield more stable de-identification performance than large pretrained transformer models or LLM-based one-shot approaches. The experiment thus examines whether compact recurrent or transformer models are more practical for de-identification in GSA-region hospitals, where data and infrastructure are limited.

III. RESULTS

In the following, the results of the literature review and the experimental evaluation on GraSCCo are presented.

A. Literature Review

1) Overall trend

Building on prior comparative reviews [1] and [2] and the shared tasks [5] and [11], our review of publications from January 2023 to April 2025 confirms the continued shift towards transformer-based approaches. Encoder-only transformer families (BERT, RoBERTa, DeBERTa, XLM-R, CamemBERT) dominate recent work (e.g., [37][42][46]). In many publications, the strongest systems appear to be hybrid pipelines that combine transformer encoders with rule-based modules, which reach top F1-scores above 0.98 (e.g.,

[37][42]). In contrast, purely rule-based systems perform substantially worse on modern benchmarks (e.g., MITDeID around F1 ≈ 0.64 [23]).

2) Model performance and state of the art

Specialized transformer variants (e.g., [37][42]) and strong BiLSTM-CRF baselines (e.g., [39][40]) both achieve competitive performance with F1-scores ≥ 0.99, indicating that recurrent architectures can still be viable. The top 25 systems by F1 are summarized in Table I, with hybrid and compact encoder models prevalent among the leaders.

TABLE I. TOP 25 STUDIES BY F1 SCORE (JAN 2023 – MAY 2025)

model	lang	data	cat	eval	F1
PubMedBERT [37]	EN	999 texts, i2b2 2014	21	T, B	0.995
Encoder with self-attention [38]	EN	MIMIC-III	18*	T, B	0.994
LSTM-CRF [39]	EN	MIMIC-III	18*	T, B	0.993
BiLSTM-CRF [40]	EN	600 texts, i2b2 2014	5	T, Mi	0.992
BiLSTM-CRF [41]	EN	MIMIC-III	21**	T, B	0.991
mDeBERTaV3 [42]	ES	MEDDOCAN [43]	8	E, Mi	0.990
BiLSTM-CRF [44]	EN	i2b2 2014	18*	T, B	0.990
Transformer-Ensemble [45]	EN	15'716 texts i2b2 2014	21	E, Mi	0.989 0.970
BERT-Large [46]	EN	i2b2 2014	21**	T, B	0.988
BiLSTM [47]	HU	15000 texts	10	T, B	0.987
Seq2Seq [48]	EN	i2b2 2014	18*	T, B	0.985
BERT-Ensemble [49]	EN	i2b2 2014	18	E, Mi	0.985
BiLSTM-CRF [50]	EN	i2b2 2014	5	T, B	0.983
GRU [51]	EN	i2b2 2014	21**	T, Mi	0.981
LSTM [52]	EN	i2b2 2014	7	T, B	0.980
CamelBERT [53]	AR	i2b2 2014	17	T, B	0.980
PI-RoBERTa [54]	EN	i2b2 2014 (translated)	8	T, Mi	0.980
XLM-RoBERTa large [55]	ES	MEDDOCAN	9	T, Mi	0.976
BERT(med)-BiLSTM-CRF [56]	ZH	33'107 texts	21**	E, Mi	0.976
KoBERT [57]	KO	11'281	6	T, B	0.971
BiLSTM-CRF [58]	FR	878'217	8	T, Mi	0.970
ClinicalBERT [59]	EN	i2b2 2014	18*	E, Mi	0.967
Mistral-7b [27]	FR	9'097 texts	6	T, B	0.967
BiLSTM-CRF [60]	EN	i2b2, CEGS N-GRID	21	E, Mi	0.965
Bert-base-german-cased [30]	DE	i2b2 2014 (translated)	21**	E, Ma	0.960

*18 HIPAA categories; ** i2b2 extended HIPAA categories; lang: language; cat: number of categories; eval: evaluation level (T: Token, E: Entity, B: Binary, Mi: Micro, Ma: Macro); EN: English; ES: Spanish; HU: Hungarian; AR: Arabic; ZH: Chinese; KO: Korean; FR: French; DE: German.

3) Generative LLMs

Decoder-based LLMs are a fast-moving research area, but remain limited in reliability, precision, and clinical usability. Zero-shot named entity recognition is especially challenging. LLaMA-3-8B, for instance, can reach very high recall (≈ 0.99) while yielding extremely low precision (entity-level F1 < 0.20), reflecting hallucinations and prompt sensitivity [24]. GPT-4 can occasionally match the performance of fine-tuned BERT-class models on small datasets, but still struggles with complex identifiers (e.g., hospital names), risking semantic information loss [25]. Prompting and fine-tuning help, yet privacy and compliance concerns constrain the real-world use of cloud-hosted GPT (Generative Pre-Trained Transformer) models [26]. Open-source LLMs like Mistral-7B show promise for local deployment when fine-tuned and quantized, though coverage across Protected Health Information (PHI) categories is often limited [27]. Overall, LLMs yield interesting experimental results, especially in few-shot settings, but remain less robust and transparent than compact encoder models, such as BERT.

4) Datasets, Evaluation and Transferability

Most studies rely on the i2b2-2014 benchmark, often combined with MIMIC-III or institutional hospital data. While this allows a certain degree of comparability, these benchmarks are repeatedly criticized for limited domain coverage and weak generalizability beyond the English-speaking context [28]. Evaluations further differ in whether they are token- or entity-based and in the use of micro- vs. macro-F1, hindering cross-study comparability [1][2]. This heterogeneity also weakens cross-language transfer claims and motivates German-specific validation.

Importantly for German, promising results were reported by Arzideh et al. [29] with gELECTRA and by Gunay et al. [30] using Bert-base-german-cased on machine-translated i2b2 2014 texts, both exceeding F1 > 0.95. These studies demonstrate that compact transformer models can achieve competitive performance even with limited resources, making them particularly relevant for the German-speaking clinical domain, where annotated corpora are small and access to real data remains highly restricted.

B. Research Status for German De-Identification

Research on German-language clinical text de-identification has evolved from early rule-based and hybrid approaches to modern transformer-based and LLM-driven methods. Early work included the Averbis system [31][32], which combined metadata, rule-based tagging, and machine learning and reported near-perfect performance in specific settings. Subsequent studies explored regex-based methods [33] and sequence models, such as CRF and BiLSTM, with BiLSTM reaching up to 96% F-scores [34].

More recent efforts integrated hybrid pipelines, such as Maskeeter [10], combining dictionaries, regex, and manual checks, while Gunay et al. [30] demonstrated strong performance (F1 ≈ 0.96) using German BERT model trained on synthetic and translated corpora.

Arzideh et al. [29] benchmarked multiple transformer models (mBERT, medBERTde, gBERT, gELECTRA, XLM-RoBERTa) on over 10'000 clinical documents, with ensemble strategies and gELECTRA achieving F1-scores up to 0.95, surpassing human annotators. At the same time, Sousa et al.

[35] tested zero- and one-shot prompting with LLMs (GPT-3.5, GPT-4, LLaMA), which showed promising recall but still lagged behind classical architectures in precision and robustness. Finally, Wiest et al. [36] proposed a locally deployable, privacy-preserving pipeline based on quantized LLMs, achieving over 99% sensitivity and 98% specificity while offering end-to-end de-identification workflows, but PHI coverage breadth and reliability remain open issues.

TABLE II. BILSTM-CRF AND GELECTRA PERFORMANCE ON TEST-SET-OBSERVED ENTITY TYPES

Entity	Precision		Recall		F1 Score		S*
Age	1.00	1.00	0.67	1.00	0.80	1.00	3
Date	0.97	0.98	0.95	0.98	0.96	0.98	316
ID	1.00	0.78	0.82	0.78	0.90	0.78	11
City	1.00	0.89	0.92	0.89	0.96	0.89	12
Hospital	1.00	0.57	1.00	0.67	1.00	0.62	19
Organization	1.00	0.00	1.00	0.00	1.00	0.00	2
Street name	0.80	1.00	1.00	1.00	0.89	1.00	12
Postal code	1.00	1.00	1.00	1.00	1.00	1.00	5
Clinician name	1.00	0.67	0.94	0.63	0.97	0.65	68
Patient name	1.00	0.89	0.87	0.94	0.93	0.91	55
Title	1.00	0.78	1.00	0.85	1.00	0.81	72
Profession	1.00	0.00	1.00	0.00	1.00	0.00	1
Micro avg	0.98	0.89	0.94	0.91	0.96	0.90	
Macro avg	0.98	0.66	0.93	0.67	0.95	0.66	

*S: Support

C. Laboratory results on small German corpus

The experimental evaluation on the GraSCCo corpus (see Table II) revealed clear performance differences under data-scarce conditions.

- BiLSTM-CRF: Achieved the best overall entity-level performance with Micro F1 = 0.96 (Micro P = 0.98, Micro R = 0.94; Macro F1 = 0.95). Except for the entities age (F1 = 0.80) and street names (F1 = 0.89), all categories reached Micro F1 scores between 0.90 and 1.00, indicating strong robustness, including on rare entities.
- gELECTRA: Delivered Micro F1 = 0.90 (Micro P = 0.89, Micro R = 0.91), but a substantially lower Macro F1 = 0.66, reflecting weaknesses on infrequent classes. Performance was near-perfect on structured entities (e.g., date: 0.98; postal code: 1.00), yet lower on semantically variable or sparse categories (e.g., clinician name: 0.65, hospital: 0.62) and missed organization (support = 2) and profession (support = 1).
- LlaMA-3.1-SauerkrautLM: The decoder model proved impractical in a one-shot setting. Although the output format was consistent, label assignment was unreliable, with frequent hallucinations, misclassifications, and omissions.

IV. DISCUSSION

On the small synthetic GraSCCo corpus (n = 64) [8][21], BiLSTM-CRF emerges as a reliable baseline (Micro-F1 = 0.96; Macro-F1 = 0.95), consistent with prior evidence for recurrent models in data-scarce sequence tagging [15][16]. gELECTRA performs strongly on structured identifiers, while its Macro-F1 drop (Micro-F1: 0.90; Macro-F1: 0.66) for rare or semantically variable categories might be influenced by data scarcity and label sparsity and should not be taken as definitive evidence of an inherent encoder limitation [1][29][61]. Based on the results in Table I, we expect encoder-based transformers, such as gELECTRA to achieve substantially higher performance and greater stability on a larger, more realistic dataset. Given compliance and infrastructure constraints, BiLSTM-CRF provides a prudent baseline; hybrid systems that combine encoders with deterministic rules merit further exploration [1][2][10].

Recent work exhibits heterogeneous protocols (ranging from token- to entity-level metrics, and micro- vs. macro-averaging) limiting comparability and possibly exaggerating perceived performance gains [1][2]. Our own experiment, though standardized, faces similar constraints due to its small synthetic corpus (GraSCCo, n = 64). In addition, implementation details, such as preprocessing and hyperparameters are often not mentioned in the publications. Together, these factors underline the need for transparent, harmonized benchmarking frameworks to ensure reproducibility and meaningful cross-study comparison.

English shared Tasks and benchmarks catalyzed progress [5][11], yet linguistic/institutional differences and scarce German corpora limit direct transfer [3][7][8][9]. Consequently, German-specific validation is indispensable. Future resources will only be impactful if paired with consistent, openly documented evaluation protocols [1][2] [12][13].

V. CONCLUSION AND FUTURE WORK

Under small-corpus constraints, partial adaptation of English SoTA is feasible. When data are severely lacking, BiLSTM-CRF is currently the most reliable option. gELECTRA remains promising, but our present evaluation is insufficient to judge its full potential.

Implications. For institutions in the GSA region with stringent privacy rules and limited annotation capacity, the most reliable near-term path appears to be a compact encoder (BiLSTM-CRF now and gELECTRA as data grow) complemented by deterministic components for highly structured PHI, deployed fully on-premises [1][2][10][16]. Results are preliminary given the synthetic, small-scale evaluation, and additional data plus careful tuning are likely

to narrow the gap for transformer encoders [1][8][16][17][20].

Looking ahead, further evaluation of gELECTRA on a real annotated corpus, potentially in combination with resources like the currently emerging German Medical Text Corpus (GeMTeX) [62], could provide a solid foundation for developing a privacy-compliant de-identification tool tailored to hospitals in the GSA region.

REFERENCES

[1] A. Kovacevic, B. Basaragin, N. Milosevic, and G. Nenadic "De-identification of clinical free text using natural language processing: a systematic review of current approaches," Artificial Intelligence in Medicine, 2024, ISSN: 0933-3657.

[2] B. Negash et al., "De-identification of free text data containing personal health information: a scoping review of reviews," International Journal of Population Data Science, vol. 8, 2023.

[3] C. Moore, J. Ranisau, W. Nelson, J. Petch, and A. Johnson, "PyCLIPSE: a library for de-identification of free-text clinical notes," arXiv, 2023.

[4] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding" Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, vol. 1, 2019, pp. 4171-4186, doi: https://doi.org/10.18653/v1/N19-1423.

[5] A. Stubbs, C. Kotfila, and Ö. Uzuner, "Automated systems for the de-identification of longitudinal clinical narratives: Overview of 2014 i2b2/UTHealth shared task track 1," Journal of Biomedical Informatics, vol. 58, pp. S11–S19, 2015.

[6] A. E. W. Johnson et al., "MIMIC-III, a freely accessible critical care database," Scientific Data, vol. 3, 2016.

[7] J. L. Leevy, T. M. Khoshgoftaar, and F. Villanustre, "Survey on RNN and CRF models for de-identification of medical free text," Journal of Big Data, vol. 7, no. 73, 2020.

[8] L. Modersohn, S. Schulz, C. Lohr, and U. Hahn, "GRASCCO-The first publicly shareable, multiply-alienated German clinical text corpus," Studies in Health Technology and Informatics, vol. 296, pp. 66–72, 2022.

[9] T. Kolditz et al., "Annotating German Clinical Documents for De-Identification" Studies in Health Technology and Informatics, vol. 264, pp. 203–207, 2019.

[10] M. Baumgartner et al., "Masketeer: An Ensemble-Based Pseudonymization Tool with Entity Recognition for German Unstructured Medical Free Text," Future Internet, vol. 16, 2024.

[11] A. Stubbs, M. Filannino, and Ö. Uzuner, "De-identification of psychiatric intake records: Overview of 2016 CEGS N-GRID shared tasks track 1," Journal of Biomedical Informatics, vol. 75S, pp. S4–S18, 2017.

[12] H. M. Cooper, "Organizing knowledge syntheses: A taxonomy of literature reviews," Knowledge in Society, vol. 1, pp. 104–126, 1988.

[13] C. Wohlin et al., "Experimentation in Software Engineering," Berlin, Germany: Springer, pp. 9–20, 2012.

[14] Z. Huang, W. Xu, and K. Yu, "Bidirectional LSTM-CRF models for sequence tagging," arXiv, 2015.

[15] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer, "Neural architectures for named entity recognition," Proc. NAACL-HLT, San Diego, CA, USA, pp. 260–270, 2016.

[16] K. Clark, M.-T. Luong, Q. V. Le, and C. D. Manning, "ELECTRA: Pre-training text encoders as discriminators rather than generators," ICLR, Addis Ababa, Ethiopia, 2020.

[17] Deepset. *German ELECTRA (gelectra-base).* [Online]. Available from: https://huggingface.co/deepset/gelectra-base [retrieved: January 2026]

[18] Meta AI, *Llama 3.1.* [Online]. Available from: https://ai.meta.com/blog/llama-3-1 [retrieved: January 2026]

[19] VAGOsolutions. *Llama-3.1-8B-Instruct-German-SauerkrautLM*. [Online]. Available from: https://huggingface.co/VAGOsolutions/Llama-3.1-8B-Instruct-German-SauerkrautLM [retrieved: January 2026]

[20] C. Lohr et al., "GraSCCo_PHI - Graz Synthetic Clinical text Corpus with Protected Health Information Annotations," Zenodo, doi: 10.5281/zenodo.11502329.

[21] Explosion AI. *Training Pipelines & Models*. [Online]. Available from: https://spacy.io/usage/training#ner [retrieved: January 2026]

[22] Ollama. *Ollama.* [Online]. Available from: https://www.ollama.com [retrieved: January 2026]

[23] I. Neamatullah et al., "Automated de-identification of free-text medical records," BMC Medical Informatics and Decision Making, vol. 8, p. 32, 2008, doi: 10.1186/1472-6947-8-32.

[24] R. Kuo et al., "Comparative evaluation of large-language models and purpose-built software for medical record de-identification," Research Square, 2024, doi: 10.21203/rs.3.rs-4870585/v1. (*preprint*)

[25] F. J. Moreno-Barea et al, "Named entity recognition for de-identifying Spanish electronic health records," Comput. Biol. Med., vol. 185, p. 109576, 2025, doi: 10.1016/j.compbiomed.2024.109576.

[26] Z. Liu et al., "DeID-GPT: Zero-shot Medical Text De-Identification by GPT-4," arXiv, 2023, doi: 10.48550/arXiv.2303.11032.

[27] O. Dorémus et al., "Harnessing Moderate-Sized Language Models for Reliable Patient Data De-identification in Emergency Department Records: Algorithm Development, Validation, and Implementation Study," JMIR AI, 2025, doi: 10.2196/57828.

[28] J. L. Leevy and T. M. Khoshgoftaar, "A Short Survey of LSTM Models for De-identification of Medical Free Text," Proc. IEEE CIC 2020, Atlanta, GA, USA, pp. 117–124, 2020, doi: 10.1109/CIC50333.2020.00023.

[29] K. Arzideh, et al., "A Transformer-Based Pipeline for German Clinical Document De-Identification," Appl. Clin. Inform., vol. 16, no. 1, pp. 31–43, 2025, doi: 10.1055/a-2424-1989.

[30] M. Gunay, B. Keles, and R. Hizlan, "LLMs-in-the-Loop Part 2: Expert Small AI Models for Anonymization and De-identification of PHI Across Multiple Languages," arXiv, 2024, doi: 10.48550/arXiv.2412.10918.

[31] K. Tomanek, D. Wermter, and U. Hahn, "An Interactive De-Identification-System," Proceedings of the 5th International Symposium on Semantic Mining in Biomedicine (SMBM 2012), Zurich, Switzerland, 2012, doi: 0.5167/UZH-64476.

[32] H. Seuss et al., "Semi-automated De-identification of German Content Sensitive Reports for Big Data Analytics," Rofo, vol. 189, no. 7, pp. 661–671, 2017, doi: 10.1055/s-0043-102939.

[33] P. Richter-Pechanski, S. Riezler, and C. Dieterich, "De-Identification of German Medical Admission Notes," German Medical Data Sciences: A Learning Healthcare System, vol. 253, pp. 165–169, 2018, doi: 10.3233/978-1-61499-896-9-165.

[34] P. Richter-Pechanski, A. Amr, H. A Katus, and C. Dieterich, "Deep Learning Approaches Outperform Conventional Strategies in De-Identification of German Medical Reports," Studies in health technology and informatics, pp. 101-109, 2019, doi: 10.3233/SHTI190813.

[35] S. Sousa, A. M. Jantscher, M. Kröll, and R. Kern, "Large Language Models for Electronic Health Record De-Identification in English and German," Information, vol. 16, no. 2, p. 112, 2025, doi: 10.3390/info16020112.

[36] I. C. Wiest et al., "Deidentifying Medical Documents with Local, Privacy-Preserving Large Language Models: The LLM-Anonymizer," NEJM AI, vol. 2, no. 4, 2025, doi: 10.1056/aidbp2400537.

[37] P. J. Chambon, S. Bluethgen, J. Dreyfuss, A. Lungren, and D. Rubin, "Automated deidentification of radiology reports combining transformer and 'hide in plain sight' rule-based methods," JAMIA, vol. 30, no. 2, pp. 318–328, 2023, doi: 10.1093/jamia/ocac219.

[38] T. Ahmed, M. M. A. Aziz, and N. Mohammed, "De-identification of electronic health record using neural network," Scientific reports, vol. 10, no. 1, p. 18600, 2020, doi: 10.1038/s41598-020-75544-1.

[39] J. Y. Lee, F. Dernoncourt, O. Uzuner, and P. Szolovits, "Feature-Augmented Neural Networks for Patient Note De-identification," ClinicalNLP (COLING 2016), Osaka, Japan, 2016, pp. 17–22. doi: 10.48550/arXiv.1610.09704.

[40] L. Liu, O. Perez-Concha, A. Nguyen, V. Bennett, and L. Jorm, "De-identifying Australian hospital discharge summaries: An end-to-end framework using ensemble of deep learning models," Journal of biomedical informatics, vol. 135, 2022, doi: 10.1016/j.jbi.2022.104215.

[41] F. Dernoncourt, J. Y. Lee, O. Uzuner, and P. Szolovits, "De-identification of Patient Notes with Recurrent Neural Networks," arXiv, 2016, doi: 10.48550/arXiv.1606.03475.

[42] C. Aracena et al., "A Privacy-Preserving Corpus for Occupational Health in Spanish: Evaluation for NER and Classification Tasks," Proceedings of the 6th Clinical Natural Language Processing Workshop, Mexico City, Mexico: ACL, pp. 111–121, 2024, doi: 10.18653/v1/2024.clinicalnlp-1.11.

[43] M. Montserrat et al, "Automatic De-identification of Medical Texts in Spanish: the MEDDOCAN Track, Corpus, Guidelines, Methods and Evaluation of Results," Proceedings of the Iberian Languages Evaluation Forum, 2019.

[44] A. Aloqaily et al., "Deep Learning Framework for Advanced De-Identification of Protected Health Information," Future Internet, vol. 17, no. 1, p. 47, 2025, doi: 10.3390/fi17010047.

[45] K. Murugadoss et al., "Scaling text de-identification using locally augmented ensembles," medRxiv, 2024. (*preprint*)

[46] A. E. W. Johnson, L. Bulgarelli, and T. J. Pollard, "Deidentification of free-text medical records using pre-trained bidirectional transformers," ACM CHIL '20, Toronto, ON, Canada, pp. 214–221, 2020, doi: 10.1145/3368555.3384455.

[47] A. Berzi et al., "NLP-based removal of personally identifiable information from Hungarian electronic health records," Front. Artif. Intell., vol. 8, 2025, doi: 10.3389/frai.2025.1585260.

[48] M. M. Anjum, N. Mohammed, and X. Jiang, "De-identification of Unstructured Clinical Texts from Sequence to Sequence Perspective," Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security, pp. 2438–2440, 2021, doi: 10.1145/3460120.3485354.

[49] K. Murugadoss et al., "Building a best-in-class automated de-identification tool for electronic health records through ensemble learning," Patterns, vol. 2, no. 6, p. 100255, 2021, doi: 10.1016/j.patter.2021.100255.

[50] L. Liu et al., "Web-Based Application Based on Human-in-the-Loop Deep Learning for Deidentifying Free-Text Data in Electronic Medical Records: Development and Usability Study" Interactive journal of medical research, vol. 12, 2023, doi: 10.2196/46322

[51] Y.-S. Zhao, K.-L. Zhang, H.-C. Ma, and K. Li, "Leveraging text skeleton for de-identification of electronic medical records," BMC Med. Inform Decis Mak, vol. 18, suppl. 1, p. 18, 2018, doi: 10.1186/s12911-018-0598-6.

[52] K. Li, Y. Chai, H. Zhao, X. Nan, and Y. Zhao, "Learning to Recognize Protected Health Information in Electronic Health Records with Recurrent Neural Network," Natural Language Understanding and Intelligent Applications, Cham: Springer, pp. 575–582, 2016, doi: 10.1007/978-3-319-50496-4_51.

[53] V. Kocaman, Y. Mellah, H. Haq, and D. Talby, "Automated De-Identification of Arabic Medical Records," Proceedings of ArabicNLP 2023, pp. 33–40, Singapore, 2023, doi: 10.18653/v1/2023.arabicnlp-1.4.

[54] S. Singh et al., "Generation and De-Identification of Indian Clinical Discharge Summaries using LLMs," Proceedings of the 23rd Workshop on Biomedical Natural Language Processing, Bangkok, Thailand, pp. 342–362, 2024, doi: 10.18653/v1/2024.bionlp-1.26.

[55] G. López-García et al., "Named Entity Recognition for De-identifying Real-World Health Records in Spanish," Computational Science – ICCS 2023, Prague, Czech Republic, pp. 228–242, 2023, doi: 10.1007/978-3-031-36024-4_17.

[56] K. Xu, Y. Song, and J. Ma, "Identifying protected health information by transformers-based deep learning approach in Chinese medical text," Health Informatics Journal, vol. 31, no. 1, 2025, doi: 10.1177/14604582251315594.

[57] J. An et al., "De-identification of clinical notes with pseudo-labeling using regular expression rules and pre-trained BERT," BMC medical informatics and decision making, vol. 25, no. 1, 2025, doi: 10.1186/s12911-025-02913-z.

[58] M. E. Azzouzi et al., "Automatic de-identification of French electronic health records: a cost-effective approach exploiting distant supervision and deep learning models," BMC medical informatics and decision making, vol. 24, no. 1, p. 54, 2024, doi: 10.1186/s12911-024-02422-5.

[59] A. Paul, D. Shaji. L. Han, W. Del-Pinto, and G. Nenadic, "DeIDClinic: A Multi-Layered Framework for De-identification of Clinical Free-text Data," arXiv, 2024. doi: 10.48550/arXiv.2410.01648.

[60] S. Meystre and P. Heider, "High Accuracy Open-Source Clinical Data De-Identification: The CliniDeID Solution," Studies in Health Technology and Informatics, vol. 310, pp.1370-1371, 2024, doi: 10.3233/SHTI231199.

[61] A. Ezen-Can, "A Comparison of LSTM and BERT for Small Corpus," arXiv, 2020. doi: 10.48550/arXiv.2009.05451.

[62] Medizininformatik-Initiative. *GeMTeX – Medizinische Texte für die Forschung automatisiert erschließen [GeMTeX – Automated extraction of medical texts for research]*. [Online]. Available from: https://www.medizininformatik-initiative.de/de/gemtex-medizinische-texte-fuer-die-forschung-automatisiert-erschliessen [retrieved: January 2026]

Evaluating Different Explainability Methods for Coronary Artery Segmentation

Apostolos Stogiannis
Democritus University of Thrace
Xanthi, Greece
e-mail: aposstog@ee.duth.gr

Nikos Tsolakis
Information Technologies Institute
Centre for Research & Technology Hellas
School of Informatics
Aristotle University of Thessaloniki
Thessaloniki, Greece
e-mail: ntsolaka@csd.auth.gr

Miriam Gutierrez
Vicomtech, Basque Research and Technology Alliance
Spain
e-mail: mgutierrezf@vicomtech.org

Laura Valeria Pérez
Vicomtech, Basque Research and Technology Alliance
Spain
e-mail: lvperez@vicomtech.org

Karen López-Linares
Vicomtech, Basque Research and Technology Alliance
Spain
e-mail: klopez@vicomtech.org

Christoniki Maga-Nteve
Information Technologies Institute
Centre for Research & Technology Hellas
Thessaloniki, Greece
e-mail: chmaga@iti.gr

Georgios Meditskos
School of Informatics
Aristotle University of Thessaloniki
Thessaloniki, Greece
e-mail: gmeditsk@csd.auth.gr

Stefanos Vrochidis
Information Technologies Institute
Centre for Research & Technology Hellas
Thessaloniki, Greece
e-mail: stefanos@iti.gr

***Abstract*—Deep learning achieves high accuracy in coronary artery segmentation but lacks interpretability for clinical deployment. We present a benchmark of five Explainable Artificial Intelligence (XAI) methods (Gradient-weighted Class Activation Mapping (Grad-CAM), Grad-CAM-plus-plus, Score-CAM, Integrated Gradients, Local Interpretable Model-agnostic Explanations (LIME)) on coronary angiography images from the ARCADE dataset. We introduce a vessel-aware evaluation framework with four metrics, Pointing Game, Average Precision, Intersection over Union, Energy Concentration and systematically optimize layer selection and scoring strategies for each method across three patients. Gradient-based CAM methods achieved an aggregate score of 0.420, with consistent layer preferences across patients, while perturbation methods fail. Our findings establish Grad-CAM as optimal for coronary vessel explanation and demonstrate that layer optimization is method-specific but patient-invariant.**

Keywords*-Explainable Artificial Intelligence (XAI); Coronary Artery Segmentation; Medical Imaging; Healthcare.

I. INTRODUCTION

Cardiovascular Diseases (CVDs) remain the leading cause of mortality worldwide [1]. A critical aspect in both diagnosis and treatment planning is the analysis of coronary arteries, where stenosis, occlusions, or anatomical abnormalities can have life-threatening consequences. X-ray Coronary Angiography (XCA) is the clinical gold standard for visualizing the coronary vasculature [2], yet manual interpretation is labor-intensive, subjective, and prone to inter-observer variability [3]. Automated vessel segmentation can accelerate clinical workflows and provide consistent anatomical delineation, directly supporting decision making in coronary interventions and surgical planning.

Deep learning has achieved state-of-the-art performance in medical image segmentation. Architectures such as U-Net [4] and its variants effectively capture fine vessel structures, enabling automated extraction of coronary artery trees from angiographic images [5][6][7]. However, despite these advances, most models are often treated as black boxes, offering limited insight into how predictions are made [8]. This opacity raises concerns in safety-critical clinical applications, where transparency and trustworthiness are as important as accuracy. In coronary imaging, a lack of explainability can hinder clinical adoption, as physicians must understand and validate the rationale behind automated vessel delineations before integrating them into diagnostic or interventional decisions. Misinterpretation of coronary anatomy, such as inaccurate boundary detection or missing

lesions, could directly affect treatment planning, stent placement, or surgical strategy [9].

Explainable Artificial Intelligence (XAI) aims to bridge this gap by enhancing the interpretability of deep learning models. In coronary artery segmentation, XAI can highlight which image regions most influenced the model's output, providing a visual rationale for automated predictions. While several XAI methods have been proposed, such as gradient-based (Gradient-weighted Class Activation Mapping (Grad-CAM) [10], Grad-CAM++ [11]), perturbation-based (Score-CAM [12]), and surrogate-model approaches like Local Interpretable Model-agnostic Explanations (LIME) [14], their effectiveness for thin, branching, and sparse vascular structures has not been systematically evaluated.

In this work, we address this gap by implementing and benchmarking multiple XAI methods for coronary artery segmentation on the publicly available ARCADE dataset [6] using a U-Net model. Our goal is to provide a reproducible framework for evaluating explanation methods tailored to vascular imaging. We apply a set of vessel-aware quantitative metrics: Pointing Game [15], Average Precision [16], Intersection over Union [17], and Energy Concentration Ratio [18], that complement visual inspection and enable rigorous comparison. By analyzing results across multiple patients, we highlight which methods are most informative and practical for clinical use, ultimately aiming to improve the trustworthiness and adoption of AI in cardiology.

The remainder of this paper is organized as follows. Section II reviews related work on XAI in medical imaging and coronary artery analysis. Section III describes the proposed methodology, including the dataset, model architecture, explainability methods, and evaluation metrics. Section IV presents the explanation evaluation framework. Section V reports the experimental results and quantitative comparisons. Section VI discusses the obtained findings and their implications. Finally, Section VII concludes the paper and outlines directions for future work.

II. Related work

In recent years, considerable research has focused on integrating XAI into medical imaging to enhance interpretability and clinical trust. These approaches aim not only to make deep learning models transparent but also to create conditions that support their safe deployment in clinical workflows.

Do et al. [19] proposed a deep learning framework for diagnosing Coronary Artery Disease (CAD) using Single-Photon Emission Computed Tomography (SPECT) Myocardial Perfusion Imaging (MPI) polar maps. Their ResNet152V2 model incorporated LIME, Grad-CAM, and RISE for interpretability, evaluated via deletion and insertion metrics to overcome limitations of qualitative heatmap inspection. Similarly, Papandrianos et al. [20] developed an RGB-Convolutional Neural Network (CNN) model for automatic CAD classification from SPECT MPI images, achieving 93.3 % accuracy and 94.6 % AUC, and used a Grad-CAM-based color visualization to explain model decisions. Beyond imaging, Goettling et al. [21] introduced xECGArch, an interpretable deep learning architecture for ECG analysis that combines short- and long-term CNN branches. They compared 13 XAI methods using perturbation analysis, demonstrating the value of systematic XAI evaluation for physiological signal interpretation.

Bhandari et al. [22] applied Grad-CAM, LIME, and SHAP to classify chest X-ray images into COVID-19, pneumonia, and tuberculosis, showing that combining complementary XAI methods can improve visual interpretability. Their work highlighted the importance of multi-method comparison for clinical validation of deep learning models.

Anand et al. [23] employed a U-Net architecture for coronary vessel segmentation in X-ray angiography, achieving strong quantitative results (mean F1 = 0.921) but without integrating explainability. Gao et al. [24] later proposed an ensemble framework combining deep learning and filter-based features for coronary artery segmentation, reporting high precision and sensitivity across 130 angiographic images. These works demonstrate the maturity of segmentation pipelines but also the lack of explainability integration in vessel-specific contexts.

Bhati et al. [9] provided a broad survey of XAI techniques in medical imaging, categorizing gradient-based, perturbation-based, decomposition, and attention-driven methods, and discussing challenges to clinical adoption and validation.

Collectively, these studies show that while XAI is increasingly applied to medical imaging, quantitative evaluation of explanations, particularly for segmentation tasks involving thin, branching structures like coronary arteries, remains underexplored. To address this gap, our study focuses on systematically benchmarking multiple canonical XAI methods (Grad-CAM, Grad-CAM++, Score-CAM, Integrated Gradients [13], LIME) on coronary artery segmentation, introducing vessel-aware evaluation metrics (pointing game, average precision, Intersection Over Union (IoU), energy concentration ratio) and ensuring reproducibility through open, well-specified implementation details.

III. Methodology

This section describes the overall methodology adopted in this study, including the dataset selection, the preprocessing steps, the segmentation model architecture, the evaluation metrics, and the explainability methods used to interpret and evaluate the model predictions.

A. Dataset Acquisition and Preprocessing

This study employs data from the ARCADE dataset which contains 1,200 X-ray Coronary Angiography (XCA)

images with pixel-level vessel annotations. The dataset is divided into 1,000 training and 200 validation images. Each image is accompanied by a binary mask labeling vessel vs. background. While ARCADE also provides region-specific annotations for 26 SYNTAX anatomical regions, this study focuses on binary segmentation to isolate the vessel tree as a whole. The dataset was converted into TFRecord format to enable efficient training and inference.

Besides ARCADE there are other public datasets for coronary artery analysis, such as CADICA [25], XCAD [26]. CADICA provides annotated invasive coronary angiography videos; however, it is limited by a relatively small cohort size. XCAD includes segmentation masks for coronary arteries but offers fewer annotated samples and a restricted testing set. In contrast, the ARCADE dataset provides a large-scale, standardized benchmark with high-quality pixel-level vessel annotations and a diverse clinical population. Its size, annotation quality, and task diversity make ARCADE particularly suitable for robust coronary artery segmentation and explainable artificial intelligence evaluation. Therefore, ARCADE was selected as the primary dataset for this study.

To enhance vessel visibility and standardize inputs, several preprocessing techniques were applied. First, a white top-hat filter was used to enhance the contrast of bright vessels against the dark background. Through normalization, the intensity values were scaled from [0-255] to [0-1]. Lastly, Contrast Limited Adaptive Histogram Equalization (CLAHE) was applied to further enhance contrast. Collectively, these preprocessing steps enhance the visibility of thin vessel structures that are otherwise difficult to distinguish from background noise in XCA.

B. Model Architecture and training configuration

The segmentation architecture is based on the U-Net model, selected for its proven versatility and strong performance in medical image segmentation. Several segmentation architectures were evaluated during preliminary experiments, with U-Net providing the best trade-off between performance and interpretability. The network comprises five encoder and five decoder stages. Each encoder block applies two 3×3 convolutions with ReLU activation and HeNormal kernel initialization, followed by batch normalization, dropout, and max pooling for down sampling. Each decoder block performs 2× upsampling via a 3×3 transposed convolution (stride 2), concatenates the corresponding encoder skip connection, and applies two 3×3 convolutions. The final output layer uses a 1×1 convolution followed by a softmax activation to generate per-class probability maps. During training, a Centerline Cross-Entropy (clCE) loss was employed to emphasize accurate segmentation of thin and elongated coronary vessels [27]. This loss increases penalties for misclassification near vessel centerlines, helping preserve connectivity and reduce fragmentation in fine vascular branches, which is critical for clinical interpretability.

C. Model Performance Evaluation

To evaluate model performance, four metrics were selected, Precision, Recall, Dice Coefficient and IoU. Precision is the number of true positive results divided by the number of all positive results. It quantifies to avoid false detections.

$$\text{Precision} = TP/ (TP + FP) \quad (1)$$

TP, FP, and FN denote True Positives, False Positives, and False Negatives, respectively.

Recall is the number of true positive results divided by the number of all cases that should have been identified as positive; in this case it measures how many true vessel pixels are successfully identified, reflecting the ability to capture thin and peripheral branches.

$$\text{Recall} = TP / (TP + FN) \quad (2)$$

Dice Coefficient quantifies the overlap between the predicted and ground truth masks by balancing precision and recall. It is widely used in medical segmentation benchmarks.

$$\text{Dice} = 2 * TP / (2 * TP + FP + FN) \quad (3)$$

The IoU metric measures the ratio of the intersection area to the union area between prediction and ground truth. It is a stricter metric that penalizes both false positives and false negatives, providing a more conservative measure of segmentation quality.

$$\text{IoU} = TP / (TP + FP + FN) \quad (4)$$

D. Explainability Methods

To investigate how different explainability techniques perform in the context of coronary artery segmentation, five widely used XAI methods were evaluated. These methods represent different methodological families: gradient-based, perturbation-based, path integrated, and surrogate modeling approaches. Each method was selected to provide a diverse view of explainability approaches, as their underlying assumptions and outputs differ significantly.

Grad-CAM is one of the most established methods for visualizing convolutional neural networks. It computes gradients of the target class with respect to the feature maps of the selected convolutional layer to produce a coarse localization heatmap of discriminative regions. Grad-CAM was included as a baseline for explainability in medical imaging tasks.

Grad-CAM++ extends Grad-CAM by incorporating higher order derivatives into the gradient computation, enabling better capture of multiple relevant regions and small-scale structures. Given the thin and branching morphology of coronary vessels, Grad-CAM++ was

included to determine whether its improved sensitivity could lead to more precise vessel localization compared with standard Grad-CAM.

Score-CAM generates heatmaps by weighting activation maps according to the model's forward prediction scores, thereby eliminating the dependency on gradients. For each feature map, Score-CAM upsamples the activation to input resolution, masks the original image with this map, and re-evaluates the model. The importance of the feature map is then derived from how much the masked input changes the model's confidence relative to the baseline prediction. In our implementation, four scoring strategies were assessed based on the model outputs:

Increase: Measures the confidence gain compared to the baseline prediction, emphasizing features that increase the likelihood of vessel detection.

Absolute: Uses the raw predicted probability for the vessel class, directly weighting features by their contribution to class confidence.

Entropy: Computes the negative entropy of the softmax distribution, rewarding feature maps that produce more confident predictions.

Max-logit: Weighs maps according to the difference between the vessel and background logits, highlighting features that maximize class separation.

By testing multiple scoring functions, we aimed to determine whether vessel-specific signal characteristics could be better captured through alternative weighting schemes. This method has the potential to generate smoother and less noisy explanations for vessel structures, compared to the gradient-based methods analyzed previously.

Integrated Gradients (IG) explain model predictions by integrating gradients along a continuous path from a baseline input to the actual input. To reduce sensitivity to baseline selection, we averaged results over three baselines: a black image, Gaussian noise, and a mean-intensity image. IG provides theoretically grounded attributions that satisfy sensitivity and implementation invariance, and has been widely used in medical imaging. We applied IG to test how path-integrated attributions perform in coronary angiography, where vessel structures are thin and globally sparse. However, due to their gradient-based nature, IG explanations may emphasize vessel edges rather than the vessel interior, which is an important limitation explored in this study.

LIME approximates a model's decision boundary locally by fitting an interpretable surrogate model (e.g., linear regression) to perturbed versions of the input. We implemented LIME with Felzenszwalb superpixel segmentation, which groups neighboring pixels into locally homogeneous regions. This choice was motivated by the need to preserve vessel-like structures while reducing the dimensionality of the perturbation space. LIME was included because it is one of the most widely recognized model agnostic methods and is often used as a baseline in XAI research. In medical imaging, it has been applied to modalities such as chest radiographs and skin lesion images. In the context of coronary angiography, however, the superpixel-based approach introduces challenges: vessels are thin, elongated, and sparse, and a single superpixel may encompass both vessel and background regions. This can reduce the fidelity of the surrogate model and lead to fragmented or misleading explanations, which we explicitly evaluate in our experiments.

E. Explanations Performance Evaluation

To quantitatively evaluate the explanations produced by each XAI method, we implemented a set of vessel-aware metrics tailored for image segmentation. These metrics capture complementary aspects of explanation quality: localization accuracy, ranking ability under class imbalance, overlap with ground truth vessels, and concentration of attribution energy. All heatmaps were normalized to the [0, 1] range prior to evaluation, ensuring consistency across methods.

Pointing Game (PG) measures localization accuracy by testing whether the regions of highest attention overlap with true vessel structures. For each heatmap, we normalize intensities and select pixels above the 80th percentile as high-attention regions. These regions are compared against a dilated ground truth mask, which tolerates small misalignments in thin vessels. The PG score is the ratio of high-attention pixels inside vessels to the total number of high-attention pixels. A higher score indicates better vessel localization.

Average Precision (AP) evaluates the ranking ability of heatmap values under class imbalance, since vessels occupy only a small fraction of the image. We flatten the normalized heatmap into prediction scores and the binary vessel mask into labels, and compute average precision across recall levels:

$$AP = \Sigma_n (R_n - R_{n-1}) P_n \quad (5)$$

R_n and P_n are recall and precision at the n-th threshold. If the ground truth mask is empty, AP is set to 0.0. A higher AP score reflects the ability of an explanation to consistently rank vessel pixels above background pixels.

IoU quantifies the overlap between binarized attention maps and vessel structures. Each heatmap is thresholded at the 80th percentile to create a binary mask of salient regions. We then compute:

$$IoU = TP / (TP + FP + FN) \quad (6)$$

IoU penalizes both false positives and false negatives, providing a strict measure of spatial overlap.

Energy Concentration Ratio (ECR) measures how much of the explanation energy is concentrated within vessels compared to background regions. It is defined as:

$$ECR = \Sigma (H * M) / \Sigma H \quad (7)$$

where H is the normalized heatmap and M is the binary vessel mask. A higher ECR indicates that the majority of attribution energy is focused on true vessel structures rather than distributed across irrelevant regions.

Together, these four metrics provide a comprehensive framework for evaluating XAI methods. PG emphasizes localization of the most salient activations, AP tests ranking performance under imbalance, IoU measures strict spatial overlap, and ECR captures global energy distribution.

IV. EXPLANATIONS EVALUATION FRAMEWORK

The proposed explainability framework compares each method's heatmap with the ground truth vessel mask for the selected patients. For methods that require a layer choice (Grad-CAM, Grad-CAM++, Score-CAM), we systematically tested multiple candidate layers and retained the configuration that achieved the best metric performance. For Score-CAM, all four scoring strategies were applied on the tested layers, and the best-performing variant was reported. For IG, the attributions were accumulated over 100 steps, and only positive contributions were retained for the final explanation. For LIME, we implemented a vessel-focused setup. Images were segmented into superpixels using the Felzenszwalb algorithm (scale = 50, sigma = 0.3, min_size = 20). A fixed random seed was used to ensure reproducibility. The prediction function was adapted to coronary vessels by combining per-pixel probabilities with a connectivity-based adjustment, favoring explanations that highlight continuous vascular structures. LIME explanations were generated with 1500 perturbation samples and up to 150 interpretable features. To enable overall ranking of XAI methods, we computed an aggregate score defined as the unweighted average of the four evaluation metrics (PG, AP, IoU, ECR) for simplicity, clinical deployment could benefit from task-specific weighting (e.g., higher weight on PG for localization-critical applications). This evaluation framework ensures that each XAI method is tested fairly under well-specified conditions, and that reproducibility is guaranteed by reporting implementation details such as baseline choices, scoring functions, segmentation parameters, and random seeds.

V. RESULTS

This section presents the evaluation results in two parts: first, the segmentation performance of the U-Net model under different preprocessing configurations; and second, a quantitative and qualitative comparison of five XAI methods applied to three patient cases.

A. Model Performance Results

We evaluated the U-Net model under four conditions: baseline (no modifications), post processing only, filtering only, and post processing combined with filtering. The quantitative results are summarized in Table I:

TABLE I. MODEL PERFORMANCE RESULTS FROM EACH TEST

Metrics	TEST 1	TEST 2	TEST 3	TEST 4
Dice	0.653	0.653	0.677	0.679
IoU	0.484	0.485	0.512	0.514
Precision	0.712	0.712	0.720	0.721
Recall	0.603	0.604	0.640	0.642

Post processing consisted of morphological closing (radius=3) followed by removal of small connected components (<100 px). This step produced only marginal gains (+0.001 IoU, +0.001 Recall), suggesting that U-Net predictions were already smooth and contained limited spurious noise.

Filtering, implemented via White Top-Hat vessel enhancement and CLAHE, had a stronger effect. Dice increased from 0.653 to 0.677 (+3.7%), IoU from 0.484 to 0.512 (+5.8%), and recall from 0.603 to 0.640 (+6.1%), while precision remained essentially stable. These improvements indicate that filtering improved vessel contrast, enabling the network to capture more thin and peripheral branches without introducing false positives.

The combination of filtering and post processing yielded the best overall performance (Dice=0.679, IoU=0.514), but only marginally above filtering alone. This suggests that filtering is the dominant factor driving improvements, while post processing provides small refinements.

B. Explanations Performance Results

We evaluated explanation performance on patients 15, 148 and 237. Results are reported per metric (PG, AP, IoU, ECR) and summarized into a total score. We also present heatmaps for qualitative comparison.

As explained in section IV we begin by testing Grad-CAM and Grad-CAM++ on the selected layers to find which one performs the best metrics wise. Across all three patients the best performing layer was "activation" for these methods. Then we tested Score-CAM, this time the best performing layer for all patients was "conv2d_15" and the best scoring method was "entropy". Lastly, we tested IG and LIME with the implementation that was explained in section IV. In Tables II, III, and IV, we can see how each explanation performed, across all metrics for each patient.

TABLE II. EXPLANATIONS PERFORMANCE RESULTS ACROSS EACH METRIC FOR PATIENT 15

Method	PG	AP	IoU	ECR	Score
Grad-CAM	0.286	0.580	0.188	0.465	0.380
Grad-CAM++	0.286	0.580	0.188	0.465	0.380
Score-CAM	0.200	0.416	0.143	0.283	0.260
IG	0.145	0.059	0.067	0.068	0.085
LIME	0.064	0.061	0.040	0.186	0.088

TABLE III. EXPLANATIONS PERFORMANCE RESULTS ACROSS EACH METRIC FOR PATIENT 148

Method	PG	AP	IoU	ECR	Score
Grad-CAM	0.223	0.714	0.129	0.491	0.389
Grad-CAM++	0.223	0.714	0.129	0.491	0.389
Score-CAM	0.191	0.572	0.116	0.231	0.278
IG	0.112	0.038	0.038	0.056	0.061
LIME	0.048	0.054	0.027	0.165	0.074

TABLE IV. EXPLANATIONS PERFORMANCE RESULTS ACROSS EACH METRIC FOR PATIENT 237

Method	PG	AP	IoU	ECR	Score
Grad-CAM	0.268	0.691	0.179	0.541	0.420
Grad-CAM++	0.268	0.691	0.179	0.541	0.420
Score-CAM	0.229	0.597	0.160	0.299	0.321
IG	0.128	0.062	0.057	0.097	0.086
LIME	0.055	0.083	0.036	0.268	0.111

Overall, Grad-CAM and Grad-CAM++ performed the best across each metric for all three patients, while having the same values. Score-CAM achieved middling results, IG and LIME performed the worst.

Now, for each patient, we present the heatmap from each explanation method. For added context we present the original X-ray image for each patient from the ARCADE dataset and the Ground Truth (GT) mask that was used to evaluate the explanations.

We present images for patients 15, 148 and 237 in Figures 1, 2, and 3:

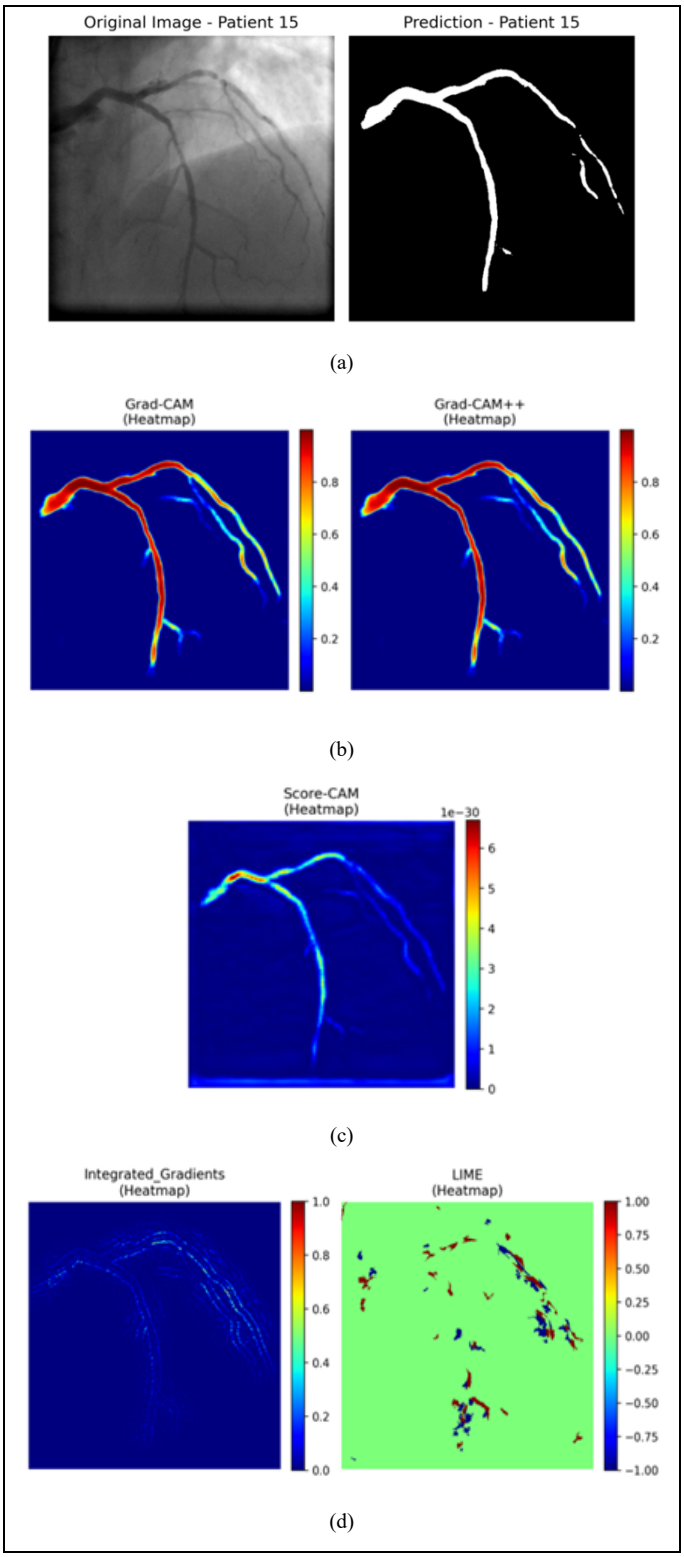


Figure 1. Comparison of heatmaps for patient 15: (a) Original image and GT mask, (b) Grad-CAM and Grad-CAM++, (c) Score-CAM, (d) IG and LIME.

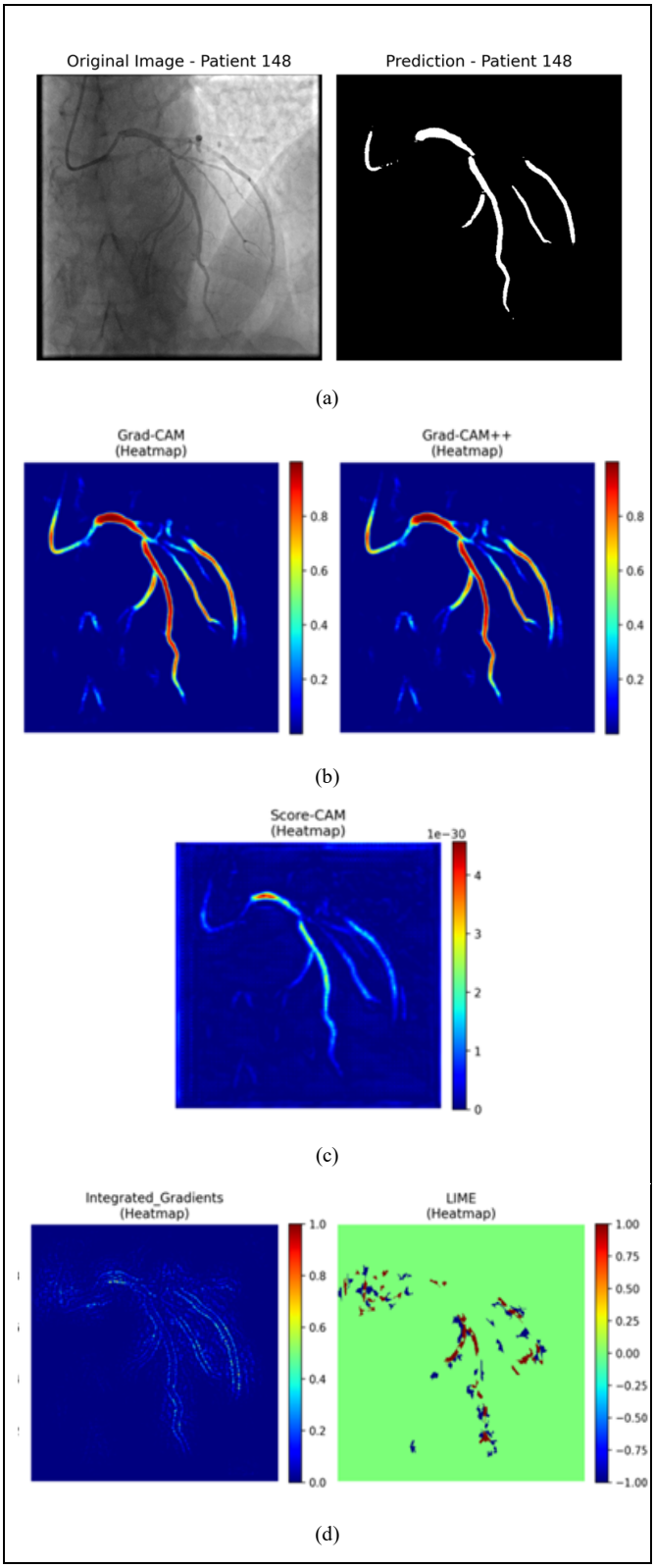


Figure 2. Comparison of heatmaps for patient 148: (a) Original image and GT mask, (b) Grad-CAM and Grad-CAM++, (c) Score-CAM, (d) IG and LIME.

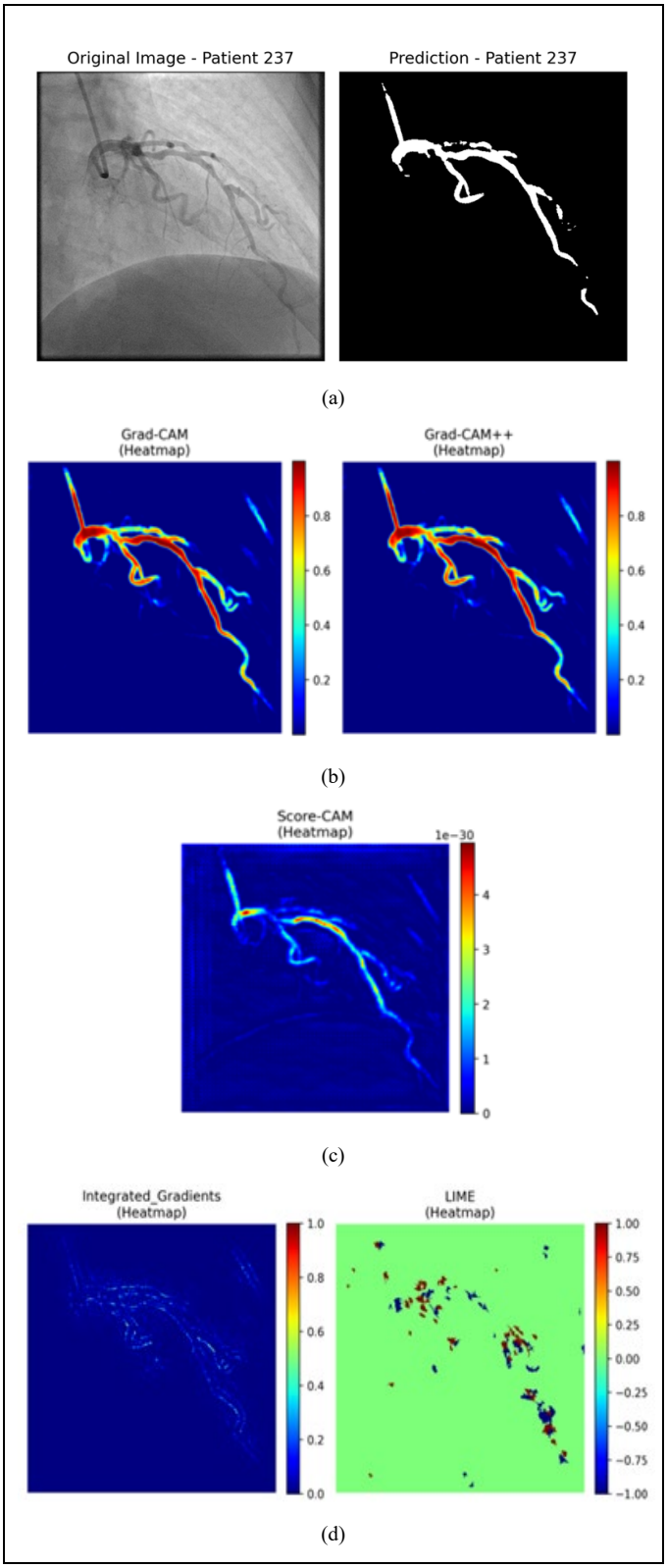


Figure 3. Comparison of heatmaps for patient 237: (a) Original image and GT mask, (b) Grad-CAM and Grad-CAM++, (c) Score-CAM, (d) IG and LIME.

 ISBN: 978-1-68558-362-0

We observe that each ground truth mask is able to depict major coronary arteries with good continuity, making most of the vessel tree structure visible. However, some terminal branches don't appear in them and there seem to be some discontinuities in some parts of each vessel structure.

The heatmaps generated by Grad-CAM and Grad-CAM++ are able to highlight the main coronary arteries with high intensity values (0.8–1.0) and assigning lower activations (0.4–0.6) to peripheral branches. For each patient these two methods generate almost identical heatmaps. Score-CAM is able to some extent, portray the vessel tree, but from the range of values on each heatmap, it has problems giving more attention to bigger vessels than the smaller branches. It also has a lot of background noise. IG primarily emphasizes vessel edges, while LIME highlights isolated regions rather than reconstructing the full vascular structure.

VI. DISCUSSION

The results indicate that Grad-CAM and Grad-CAM++ produced the most clinically meaningful explanations, achieving the highest aggregate scores and generating continuous activations along coronary centerlines. Grad-CAM++ did not outperform Grad-CAM, suggesting that higher-order derivatives add limited value for thin vessel structures. Score-CAM localized vessels moderately well, but struggled to distinguish main vessels from smaller branches. Its superior performance of the entropy-based scoring suggests that uncertainty reduction is a better proxy for vessel importance. The optimal layer selection proved consistent across all three patients, suggesting that layer preferences are method-specific rather than image-specific, which simplifies clinical deployment. Moreover, Grad-CAM benefits from final high-level features, while Score-CAM's perturbation approach works better at intermediate feature levels where spatial resolution is higher. IG and LIME failed to capture vessel interiors or continuity, instead producing edge- or spot-like activations, reflected in very low composite scores.

The four evaluation metrics offered complementary insights. PG assessed the localization of peak activations, AP measured ranking under class imbalance, IoU captured strict overlap after binarization, and ECR quantified the distribution of energy attribution. Using both continuous and thresholded metrics revealed distinct failure modes, such as diffuse attention yielding high ECR but low IoU. These findings highlight the need for multi-metric evaluation in medical XAI.

Finally, qualitative inspection showed that CAM-based methods consistently highlighted large vessel trunks, while smaller branches were less emphasized. In some cases, CAM heatmaps appeared to reveal details not present in the ground truth, suggesting potential annotation limitations.

VII. CONCLUSIONS AND FUTURE WORK

We presented a systematic framework for evaluating explainability methods in coronary artery segmentation using the ARCADE dataset and a U-Net backbone. Our findings show that CAM-based methods, particularly Grad-CAM and Grad-CAM++, provide the most reliable explanations, while perturbation and gradient-integration approaches, such as Score-CAM, IG, and LIME were less effective for thin vascular structures.

Future work will expand this evaluation state-of-the-art architectures such as TransUNet or nnU-Net, larger and more diverse datasets, incorporate additional XAI methods (e.g., SHAP, concept-based explanations), and include expert evaluation by cardiologists to assess clinical usability while they also assess whether CAM-based explanations identify genuine anatomical structures missed during manual annotation. Integrating explainability into real-time clinical workflows and extending the framework to multi-class vessel segmentation are also promising directions.

REFERENCES

[1] G. A. Roth *et al.*, "Global burden of cardiovascular diseases and risk factors, 1990–2019: Update from the GBD 2019 study," *J. Am. Coll. Cardiol.*, vol. 76, no. 25, pp. 2982–3021, 2020.

[2] M. Kolossváry, B. Szilveszter, P. Kolossváry, I. Karády, and P. Maurovich-Horvat, "Radiomic features are superior to conventional quantitative computed tomographic metrics to identify coronary plaques with napkin-ring sign," *Circ. Cardiovasc. Imaging*, vol. 10, no. 12, pp. 1–10, 2019.

[3] B. K. Nallamothu, M. H. Spertus, D. A. Lansky, J. D. Hofer, and E. R. Bates, "Comparison of clinical interpretation with visual assessment and quantitative coronary angiography in patients undergoing percutaneous coronary intervention in contemporary practice," *Circulation*, vol. 127, no. 17, pp. 1793–1800, 2013.

[4] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: convolutional networks for biomedical image segmentation," in Proc. Int. Conf. Medical Image Computing and Computer-Assisted Intervention (MICCAI), 2015, pp. 234–241.

[5] T. T. D. Mahendiran, P. Rajan, S. P. Singh, and A. Verma, "AngioPy segmentation: Deep learning tool for coronary segmentation," *Int. J. Cardiol., in press, 2025.*

[6] M. A. Popov, V. A. Soldatov, and I. A. Solovyev, "Dataset for automatic region-based CAD diagnostics using X-ray angiography images," *Sci. Data*, vol. 20, 2024.

[7] C. Zhao, C. Yang, J. Gao, and S. Xia, "AGMN: Graph matching network for coronary artery semantic labeling," *Pattern Recognit.*, vol. 143, 2023.

[8] C. Rudin, "Stop explaining black-box machine learning models for high-stakes decisions and use interpretable models instead," *Nat. Mach. Intell.*, vol. 1, no. 5, pp. 206–215, 2019.

[9] D. Bhati, F. Neha, and M. Amiruzzaman, "A survey on explainable artificial intelligence (XAI) techniques for visualizing deep learning models in medical imaging," *J. Imaging*, vol. 10, 2024.

[10] R. R. Selvaraju, A. Das, R. Vedantam, M. Cogswell, and D. Parikh, "Grad-CAM: visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2017, pp. 618–626, doi: 10.1109/ICCV.2017.74.

[11] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-CAM++: generalized gradient-based visual explanations for deep convolutional networks," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, 2018, pp. 839–847, doi: 10.1109/WACV.2018.00097.
[12] H. Wang, Z. Wang, H. Du, Y. Shen, and Y. Pu, "Score-CAM: Score-weighted visual explanations for convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, 2020, pp. 111–119, doi: 10.1109/CVPRW50498.2020.00020.
[13] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2017.
[14] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?' Explaining the predictions of any classifier," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, 2016, pp. 1135–1144.
[15] J. Zhang and K. A. Chan, "Top-down neural attention by excitation backprop," *Int. J. Comput. Vis.*, vol. 126, no. 10, pp. 1084–1102, 2018.
[16] J. Davis and M. Goadrich, "The relationship between precision–recall and ROC curves," in *Proc. 23rd Int. Conf. Mach. Learn.*, 2006, pp. 233–240.
[17] G. Csurka, D. Larlus, F. Perronnin, and F. Meylan, "What is a good evaluation measure for semantic segmentation?" in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, 2013, pp. 32.1–32.11.
[18] V. Petsiuk, A. Das, and K. Saenko, "RISE: Randomized input sampling for explanation of black-box models," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, 2018.
[19] T. Do, P. Huynh, M. Nguyen, and V. Nguyen, "An XAI-based deep learning framework for coronary artery disease diagnosis using SPECT MPI polar map images," in *Proc. Seventh Int. Conf. Res. Intell. Comput. Eng.*, 2022, pp. 235–241, doi: 10.15439/2022R06.
[20] N. I. Papandrianos, I. A. Panagiotopoulos, G. T. Papageorgiou, and K. N. Sidiropoulos, "An explainable classification method of SPECT myocardial perfusion images in nuclear cardiology using deep learning and Grad-CAM," *Appl. Sci.*, vol. 12, p. 7592, 2022.
[21] M. Goettling, J. M. Schwenk, F. Kragness, J. Tomaszewski, and R. F. Speier, "xECGArch: A trustworthy deep learning architecture for interpretable ECG analysis considering short-term and long-term features," *Sci. Rep.*, vol. 14, no. 13122, 2024.
[22] M. Bhandari, R. Singh, S. Gupta, A. Kumar, and R. Sharma, "Explanatory classification of CXR images into COVID-19, pneumonia, and tuberculosis using deep learning and XAI," *Comput. Biol. Med.*, vol. 150, p. 106156, 2022.
[23] H. S. Anand, R. K. Sharma, and P. Gupta, "Coronary vessel segmentation in X-ray using U-Net," in *Lecture Notes in Networks and Systems*, vol. 969, Springer, 2024, pp. 57–66.
[24] Z. Gao, Z. Li, H. Ma, and L. Zhang, "Vessel segmentation for X-ray coronary angiography using ensemble methods with deep learning and filter-based features," *BMC Med. Imaging*, vol. 22, p. 10, 2022.
[25] J. Jiménez-Carretero, R. Ruiz-Sarmiento, J. M. Górriz, and J. Ramírez, "CADICA: A dataset for coronary artery disease analysis in invasive coronary angiography," arXiv preprint arXiv:2402.00570, 2024.
[26] Z. Wang, Y. Guo, X. Yang, and J. Liu, "XCAD: A dataset for coronary artery segmentation in X-ray angiography,"
IEEE Access, vol. 8, pp. 189030–189041, 2020.
[27] C. Acebes, C. Tejos, and P. Irarrazaval, "The centerline–cross entropy loss for vessel-like structure segmentation," in *Proc. MICCAI*, 2024, doi: 10.1007/978-3-031-72117-5_30.

Assessing the Effectiveness of an Artificial Intelligence Tool for Note-taking in a General Practice Setting

An evaluation of the Heidi AI note-taking tool

Shreya Shah, Aarya Shetye, Carol Habib
Guy's, King's and St Thomas' School of Medical Education
King's College London
London, United Kingdom
e-mail: shreya.2.shah@kcl.ac.uk, aarya.shetye@kcl.ac.uk, carol.habib@kcl.ac.uk

Abstract— **Documentation burden in general practice reduces efficiency and productivity. This is particularly important as clinicians face increasing time pressures, and inaccurate documentation can compromise patient safety and continuity of care. Large language models may assist with automated note-generation, but evidence of practical performance remains very limited. Existing studies are limited in number and often rely on simulated consultations or non-clinical settings, leaving uncertainty regarding real-world performance in primary care. This study assessed the accuracy of Heidi, an artificial intelligence medical scribe software, compared to manual note-taking by medical students in a general practice setting. Across six consultations, Heidi achieved a mean accuracy of 84%, with most errors being minor or clinically non-significant, suggesting that AI-assisted documentation may safely support clinical workflows when combined with appropriate human oversight.**

Keywords-Artificial Intelligence; General Practice; Documentation; Efficacy; Heidi; Large Language Models; Note-taking; Primary Care; Digital Health.

I. INTRODUCTION

Effective documentation is a cornerstone of healthcare, requiring both accuracy and efficiency. It is crucial for patient safety, ongoing care, and adherence to legal standards. Generally, general practitioners rely on manual note-taking either during or after a consultation. However, this can be time-intensive and may interrupt the natural flow of doctor–patient communication. Recently, artificial intelligence has become more prominent in healthcare. It is being used to support clinicians in optimising documentation without compromising the quality of patient communication. Among these tools is Heidi, a clinical documentation system powered by artificial intelligence, which is designed to improve the accuracy, consistency, and speed of consultations.

Started in 2019 [1], Heidi is a medical scribe that uses Artificial Intelligence (AI) and Natural Language Processing (NLP) to convert real-time or recorded speech into structured clinical notes. It aims to streamline administrative tasks for medical professionals by generating formatted, editable documentation aligned with clinical standards. Therefore, it gives clinicians more time to prioritise patient-centred care and minimise post-consultation administrative tasks.

There have been few studies comparing AI note-taking in comparison to manual note-taking in consultations, one of which compared ChatGPT, Heidi and manual notes in accuracy, readability and efficiency to generate dermatology consultation letters [2]. Heidi was found to be the most consistent and reliable for clinical implementation [2]. The aim of our study was to compare the efficacy of AI note-taking software with healthcare professionals' note-taking in a general practice setting in England, particularly when being utilised by medical students. It explores whether Heidi can be practically and advantageously integrated into general practice, contributing to the wider conversation about the growing role of technology in healthcare settings.

In Section 2, we outline our methods of evaluating Heidi's performance. In Section 3, we start looking at trends seen in our data which we put in a wider clinical context in Section 4, alongside beginning to predict future work and implementation.

II. METHODS

Heidi, an AI medical scribe designed to automate clinical documentation, was used for the purpose of this study. Heidi was chosen due to the ease of use, low running cost and it being one of the few AI tools being particularly developed for use in healthcare at the time. We had no prior association with nor were we approach by Heidi for the purposes of this study.

Across six consultations, AI generated and medical student generated notes were composed. Three reviewers used a four-section classification and 18-point rubric to assess the accuracy of both types of notes for each patient. Scores were also analysed qualitatively depending on the length and complexity of the consultation. Heidi produced complete notes for all six consultations. The mean accuracy score was 84% when excluding irrelevant inaccuracies, with two consultations receiving perfect scores. Most inaccuracies reflected omissions or minor contextual misunderstandings, and clinically significant inaccuracies were rare. Heidi demonstrated high accuracy and potential as an assistive documentation tool in general practice.

Before the study, test runs were conducted to assess Heidi's features and its ability to recognise different accents and multiple people talking in the same consultation. The free version of Heidi was utilised. Six patients participated in the

 ISBN: 978-1-68558-362-0

study. Five in-person patients provided written informed consent to use the Heidi software, have the consultation recorded and use their consultation notes for the study. One patient who had a telephone consultation provided informed verbal consent. Heidi is in compliance with the Data Protection Act and ensures the secure management of data for both the NHS and private practice. Consultations were carried out by two medical students at a time. The primary medical student carried out the consultation while the second medical student monitored software functionality. The primary medical student took brief notes in a notebook during the consultation. They were given three minutes at the end to collate and finalise their consultation notes on a computer. This time pressure was utilised to simulate a realistic setting in a general practice. All primary medical students used this method to standardise note-taking style. Heidi was activated at the start of the consultation and stopped before the general practitioner arrived and reviewed the patient. The AI notes were transcribed and generated in real time by Heidi. The note template used on Heidi for all patients was the "General Practitioner's note". Management plans were not included or analysed in this study.

The patient consultations of this study spanned over three months, during which Heidi updated its note template and transcript formats. Variations in consultation note formats were taken into account during analysis. Three medical students independently reviewed both the AI and human-generated consultation notes. The following rubric was used:
The consultation was divided into six sections:
Presenting Complaint, Past Medical History, Medication History (Including Allergies), Social History, Family History, Examination (If Applicable).
The consultation was marked under four classifications:
1. Accurate
2. Inaccurate, Will Not Make a Difference (Inaccurate WNMD)
3. Inaccurate, Will Make a Difference (Inaccurate, WMD)
4. Not Applicable (N/A).

Each section of the consultation was marked under one of the four classifications. The results were collated, and a consensus was reached after discussion.

Next, a quantitative analysis was performed. Each transcript was scored under a 3-2-1 system which was discussed and agreed upon by three reviewers. This system assigned each of the classifications points: Accurate (3 points), Inaccurate WNMD (2 points), Inaccurate WMD (1 point), N/A (3 points). A section was classified as inaccurate if there was a discrepancy in it between student notes and Heidi's notes. A section was classified as N/A if it was not covered in the consultation, and this absence was confirmed in the transcript, student notes, and Heidi's notes. N/A was assigned 3 points because these items were genuinely not applicable to the consultation, and therefore could not be classified as inaccurate. Treating them as maximally accurate prevented artificial deflation of overall accuracy scores. Each patient transcript was scored out of 18. The frequency of each classification under each section was calculated (e.g., the presenting complaint was accurate in 3 out of 6 patients). Additionally, the total frequency of each classification was calculated (e.g., 5 sections were classified as Inaccurate WMD). While the approach to evaluating AI-generated documentation was similar to that described by Farooq et al. [2], the human comparator differed, with this study using simultaneous manual note-taking by medical students rather than clinician dictation and medical transcription.

III. RESULTS

Overall, 36 sections were assessed, of which 4 were labelled as N/A. These sections were excluded from analysis as they do not influence the evaluation of Heidi's efficacy, leaving 32 total sections classified (Table 1). The amended classification yielded an overall accuracy rate of 59% (Figure 1). When Inaccurate WNMD sections, which do not affect the overall quality of patient notes, are included, the effective accuracy by section rises to 84%. Despite this, 16% of sections remain inaccurate, highlighting areas where Heidi's notes may require verification (Figure 1).

Each patient note was scored based on the classification of its individual sections, with a maximum possible score of 18 (see Methods). Because patients presented with varying complaints and consultation complexities, this scoring allowed Heidi's efficacy to be assessed in the context of each patient. Patients 5 and 6 both achieved perfect scores, reflecting simpler consultations compared to patients 1–4 (Table 2). Overall, Heidi scored 90 out of a possible 108 points, corresponding to an overall accuracy by patient of 83%, a very similar result (Table 2).

TABLE I: CLASSIFICATION OF 32 SECTIONS

Classification	Number
Accurate	19
Inaccurate WNMD	8
Inaccurate WMD	5
N/A	4
Total	36
Total excluding N/A	32

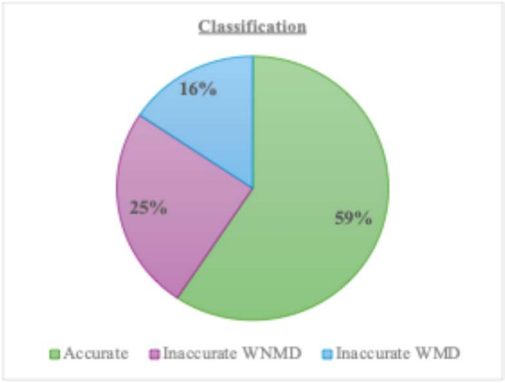


Figure 1: Pie chart showing percentage by classification

TABLE II: POINTS SCORED PER PATIENT

Patient Note	Score
Patient 1	13
Patient 2	14
Patient 3	13
Patient 4	14
Patient 5	18
Patient 6	18
Overall Score	90
Total Possible Points	108

Figure 2: Bar chart demonstrating numerical performance by section

TABLE III: CLASSIFICATION OF PRESENTING COMPLAINT

PC	Number of Sections
Accurate	3
Inaccurate WNMD	1
Inaccurate WMD	2
N/A	0

TABLE IV: CLASSIFICATION OF PAST MEDICAL HISTORY

PMH	Number of Sections
Accurate	3
Inaccurate WNMD	3
Inaccurate WMD	0
N/A	0

TABLE V: CLASSIFICATION OF MEDICATION HISTORY

MH	Number of Sections
Accurate	4
Inaccurate WNMD	1
Inaccurate WMD	1
N/A	0

TABLE VI: CLASSIFICATION OF SOCIAL HISTORY

SH	Number of Sections
Accurate	4
Inaccurate WNMD	2
Inaccurate WMD	0
N/A	0

TABLE VII: CLASSIFICATION OF FAMILY HISTORY

FH	Number of Sections
Accurate	3
Inaccurate WNMD	1
Inaccurate WMD	1
N/A	1

TABLE VIII: CLASSIFICATION OF EXAMINATION

Examination	Number of Sections
Accurate	2
Inaccurate WNMD	0
Inaccurate WMD	1
N/A	3

Each section (as outlined in Methods) was scored according to the number of classifications received across the six patient notes (Accurate, Inaccurate WNMD, etc.). Using the same scoring system, each section was assigned a maximum score of 18, with each section repeating six times for six patients and being scored a maximum of 3 points (Table 2). (Tables 3-8) portray accuracy by section (e.g., presenting complaint). Comparison of these scores highlighted the Presenting Complaint (PC) as the poorest performing section, with a score of 13 (Figure 2). The remaining sections performed similarly, each scoring 15 or 16 points, indicating consistency in Heidi's accuracies and inaccuracies.

IV. DISCUSSION

Looking at the data that was collected, Heidi performed consistently better than initial expectations. Large Language Models (LLM) had been garnering attention for their hallucinatory tendencies and how they frequently produced incorrect or fictional information while generating text [4]. Heidi, as an LLM, was expected to exhibit some of the hallucinatory tendencies seen in other non-healthcare specific LLMs. This was one of the main concerns explored, as inaccurate information in patient notes poses a direct risk to patient safety. According to NHS England, "missing, inaccurate, or non-standard information can lead to inconsistent care and risk the quality and safety of care delivered" [3]. If inaccuracies in Heidi's generated notes reach a significant threshold, this raises concerns that attempts to improve productivity and reduce administrative burden may inadvertently compromise patient safety.

During transcription, we observed one instance of Heidi 'hallucinating' where it recorded that a patient had a sister instead of a brother. This was classified as 'Inaccurate WNMD', as although the detail was incorrect; it did not directly affect patient care. This raises the question of whether such inaccuracies may affect patient confidence. Patients now routinely access their consultation records via the 'Patient Access' and 'NHS' applications. Although inaccuracies within the WNMD section do not directly influence clinical management, they may nevertheless shape patients' perceptions of the overall quality of care received. Such perceptions have the potential to alter the dynamics of

the doctor–patient relationship, where even subtle erosions of trust could contribute to suboptimal patient engagement and, consequently, less favourable long-term outcomes.

Overall, Heidi's outputs were consistently accurate or very close to accurate, with only 5 out of 36 sections formally classified as inaccurate and making a difference. It raises the question as to what is considered an acceptable standard in health. When combined with a professional or student performing a final review before integration into patient records, these findings suggest that Heidi could substantially reduce the administrative burden while maintaining accuracy and ensuring patient errors are not made.

When analysing scores by patient transcript, it was observed that Patients 1–4 presented with more complex conditions, either as first-time presentations or due to extensive past medical histories. In contrast, Patients 5 and 6 represented comparatively 'simpler' consultations, without complex comorbidities or the need for extensive physical examinations. Heidi performed at least on par with, and in some cases better than, the student's notes for Patients 5 and 6. The lowest scores were seen in Patients 1 and 3, each receiving 13/18, which aligned with the greater complexity of their cases.

During the review of the independent analyses, occasional disagreements arose regarding whether certain points should be classified as 'Inaccurate WMD' or as 'Inaccurate WNMD'. Because the overall evaluation focused on Heidi's accuracy in the context of patient safety, the final decision consistently leaned toward the more conservative option by classifying points as 'Inaccurate WMD' when there was uncertainty. The category of 'Inaccurate WNMD' was used for minor errors or irrelevant details that a student would not typically record. Considerable discussion centered on what degree of inaccuracy constitutes a genuine patient safety concern. Consequently, when calculating final accuracy percentages, sections labelled as 'Inaccurate WNMD' were not regarded as posing a direct risk to patient safety and were therefore included in the overall accuracy score.

From a medical education perspective, Heidi can help students focus more on developing the soft skills of patient engagement, while also providing a secondary reference against which they can check their own notes during placement. Nevertheless, this raises an important concern regarding how to ensure that clinicians consistently conduct thorough reviews and whether insufficient oversight could inadvertently increase risks to patient safety. Addressing this question lies beyond the scope of the present study but provides scope for further research.

While Heidi exceeded expectations in this study, it comes with its limitations. As with any AI input system, Heidi could not interpret when a patient would point to body parts to explain their symptoms. Heidi transcribed statements verbatim without contextual interpretation on certain occasions. For instance, if a patient said they had a 'dry cough' but produced 'green mucous', Heidi would state the same in the notes. However, a dry cough by definition does not produce mucus. Additionally, Heidi had limited capability to correct information. It frequently stated the first thing a patient said, even if they changed their description throughout the duration of the consultation. Heidi's notes described negative signs and symptoms that a patient had been asked about (no fever, no chest pain) while notes taken by healthcare professionals generally only include positive signs and symptoms unless negative symptoms are diagnostically relevant. While these may be accurate, the question arises of whether recording too much could lead to an information overload where important details are overlooked.

Our study has several limitations. We had a small sample size of six patients, which limits statistical power and the ability to generate findings or test for significance. A reviewer bias may prevail over the descriptive analysis carried out, as whether a point will or will not make a difference can be subject to interpretation. The reviewers were medical students, so judgments about clinical significance may differ from experienced practitioners. Natural variation in typing speed and computer proficiency is expected between medical students, as would be expected between different general practitioners. Additionally, medical students were aware that their documentation would be evaluated, which may have introduced a Hawthorne effect, potentially inflating note quality relative to routine practice. Other factors, such as usability and clinician/patient satisfaction were not measured. This study was only carried out in a general practice setting, but further research should be carried out in different medical settings where outcomes may differ. This study is biased towards one AI tool. Finally, our findings relate to the version of Heidi used during the study. As it is a continually updated software, updates may change accuracy and performance

V. Conclusion and Future Work

All things considered, Heidi achieved an impressive accuracy rate of 84%. While complete accuracy may be an ambitious expectation, even small inaccuracies carry potential implications for patient safety. This reinforces the necessity (emphasised by Heidi itself upon login) that all generated notes must be reviewed before integration into patient records. Heidi functions effectively as a supplement to human documentation but is not yet capable of replacing a general practitioner. Its main strength lies in reducing administrative burden: by generating draft notes, it allows clinicians - or medical students, in this case, to concentrate more fully on patient interaction during consultations, with their role shifting to reviewing and correcting the output. Further research needs to be carried out with larger sample sizes and in a variety of settings to gain more clarity on this subject. It could place attention on long-term effect studies to evaluate how medical scribe use influences clinician burnout, efficiency, and patient outcomes over months or years.

Larger multi-centre studies would generate stronger evidence of long-term advantages and possible drawbacks of AI medical scribes. This trial was not blinded, but future iterations could address this by using practices that routinely record calls and document consultations manually. These recordings and notes could then be randomly selected for review, helping to minimise perfectionism bias that may arise when clinicians or students are aware that their documentation is being assessed. As well as this, comparative studies should be carried out to compare Heidi against other AI-driven scribes, such as ChatGPT, DeepScribeAI, Nuance DAX, Suki AI, Abridge, evaluating which system is the most effective in terms of efficiency, accuracy, and user satisfaction. Given that AI medical scribes vary in their intended purpose, from primary care workflow support to cost efficiency, and clinician explainability, further studies should have evaluation frameworks that measure outcomes aligned with the scribe's intended use. These studies could compare their use in a general practice setting versus a hospital setting, as well as comparison across different specialities, including psychiatry, paediatrics and surgery, to assess whether effectiveness differs by clinical context. It's important that future research investigates cost-benefit analyses to assess if the increase in efficiency from an AI medical scribe outweighs the implementation costs. The wider financial impact of adopting the system at a regional or national scale should also be determined. This would provide crucial data to optimise resource allocation and support decisions on large-scale implementation.

Research should evaluate the impact of AI medical scribes on both patient and staff experience. Regarding patients' experience, patient satisfaction, communication, and privacy concerns should be explored. Notably, patients attending for mental health–related consultations were less willing to participate, highlighting important considerations regarding patient attitudes toward the use of AI in clinical settings. In reference to staff experience, clinician burnout is often linked to excessive administrative workload, so further studies can demonstrate whether the use of AI scribes can reduce administrative burden and therefore improve clinician wellbeing. The current landscape underscores the need for robust evidence on the safety, efficiency, and real-world performance of large language models, as highlighted by the NHS [3], before their broader deployment in medical documentation. Given the rapid advancement of digital technologies, this study aims to evaluate whether such tools can enhance clinical efficiency by reallocating clinician time toward direct patient care. This question is increasingly pertinent as AI becomes more accessible and further embedded within healthcare systems. Another important area of research is the effectiveness of the integration of Heidi with different Electronic Health Records (EHR), such as Epic and EMIS, as smooth integration is needed for its practical use. Such studies should investigate whether integrating Heidi with EHR systems would impact the accuracy of documenting records and operational workflow.

Occasional errors, such as missing medications from the patient's history, could lead to incomplete or inaccurate records and pose risks to patient safety. This does not exclude the possibility that Heidi may eventually achieve full accuracy; however, in its current form, the system still necessitates a degree of clinician oversight, as acknowledged by the software itself. From a psychological standpoint, increasing software accuracy may paradoxically reduce the robustness of final human review over time, as clinicians become less vigilant when errors appear infrequent. It could therefore be argued that the software is, in some respects, safer when errors are more common, as this heightened error frequency prompts greater clinician alertness. Nevertheless, an inaccurate scribe ultimately presents a significant safety concern. Implementing structured safety checks similar to those embedded in electronic prescribing systems. may provide a viable mitigation strategy.

In conclusion, this study adds to the growing body of research demonstrating the promise of AI-assisted clinical documentation. Ensuring that AI scribes genuinely enhance patient care, support clinician workflow, and maintain high standards of clinical governance will be critical to their responsible integration into routine practice.

ACKNOWLEDGMENTS

Ethical approval was not required for this small-scale service evaluation. All patients provided informed written or verbal consent for participation and the use of their consultation data. This research did not receive funding from any agency in the public, commercial or not-for-profit sectors. The authors thank the general practice staff in North Central London for facilitating data collection, and the patients who generously consented to participate. The authors would also like to express their gratitude to their supervisor, Dr. James Rusius, for his guidance and oversight during the study.

AUTHOR CONTRIBUTIONS

Conceptualisation: SS; Methodology: SS, AS, CH; Validation: SS, AS, CH; Investigation (data collection): SS, CH; Data curation: SS, AS, CH; Formal analysis: SS, AS, CH; Writing - Original Draft: SS, AS, CH; Writing – Formatting and Editing: SS, AS. All authors reviewed and approved the final manuscript.

REFERENCES

[1] Heidi Health, "About Heidi Health – Company Story | UK." *Heidi Health*, 2019. [Online]. [retrieved: August, 2025] Available: https://www.heidihealth.com/uk/about-us/company

[2] F. Farooq, H. Cooper, A. Shipman, and C. D. Michell, "Artificial intelligence verses traditional method in generating dermatology consultation letters: a pilot study comparing accuracy, readability, and efficiency," *Clin. Exp. Dermatol.* 2025. [Online]. [retrieved: August, 2025] doi: https://doi.org/10.1093/ced/llaf323,

[3] NHS England, "High quality patient records," *NHS England*, 2022. [Online]. [retrieved: August, 2025]

Available: https://www.england.nhs.uk/long-read/high-quality-patient-records/

[4] Y. Sun, D. Sheng, Z. Zhou, and Y. Wu, “AI hallucination: towards a comprehensive classification of distorted information in artificial intelligence-generated content,” *Humanit. Soc. Sci. Commun.*, vol. 11, no. 1, 2024. [Online]. [retrieved: August, 2025] doi: https://doi.org/10.1057/s41599-024-038

SmartCHANGE: From Risk Prediction to Daily Habits Through AI and Gamified Lifestyle changes

Valentina Di Giacomo, Federica Saccà
Elena Mancuso
AI&Data for Health
Engineering Ingegneria Informatica
Rome, Italy
email: valentinadigiacomo@eng.it,
federicasacca@eng.it, elena.mancuso@eng.it

Lotte van der JagdJagt, Harm op den Akker
Martijn Vastenburg
Research & Development
ConnectedCare
Nijmegen, The Netherlands
email:l.v.d.jagt@connectedcare.nl,
h.opdenakker@connectedcare.nl,
m.h.vastenburg@connectedcare.nl

***Abstract* — Lifestyle-related chronic conditions are increasingly emerging during childhood and adolescence, highlighting the need for effective early prevention strategies beyond traditional care settings. Although digital health technologies enable continuous monitoring and personalized support, many fail to sustain engagement or deliver actionable insights to Health-Care Professionals (HCPs). This paper presents SmartCHANGE, an integrated digital ecosystem that combines AI-based risk prediction with lifestyle change support. The system includes *HappyPlant*, a gamified mobile app for children and families, and a web-based application for clinicians. While the app promotes healthy behaviors through goal-setting and personalization, the HCPs web app supports data interpretation and tailored feedback. Preliminary findings indicate strong user acceptance and improved clinical efficiency, suggesting potential for scalable preventive care.**

Keywords: non-communicable diseases; digital tools; lifestyle change; preventive healthcare.

I. INTRODUCTION

Non-communicable Chronic Diseases (NCDs), including diabetes and cardiovascular conditions, are increasingly diagnosed during childhood and adolescence and are strongly associated with poor nutrition, physical inactivity, and other modifiable lifestyle factors. These behaviors often develop early in life and persist into adulthood, increasing long-term health risks and healthcare burden [1]. Effective early prevention is therefore essential; however, conventional healthcare models provide limited insight into young patients' daily behaviors between consultations, constraining timely intervention. Digital health technologies offer opportunities for continuous monitoring and personalized support outside clinical environments. Despite this potential, many existing solutions struggle to maintain engagement among young users and rarely provide HCPs with clinically actionable insights [2] [3]. Adolescents respond more positively to supportive, personalized, and playful experiences than to risk-based messaging, while clinicians require interpretable analytics that integrate efficiently into clinical workflows.

To address these challenges, this paper introduces SmartCHANGE, an integrated digital ecosystem that combines AI-based risk prediction with lifestyle change support across two interconnected platforms: *HappyPlant*, a mobile application for children and families, and a web-based application for HCPs. In Section II, the methodology adopted is presented, while in Section III we describe the system design and architecture, and report initial findings from early deployment and formative evaluation. A discussion about the current findings is reported in Section IV along preliminary conclusion drawn in Section V.

II. METHODOLOGY

SmartCHANGE was developed through an iterative co-design process involving children, families, and HCPs across Europe to identify user needs, engagement barriers, and expectations for preventive technologies. Children and families preferred systems that emphasize positive reinforcement, offer clear and achievable goals, and enable playful, personalized interaction. Clinicians highlighted challenges such, as limited consultation time, limited insight into patients' daily behaviors, difficulties translating data into actionable guidance, and maintaining continuity of care beyond clinical visits. These inputs directly informed the system's design and functional requirements, ensuring alignment with real-world workflows and developmental needs.

The SmartCHANGE solution (see Figure 1) adopts a modular, microservices-based architecture to enable scalable and secure deployment. It includes the *HappyPlant* mobile app, a web application for HCPs, and a backend for data processing, AI services, and security. Data from wearables and questionnaires are securely transmitted to the backend, where predictive models update risk estimates. The platform complies with GDPR, using encryption, pseudonymization, and audit logging, while OpenID Connect supports secure, role-based access.

 ISBN: 978-1-68558-362-0

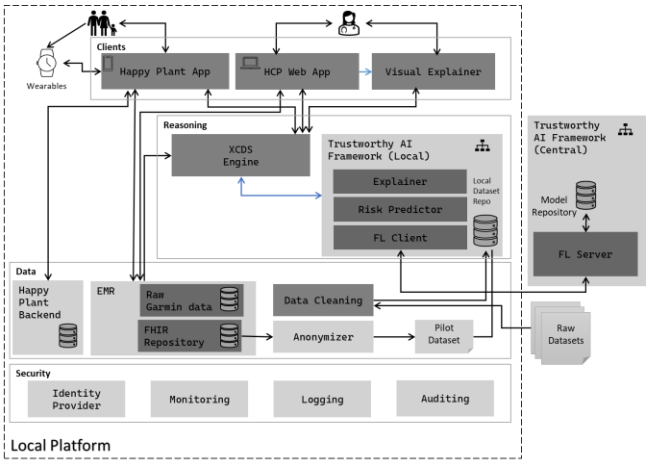


Figure 1: SmartCHANGE overall solution architecture

Machine-learning models trained via federated learning estimate health risk from longitudinal data. The models are neural networks that take primarily behavioural risk factors as input (physical activity, diet, fitness, sleep, smoking, alcohol consumption and body mass index) and output the risk of cardiometabolic disease until the age of 65 years. The models were trained on seven datasets that track health variables of thousands of people over more than 40 years. Clinicians receive interpretable outputs, and users receive age-appropriate, goal-oriented feedback to encourage lifestyle change.

Feasibilty of SmartCHANGE tools has been planned through a 3-months observational study across five sites (Portugal, Slovenia, Finland, the Netherlands and Taiwan) and will engage around 100 children aged 6-10 years and their families, plus about 150 adolescents aged 11-14 years. Real-world settings will be used, which means that each of the sites will involve type of HCPs depending on the standard of care in a particular setting (i.e., school nurses, family physicians or pediatricians).

III. RESULTS

A. *HappyPlant*

The *HappyPlant* mobile app was developed as a lifestyle change companion for adolescents and families, translating complex lifestyle guidance into an engaging, game-based experience. *HappyPlant* designs were evaluated during several co-design sessions with end-users and subsequently tested during an initial internal deployment phase in which more than 30 testers used the app over approximately three months. In the final end-user evaluation, usage data and user feedback will be collected to assess feasibility, engagement, and usability.

Users adopt a digital plant that grows as they earn “Growth Water” through positive behaviors, such as being physically active and completing healthy lifestyle challenges. To support behavior change, the app provides personalized daily and weekly goals related to movement, sleep, nutrition, and mindfulness. These goals are framed positively, emphasizing achievability and autonomy. Progress toward goals is visualized through Growth Water, which users collect and use to nurture their plant. Gamified elements, including accessories, streaks, reward gardens, and playful notifications are designed to sustain long-term engagement.

To monitor lifestyle behaviors, the application collects data on physical activity and sleep through a consumer wearable device, while dietary habits and body mass index are assessed using questionnaires. Importantly, although goals and challenges are ultimately tailored based on a calculated risk prediction (with a human HCP in the loop), the app avoids communicating medical risk scores directly and instead focuses on positive reinforcement and actionable steps.

B. *Web-based application*

Co-design activities informed development priorities and shaped the functionalities of the SmartCHANGE web app, which was piloted across the target user groups (primarily HCPs, teachers and administrators) across 3 clinical sites. HCPs used the dashboard to review behavioral data trends, track personalized goal progress, and AI-assisted risk estimates. The application collects the information on fitness, smoking and alcohol consumption, as well as non-behavioural traditional cardiovscular risk factors such as blood pressure and lipid when available. This additional information helps to improve the quality of risk prediction. Preliminary evaluation was conducted using structured feedback forms and semi-structured interviews.

By testing the web app, HCPs reported improved visibility into patients’ daily behaviors between visits and greater confidence in identifying lifestyle-related risk patterns. Structured data presentation and visualization of longitudinal trends support more focused clinical discussions and helps reducing time spent interpreting raw data. Preliminary pilot feedback indicates that consultations are more efficient and goal-oriented when supported by the web app, facilitating shared decision-making with patients and families. The most valued features included trend analysis, early risk flagging, and support for tailored feedback. Although long-term clinical outcomes were not assessed in this phase, early findings suggest that the system improves workflow efficiency and enhances the clinical relevance of lifestyle data.

IV. DISCUSSION

SmartCHANGE addresses core limitations of current preventive health technologies by combining predictive modelling with user-centered design and clinician integration [2]. Unlike many digital interventions that motivate users or analyze risk in isolation, SmartCHANGE integrates these functions within a single interoperable system, enabling personalized intervention informed by

continuous lifestyle data. Early feedback indicated that explainable outputs and structured data visualization supported clinicians' interpretation of risk patterns and contributed to trust in the system's predictions. At the user level, the focus on positive reinforcement and achievable goals aimed to promote sustained engagement without increasing anxiety associated with medical risk communication. According to preliminary findings, after 5 weeks of the study in Slovenia and Finland more than 90% of participants are still engaged with the *HappyPlant* app. Moreover, initial feedback from school nurses in Finland is very positive, emphasizing the utility of the SmartCHANGE web app features for their daily clinical practice.

V. CONCLUSIONS AND FUTURE WORK

This paper highlights the potential of AI-driven digital ecosystems to advance preventive healthcare for young populations by enabling early intervention, continuous support, and data-informed care. Ongoing studies are assessing feasibility, engagement and lifestyle change impact, while future work will focus on electronic health record integration, model refinement, and broader usability evaluation.

REFERENCES

[1] World Health Organization, "Noncommunicable diseases," WHO Fact Sheet, World Health Organization, Geneva, Switzerland, 2023. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/noncommunicable-diseases [retrieved: February, 2026].

[2] A. Thacharodi, P. Singh, R. Meenatchi, et al., "Revolutionizing healthcare and medicine: The impact of modern technologies for a healthier future—A comprehensive review," Health Care Sci., vol. 3, no. 5, pp. 329–349, Oct. 2024, doi:10.1002/hcs2.115.

[3] M. Bajramagic, T. Battelino, X. Cos, et al., "Artificial intelligence driven clinical decision support systems to assist healthcare professionals and people with diabetes in Europe at the point of care: a Delphi-based consensus roadmap," Diabetologia, 2025, doi:10.1007/s00125-025-06601-5.

Interpreting Human Ambiguity through Neuro-Fuzzy Intelligence in Holistic Healthcare Lifecycle

Andrea Pitrone
Loop AI Labs Inc.
New York, USA
email: andrea@loop.ai

Intissar Haddiya
Faculty of Medicine and Pharmacy
University Mohammed First
Oujda, Morocco
email: i.haddiya@ump.ac.ma

***Abstract*—Ambiguity in modern healthcare often arises from the intrinsic imprecision of human communication, a challenge intensified by the vulnerability of patients. Existing AI systems in healthcare primarily focus on structured physiological data, often failing to adequately process the inherent imprecision and subjective ambiguity of human communication. To address this gap, we propose an intelligent framework based on Neuro-Fuzzy Systems (NFS) that unifies cognitive, aleatoric, and communication uncertainties. This paper presents a holistic framework for managing communication-driven ambiguity across the entire medical continuum, from the training of future clinicians to patient care and post-treatment monitoring. By integrating the learning capabilities of Neural Networks with the interpretability and uncertainty-handling strengths of Fuzzy Logic, Neuro-Fuzzy Systems offer a robust and trustworthy means of modeling human-like approximate reasoning for medical intelligence. The major relevant novelty introduced by this article is the treatment of human communication ambiguity as a primary modeling target, in contrast to conventional neuro-fuzzy applications focused mainly on physiological or structured data. Through a domain-specific handling of linguistic expressions—including subjective student self-assessment and patient-reported outcomes containing hedges, such as slightly or very—the approach unifies cognitive, aleatoric, and communication uncertainty. Interpretable Neuro-Fuzzy models are applied to calibrate metacognitive judgments, integrate ambiguous narratives with Internet-of-Medical-Things data, and support clinical decisions, enabling more trustworthy, patient-centered systems. Ultimately, readers will grasp how adapting linguistic fuzzification to clinical workflows enables a highly transparent decision support system that bridges the gap between patient subjectivity and objective medical data.**

Keywords-Adaptive Neuro-Fuzzy Inference System; Ambiguity; Artificial Intelligence, Deep Learning, Holistic Healthcare; Internet of Medical Things.

I. INTRODUCTION

This section provides the clinical context for the research, defines the holistic healthcare lifecycle, introduces neuro-fuzzy systems as the proposed solution, and outlines the overall structure of the paper.

A. Context

The medical field is undergoing a profound transformation driven by advancements in genomics, high-resolution imaging, and ubiquitous monitoring enabled by the Internet of Medical Things (IoMT) [1]. The proliferation of heterogeneous data requires advanced computational tools capable of supporting a shift from reactive interventions to proactive and personalized medicine. Central to this evolution is the development of robust Artificial Intelligence (AI) systems that can effectively assist clinical decision-making and administrative processes.

However, the performance of purely data-driven models, such as Deep Learning, is often limited by the intrinsic complexity of medical reality. Unlike well-structured computational problems, medicine is characterized by uncertainties that do not conform to crisp or binary classifications [2]. A fundamental challenge arises from the epistemological gap between the deterministic nature of conventional algorithms and the inherent indeterminacy of biological systems and human communication. Expert clinicians routinely synthesize imperfect, incomplete information—which is filtered through years of subjective experience—in a process that is fundamentally non-linear and tolerant of partial truths.

These uncertainties can be broadly categorized into two key types, each requiring distinct computational treatments:

- Aleatoric Uncertainty (Data Noise): Originates from randomness in data acquisition, such as sensor noise, imprecise biological measurements, and the subjective or vague manner in which patients describe symptoms or well-being. Aleatoric uncertainty is irreducible and must be quantified and propagated through the model.
- Epistemic Uncertainty (Model or Cognitive Uncertainty): Arises from gaps in knowledge, ambiguity in human reasoning, and the subjective nature of cognitive states—for example, a medical student's self-assessed readiness or a clinician's degree of diagnostic confidence. Epistemic uncertainty is potentially reducible through additional data or improved model structure. The inadequacy of traditional crisp logic to represent linguistic expressions like "slightly elevated" or "moderate pain" illustrates this challenge.

Successful AI integration into clinical workflows, therefore, requires models that address both types of uncertainty not merely through probabilistic methods, but through transparent mechanisms that emulate human-like approximate reasoning.

Despite their significant advantages, Neuro-Fuzzy Systems face certain limitations. Notably, they are susceptible to the 'curse of dimensionality'; as the number of high-dimensional medical inputs increases, the required rule base and computational overhead grow exponentially. Additionally, the initial configuration of membership functions relies heavily on domain expert knowledge, making the system initially sensitive to subjective human biases during the design phase.

B. The Holistic Healthcare Lifecycle

To effectively support the wide range of AI applications in healthcare, an agentic AI-based platform must operate holistically across the entire healthcare lifecycle, depicted in Figure 1—Training, Diagnosis, Treatment, and Monitoring—through reusable, intelligent components. Current AI systems are typically siloed, focusing on isolated phases (e.g., imaging for diagnosis or anomaly detection for monitoring), which creates discontinuities within the intelligence pipeline.

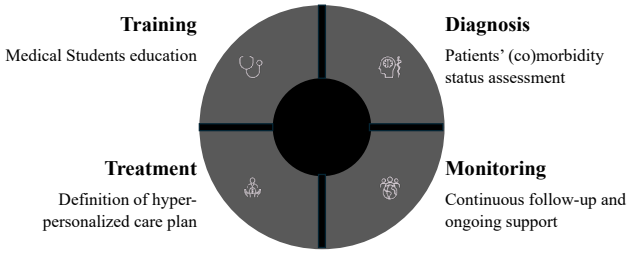


Figure 1. Holistic Healthcare Lifecycle.

This article argues that the successful integration of AI into this high-stakes ecosystem depends on its ability to explicitly model and manage ambiguity. A robust system must treat medical knowledge as a continuum rather than as a set of rigid boundaries. For example, assessing a patient's temperature requires determining its degree of membership in both the "Normal" and "Slightly Elevated" ranges, rather than forcing it into a single, arbitrary category. Similarly, interpreting "moderate pain" as a fuzzy set instead of a precise numerical value—or aligning a student's "high confidence" with actual performance risk—can significantly enhance trust and reliability. Such a holistic framework demands a system that is both adaptive to evolving data (learning) and interpretable to human users (reasoning).

C. Neuro-Fuzzy Systems as a Solution

Neuro-Fuzzy Systems (NFS) constitute a powerful class of Hybrid Intelligent Systems (HIS) that combine the strengths of Artificial Neural Networks (ANNs) and Fuzzy Logic (FL) while mitigating their respective limitations.

A comparison of ANNs and FL characteristics is shown in Table I.

TABLE I. ANNS AND FL COMPARISON

Paradigm	Strength	Weakness	NFS Contribution
ANNs	Strong learning capacity, pattern recognition, non-linear function approximation	Black-box behavior, lack of interpretability, limited handling of linguistic knowledge	Provides adaptive learning by optimizing parameters from data
FL	Effective handling of ambiguity, linguistic modeling, high interpretability	Requires expert-defined parameters, limited learning ability	Provides an interpretable rule-based structure

The synergy between these paradigms enables NFS to learn optimal parameters (e.g., membership function shapes and rule consequences) from data while preserving an interpretable, human-readable rule base for inference. This inherent transparency satisfies the growing demand for Explainable AI (XAI) in safety-critical domains such, as healthcare [3].

Two conceptually related use cases presented in this article demonstrate how NFS can effectively manage ambiguity in both human learning and doctor–patient communication.

D. Structure of the paper

This paper presents a practical application of the neuro-fuzzy approach within the holistic medical continuum to resolve and disambiguate uncertainty in complex clinical information.

- In Section II, the core neuro-fuzzy algorithmic architecture is detailed.
- Section III presents a use case mitigating cognitive ambiguity in medical training.
- Section IV explores handling communication ambiguity in remote monitoring.
- In Section V, broader applications across the medical continuum are discussed.
- Section VI outlines implementation criteria.
- Sections VII and VIII cover future directions and conclusions.

II. NEURO-FUZZY ALGORITHMS: THE ARCHITECTURAL CORE

The following discussion establishes the mathematical foundations of fuzzy logic for handling ambiguity and details the five-layer architecture of Adaptive Neuro-Fuzzy Inference Systems (ANFIS).

A. Foundations of Fuzzy Logic (FL): Handling Ambiguity

Fuzzy Logic, introduced by Zadeh [4] in the 1960s, provides the conceptual and mathematical framework required to model imprecision and gradual transitions in real-world phenomena. In classical set theory, an element x either belongs to a set A or it does not, i.e., $\mu_A(x) \in \{0,1\}$.

Fuzzy Logic extends this formulation, allowing elements to possess a degree of membership in a set. This degree is encoded by the membership function $\mu_A(x)$, which maps each element $x \in X$ to a continuous value in the interval [0, 1].

For the linguistic variable Pain Level, Gaussian membership functions are commonly employed due to their smooth differentiability, which is advantageous for the ANN training stage in ANFIS architectures. A Gaussian membership function is defined by its center c and width (standard deviation) σ:

$$\mu_A(x) = \exp\left(-\frac{1}{2}\left(\frac{x-c}{\sigma}\right)^2\right) \quad (1)$$

The Fuzzy Rule Base formalizes the inference mechanism using linguistic variables.

The standard, generalized Mamdani-type rule involving multiple inputs ($x_1, x_2, \ldots, x_n$) and a single output (y) can be written as:

IF x_1 is A_1 AND x_2 is A_2 AND… AND x_n is A_n THEN y is B

where:

- $x_1, x_2, \ldots, x_n$ are the input linguistic variables (the system's measured inputs).
- $A_1, A_2, \ldots, A_n$ are the antecedent fuzzy sets (linguistic terms like small, medium, hot, fast) that define the state of the input variables.
- y is the output linguistic variable (the system's control action or decision).
- B is the consequent fuzzy set (linguistic terms like increase, stop, high, low) that defines the desired output.
- AND is the fuzzy operator (or t-norm), typically implemented using the minimum function (min) or product function (prod).

In the context of a simple temperature control system:
IF temperature is HOT and fan_speed is LOW THEN heater_power is DECREASED_SUBSTANTIALLY

This rule structure enables approximate reasoning, a core human cognitive capability that supports inference from incomplete, ambiguous, or imprecise premises. Such reasoning directly parallels the subjective decision-making processes commonly observed in clinical assessment.

B. Adaptive Neuro-Fuzzy Inference Systems (ANFIS)

ANFIS [5] is a hybrid intelligent system that combines the adaptive learning capabilities of neural networks with the linguistic and reasoning abilities of a Fuzzy Inference System (FIS). It is the most common and robust architecture for NFS, primarily utilized to implement the Takagi-Sugeno-Kang (TSK) [6] model. The structure is a five-layer feedforward network, as shown in Table II, where each layer performs a specific stage of the fuzzy inference process, whose parameters are optimized via the hybrid learning algorithm.

TABLE II. FIVE-LAYER NETWORK STRUCTURE

Layer	Function	Parameters Adjusted by Learning
1: Fuzzification	Determines the membership grades for each input. Every node corresponds to a linguistic label and computes the membership function (μ).	Adaptive (Premise Parameters: shape, center, width of MFs).
2: Product (Rule)	Calculates the firing strength (ω_i) of each fuzzy rule. Each node represents one rule and typically performs the fuzzy AND operator (T-norm, usually multiplication).	Fixed (Π operator).
3: Normalization	Normalizes the firing strength of each rule by dividing it by the sum of all rule firing strengths.	Fixed (Normalization operator).
4: Consequent	Calculates the weighted output of each rule, $O_{4,i}$. This involves the normalized firing strength ($\bar{\omega}_i$) multiplied by the TSK rule consequence (f_i).	Adaptive (Consequent Parameters: p_i, q_i, r_i).
5: Defuzzification	Computes the final, crisp output of the entire system by summing all the weighted rule outputs from Layer 4.	Fixed (Σ operator).

A description of the role and operation of each layer is provided in the following paragraphs.

- Layer 1: Fuzzification Layer
  Node Function: $O_{1,i} = \mu_{A_i}(x)$ (or $\mu_{B_i}(y)$) (2)
  Role: Takes the crisp input (e.g., a measured temperature) and maps it to a membership degree between 0 and 1, indicating the degree to which it belongs to a fuzzy set (e.g., "Hot"). The parameters that define the shape of the membership functions (like the mean and standard deviation of a Gaussian function) are adjusted during training.
- Layer 2: Rule Layer (Product Layer)
  Node Function: $O_{2,i} = \omega_i = \mu_{A_i}(x) \cdot \mu_{B_i}(y)$ (3)
  Role: Each node here represents one TSK rule. It computes the firing strength or degree of fulfillment (ω_i) for that rule by multiplying (or taking the minimum of) the membership grades received from Layer 1. This value represents how strongly the antecedent part of the rule is satisfied.

- Layer 3: Normalization Layer ($\bar{\omega}_i$)
  Node Function: $O_{3,i} = \bar{\omega}_i = \frac{\omega_i}{\sum_k \omega_k}$ (4)
  Role: Calculates the normalized firing strength () for each rule. This step is essential because the normalized values act as the weights for the final output calculation in Layer 5.
- Layer 4: Consequent Layer
  Node Function: $O_{4,i} = \bar{\omega}_i \cdot f_i = \bar{\omega}_i \cdot (p_i x + q_i y + r_i)$ (5)
  Role: Each node computes the contribution of its corresponding rule to the overall output. It multiplies the normalized firing strength ($\bar{\omega}_i$) by the crisp TSK rule consequent (f_i), which is a linear equation of the inputs. The consequent parameters (p_i, q_i, r_i) are tuned using the Least Squares Estimation (LSE) method.
- Layer 5: Output Layer (Defuzzification)
  Node Function: $O_{5,i} = \sum_i O_{4,i} = \sum_i \bar{\omega}_i f_i$ (6)
  Role: The single node in this layer calculates the overall final output of the ANFIS model by summing the outputs of all contributing rules. This final result is the crisp, defuzzified output, which is the system's prediction or control signal.

C. Learning Mechanism

The hybrid optimization algorithm is what grants ANFIS its formidable learning capacity:

- Forward Pass (LSE): When premise parameters are fixed, the output O is a linear combination of the consequent parameters. The LSE method is used to estimate the optimal consequent parameters (p_i, q_i, r_i) that minimize the error. The solution is found by solving the matrix equation $A \cdot X = B$, where X is the vector of unknown consequent parameters, and the solution is given by $X = (A^T A)^{-1} A^T B$.
  This makes optimization fast and analytically solvable.
- Backward Pass (Gradient Descent): Once the optimal consequent parameters are found, the output error is propagated backward through the network. The gradient descent method is then applied to adjust the non-linear premise parameters (e.g., c and σ) to further minimize the error. This combination ensures a computationally efficient training process while maintaining the interpretability of the final structure.

The ANFIS approach thus offers a significant advantage over pure ANNs by allowing the system to begin with a structure informed by expert fuzzy knowledge (initial MFs and rules) and then refine that structure using the predictive power of machine learning.

III. Use Case I: Mitigating Cognitive Ambiguity in Medical Education (Student-AI Interaction)

This use case examines the challenge of cognitive uncertainty in clinical training, describes the ANFIS-based self-assessment model, and explains how adaptive learning corrects student overconfidence.

A. The Challenge of Cognitive Uncertainty in Training

Medical education requires students to achieve a high degree of confidence and competency, often involving high-stakes licensing or board exams.

A critical challenge is the inherent epistemic ambiguity in a student's self-assessment of their 'readiness'. This self-perception is a composite, fuzzy variable influenced by:

- Subjective Confidence and the Dunning-Kruger Effect [7]: Low-performing students often suffer from inflated self-assessments (overconfidence), while high-performing students may suffer from imposter syndrome (underconfidence). Traditional scalar self-assessments fail to account for this systematic cognitive bias.
- Emotional and Psychological Factors: Variables like Stress, Fatigue, and Motivation are intrinsically fuzzy, yet they profoundly impact exam performance.
- Ambiguity in Communication: The linguistic description of knowledge status (e.g., "I know enough, but not everything about cardiology") is a subjective input that must be objectively calibrated.

The goal of applying ANFIS, as presented in this article, to the Readiness Self-Assessment Tool (RSAT), that we previously developed [8], is to correct students' cognitive ambiguity, providing an objective calibration that promotes metacognition.

B. The ANFIS-Based Self-Assessment Model

The ANFIS-powered Readiness Self-Assessment Tool (RSAT) is a diagnostic tool for metacognitive correction.

1) Model Inputs (Linguistic Variables)

The system uses the following fuzzy inputs, collected via daily or weekly subjective logs:

- Confidence in Subject (e.g., Nephrology): Fuzzy sets (Low, Medium, High).
- Perceived Stress Level: Fuzzy sets (Minor, Moderate, Severe).
- Perceived Knowledge Gaps (Inverse of Mastery): Fuzzy sets (Few, Average, Many).

The crisp input (e.g., a 1-10 rating) is immediately converted into degrees of membership $\mu_{A_i}(x), \mu_{B_j}(y), \mu_{C_k}(z)$.

2) The Fuzzy Rule Base and Cognitive Calibration

The core of the system is the rule base, which models the relationship between cognitive inputs and the desired output: Exam Readiness Score (a crisp output between 0 and 100). The general structure of a rule is:
IF Confidence is A_i AND Stress is B_j AND Gaps are C_k,
THEN Readiness Score is f_i.

The consequent functions are linear TSK functions whose parameters are learned from a longitudinal training dataset composed of past students' self-assessments paired with their objective, actual exam results.

The ANFIS learning phase achieves the crucial cognitive calibration. The ANN component automatically adjusts the shape and position of the membership functions (e.g., $\mu_{HighConfidence}$). If students reporting 'High Confidence' often score only 75%, the ANFIS training algorithm shifts the center and narrows the spread of the $\mu_{HighConfidence}$ function. This forces the student to report a higher subjective score in the future to achieve the same degree of membership in the $\mu_{HighConfidence}$ set, effectively de-biasing the student's cognitive ambiguity by aligning their linguistic perception with the objective performance metric.

C. Learning and Adaptation (ANN Role) (Impact of Adaptivity)

The adaptive role of the ANN in the RSAT is critical for overcoming the psychological biases inherent in self-assessment. The ANFIS model performs a dynamic, personalized form of cognitive calibration. Over time, the NFS model transitions from being purely descriptive of the student's stated feelings to being predictive of their true competence.

This adaptation allows the RSAT to provide two layers of feedback:

- Content-Specific Feedback: Derived from the activated rules (e.g., Rule: High Gaps Low Readiness).
- Metacognitive Feedback: Derived from the changes in the learned Membership Functions. For instance, if the $\mu_{HighConfidence}$ curve shifts far to the right for a specific student, the system can provide a direct warning: "Your reported confidence levels are consistently higher than your performance history suggests. Consider recalibrating your self-assessment." This direct feedback mechanism fosters a collaborative, non-punitive AI-Student relationship.

A concrete medical application of this tool occurs during high-stakes clinical rotations, such as Cardiology. A medical student might consistently report 'High Confidence' in identifying arrhythmias despite repeatedly misinterpreting complex ECGs in practice. The ANFIS-powered tool dynamically adjusts the student's cognitive baseline, alerting them to this overconfidence bias and requiring a demonstrably higher standard of objective diagnostic performance to validate their subjective readiness.

IV. Use Case II: Handling Communication Ambiguity in Remote Patient Monitoring (Doctor-Assistant-Patient)

This section explores health status ambiguity within the IoMT, details the architecture of a neuro-fuzzy virtual assistant, and presents a formal mathematical approach for processing linguistic hedges in patient reports.

A. IoMT and the Ambiguity of Health Status

Remote Patient Monitoring (RPM) leverages the IoMT—a network of connected devices (wearables, home sensors, implanted devices)—to capture continuous streams of physiological data. This environment introduces significant challenges: aleatoric uncertainty (sensor drift, environmental noise, signal artifacts) and communication ambiguity from the patient via Patient-Reported Outcomes (PROs) [9] [10].

The patient's subjective report is often the earliest indicator of a change. A statement like, "I feel a little short of breath when I walk the dog," requires the Virtual Health Assistant (VHA) to fuse temporal, context-dependent, and linguistic information with the objective data streams (e.g., respiratory rate, heart rate variability, oxygen saturation (SpO_2)). The challenge is that the patient's language is governed by fuzzy, imprecise rules, not crisp logic.

B. NFS-Powered Virtual Health Assistant (VHA) Architecture

A VHA utilizes an NFS framework, ideally an Evolving Neuro-Fuzzy System (ENFS) for real-time, streaming data environments, to perform multi-source data fusion and risk stratification.

1) Multi-Source Fuzzy Inputs

The VHA's ANFIS core is fed two distinct sets of fuzzy inputs:

- Fuzzified Sensor Data (Aleatoric Handling): Crisp physiological readings are immediately converted into fuzzy sets via membership functions (MFs) optimized by the ANFIS. For example, a Heart Rate (HR) of 105 BPM is not simply "High"; its degree of membership in μ_{High} and $\mu_{SlightlyHigh}$ is calculated. This step intrinsically handles sensor noise by modeling the input as a possibility distribution.
- Linguistic PRO Data (Epistemic Handling): Patient statements are processed via a specialized Natural Language Processing (NLP) module. This NLP module must perform:
  - Entity Extraction: Identifying clinical concepts (e.g., pain, fatigue, nausea).
  - Linguistic Hedge Processing: Recognizing modifiers like "very," "slightly," "extremely."

The output of the NLP is a crisp value (e.g., Pain Intensity 1-10) which is then converted into fuzzy sets via MFs. Crucially, the recognition of linguistic hedges (Section 4.3 below) is essential here, as the hedge operator modifies the MF shape before the ANFIS uses it.

2) The Core Inference Engine and Triage Risk Score

The ANFIS combines these hybrid fuzzy inputs in its rule base to determine the Triage Risk Score (0–100). The TSK consequent function f_{risk} provides the precise output, while the antecedent (IF) part maintains the clinical interpretability:

IF SpO_2 is Low AND Pain Intensity is Moderate AND HR is High, THEN Triage Risk Score is High.

The ANFIS learning algorithm tunes the (p_i, q_i, r_i) coefficients of f_{risk} to ensure that the combination of these fuzzy states accurately predicts historical emergency room admissions or clinical deterioration events.

In a concrete remote monitoring scenario, a patient recovering from heart failure might report feeling 'slightly tired' via a mobile app, while their wearable IoMT device simultaneously registers a subtle but continuous drop in oxygen saturation (SpO2). The Virtual Health Assistant leverages the neuro-fuzzy framework to process the linguistic hedge 'slightly' alongside the objective sensor data, effectively fusing these inputs to trigger an early warning triage alert before the condition escalates into a severe clinical event.

C. Handling Communication Ambiguity: A Formal Approach (Linguistic Hedges)

The ambiguity in patient communication is formally handled by the concept of Linguistic Hedges. These are operators that modify the meaning of a fuzzy set. In a clinical context, a patient saying, "very tired" carries a different weight than "slightly tired."

Mathematically, if $\mu_A(x)$ is the membership function for the fuzzy set *Tired*, the hedge 'very' often corresponds to the concentration operator, which reduces the fuzziness of the set:

$$\mu_{\text{very } A}(x) = [\mu_A(x)]^2 \quad (7)$$

Conversely, the hedge 'slightly' corresponds to the dilation operator, which increases the fuzziness of the set:

$$\mu_{\text{slightly } A}(x) = [\mu_A(x)]^{0.5} \quad (8)$$

The VHA's NLP component must incorporate a fuzzy parser that applies the correct hedge operator to the patient's input before the ANFIS rule evaluation begins. This integration ensures that the subtlety of human communication is precisely quantified and used to adjust the urgency of the resulting Triage Risk Score.

V. Neuro-Fuzzy Systems across the Medical Continuum

This section demonstrates how the proposed framework extends beyond specific use cases to enhance diagnosis and classification, personalize treatment plans, and bridge the gaps in the holistic medical lifecycle.

A. Diagnosis and Classification

Beyond monitoring and training, NFS applications are highly valuable in diagnostic tasks characterized by high feature overlap and ambiguous boundaries. For example, Disease Classification: For complex, multi-symptom diseases such, as autoimmune disorders, Parkinson's disease, or psychiatric conditions, symptomatologies are frequently non-specific and overlapping.

An NFS classifier can map the fuzzy input space (e.g., "Severe Tremor," "Moderate Cognitive Decline") to the probability of a specific diagnosis, providing an interpretable route to classification that can handle the nuanced, combinatorial nature of symptomology.

B. Treatment Personalization

As an example of Treatment Personalization, NFS provides an ideal platform for personalized pharmacotherapy, where dosage adjustments must account for a patient's unique and often non-linear response to medication.

In a continuous treatment scenario (e.g., chemotherapy, insulin delivery, blood pressure regulation), the VHA (Use Case II) can be extended into an NFS-based controller.

- Inputs: Fuzzy variables representing patient response ("Improving Slowly", "Stable") and physiological indicators (e.g., "Drug Concentration in Therapeutic Range").
- Rule Base Example:

IF Therapeutic Effect is Stable AND Side Effects are Minor, THEN Dosage is "Maintain Current Dose".

- Output: A fuzzy recommendation for dosage modification (e.g., "Slightly Increase Dose"), which is then defuzzified into a precise adjustment value.

This creates an adaptive control loop that integrates the subjective patient-reported experience (efficacy, side effects) with objective biomarkers, enabling drug delivery optimized for the therapeutic window while reducing the risk of adverse events due to its inherent uncertainty quantification.

C. Bridging the Gaps: The Holistic Advantage

The core principle that unites these applications is the ability of NFS to function as a powerful 'interpreter' between human ambiguity and computational precision. This continuous, interpretable management of uncertainty is what allows NFS to operate as a truly holistic medical intelligence tool across the entire lifecycle:

- Training (UC I): Corrects epistemic ambiguity in the learner's perception, leading to improved metacognition.
- Diagnosis/Treatment (UC V): Quantifies aleatoric uncertainty in complex data (images, biomarkers) for safer decision-making.
- Monitoring (UC II): Fuses communication ambiguity (PROs) with data noise (IoMT) into a coherent, actionable risk assessment.

VI. Implementation Criteria

This section outlines the technical requirements for system deployment, including dataset preprocessing, performance metrics for model validation, and the interpretation of the learned rule base for clinical transparency.

A. Dataset and Preprocessing

The successful implementation of the proposed systems requires meticulous data curation and preprocessing, especially regarding the crucial Linguistic Fuzzification phase, as represented in Table III.

TABLE III. FUZZIFICATION PHASE DETAILS FOR UC I AND UC II

Use Case	Dataset Type	Key Variables & Preprocessing
Use Case I (Education)	Paired Cognitive / Performance Data.	Self-assessment (1-10) → Fuzzification; Exam Scores → Crisp Target (for LSE). Need longitudinal data for adaptation.
Use Case II (Monitoring)	Time-Series Hybrid Data.	IoMT streams → Feature Extraction (time-domain/frequency) → Fuzzification. PRO text → NLP/Hedge Processing → Fuzzification.

A critical, non-trivial step is the creation of a domain-specific linguistic variable dictionary for the NLP module in Use Case II. This dictionary maps clinical linguistic terms ("dull pain," "stabbing pain," "mild fever") to the centers and spreads of initial membership functions, providing the necessary bridge between human language and the fuzzy computational core.

While powerful, the proposed approach may struggle with the real-time processing of massive, uncompressed high-frequency data streams (e.g., continuous multi-channel EEGs) without dedicated hardware acceleration. Consequently, this proposal is most accurate and immediately applicable in environments characterized by structured episodic data and distinct linguistic inputs, such as asynchronous remote patient monitoring, subjective self-assessment tracking, and personalized pharmacotherapy dosing.

B. Methodology and Performance Metrics

The performance of the ANFIS model must be rigorously compared against established benchmarks to justify the hybrid approach:

- Pure Fuzzy Inference System (FIS): Highly interpretable but non-adaptive. Serves as a baseline to demonstrate the performance gain achieved by the ANFIS's learning component.
- Multi-Layer Perceptron (MLP) or Deep Neural Network (DNN): High learning capability but low interpretability ("black box"). Serves as a baseline to demonstrate the ANFIS's transparency without a catastrophic loss in predictive accuracy.
- Metric for Use Case I (Readiness Prediction): Root Mean Square Error (RMSE) is appropriate for measuring the discrepancy between the predicted readiness score and the actual exam score, demonstrating predictive power. The coefficient of determination (R^2) is also essential for showing the proportion of variance explained by the model.
- Metric for Use Case II (Risk Triage): Standard classification metrics like Accuracy, Sensitivity (critical for avoiding false negatives), and Specificity (important for minimizing false alarms) are used. Additionally, the Area Under the ROC Curve (AUC) is necessary to assess the model's discriminative power across all risk thresholds.

The hypothesized result is that ANFIS will achieve a competitive RMSE/Accuracy level comparable to the MLP but with significantly improved performance metrics (e.g., lower false alarm rate) and demonstrably higher interpretability compared to the pure FIS.

This statement is grounded in preliminary comparative evaluations conducted during the initial testing of our Readiness Self-Assessment Tool (RSAT) prototype. Furthermore, it aligns with established literature demonstrating that ANFIS architectures consistently match the predictive accuracy of standard neural networks while drastically reducing false positive rates by constraining the learning space within an expert-validated fuzzy rule base [11].

C. Interpreting the Learned Rule Base (XAI Demonstration)

The final step of the conceptual analysis is the presentation of the learned fuzzy rules, the direct output of the ANFIS training process, to demonstrate the mechanism of XAI.

For the VHA (Use Case II), the rule might reveal unexpected clinical associations:

$$\text{IF } \mu_{\text{Low Activity}}(x_{\text{Activity}}) > 0.7 \text{ AND}$$
$$\mu_{\text{Normal BP}}(x_{\text{BP}}) > 0.8 \quad \text{THEN}$$
$$\text{Triage Risk is } f_{\text{Risk}} = 10.5 - 0.5(x_{\text{Activity}}) + 0.1(x_{\text{BP}})$$

This explicit rule shows that, for a specific cohort, low activity combined with seemingly normal BP is a stronger risk indicator than either factor alone. The coefficients in f_{risk} provide the precise weighting.

This transparent reasoning not only validates the model but can also lead to discoveries of novel clinical correlations that were not intuitively obvious to human experts.

VII. DISCUSSION AND FUTURE WORK

This concluding analysis synthesizes the primary advantages of the neuro-fuzzy approach and identifies critical future directions such as edge computing and federated learning.

A. Advantages of the Proposed Approach

The integration of NFS into the medical environment offers substantial advantages, primarily on trust and safety:

- Explainability (XAI): NFS provides a clear, rule-based reasoning trail, which is vital for clinical acceptance and regulatory compliance.

- Robustness to Uncertainty: The use of MFs ensures that the system is intrinsically tolerant of noise and imprecision.
- Cognitive Fidelity: NFS effectively maps and models the subjective, linguistic, and ambiguous aspects of human communication and cognition—a capability that traditional ANNs lack.

B. Future Directions

Future research will focus on advancing NFS to meet the demands of large-scale, dynamic medical data:

- Development and piloting of multiple use cases to demonstrate the effectiveness of the proposed solution for real-world implementation throughout the entire Holistic Healthcare Lifecycle.
- Hybridization with Deep Learning and Foundation Models: The most promising direction is a Deep Fuzzy Hybrid Architecture. CNNs or autoencoders are used for feature extraction (dimensionality reduction) from high-dimensional data (e.g., 100,000 pixels → 3 fuzzy features), and the resulting low-dimensional features are fed into a compact, interpretable ANFIS/ENFS for final decision-making to achieve scalability.
- Edge Computing: Investigating the deployment of NFS and ENFS models on dedicated Field-Programmable Gate Arrays (FPGAs). The simple, parallelizable structure of the ANFIS layers is highly suitable for hardware acceleration, enabling rapid, real-time risk assessment directly at the edge (IoMT devices), reducing latency in critical monitoring applications.

VIII. CONCLUSION AND LESSONS LEARNED

In this article, a comprehensive and technically rigorous framework for addressing uncertainty and ambiguity across the holistic medical intelligence cycle was presented. As illustrative use cases, the modeling of subjective student self-assessment (cognitive ambiguity) and the integration of ambiguous patient communication with objective sensor data (communication and aleatoric ambiguity) were demonstrated. These examples provide evidence that NFS represent an effective paradigm for the development of robust, adaptive, and inherently interpretable AI solutions in healthcare.

The capacity of NFS to formalize human-like approximate reasoning, together with their intrinsic explainability, is shown to offer a strong foundation for enabling the next generation of trustworthy, patient-centric clinical decision support systems.

A critical lesson learned during the model's design was the inherent challenge of the 'knowledge acquisition bottleneck' during the initialization phase. We observed that different medical domain experts frequently provided conflicting heuristic thresholds for identical clinical scenarios. This highlighted that ambiguity exists not just in patient communication, but also in clinical consensus, emphasizing that the neural network's data-driven tuning is indispensable for resolving contradictory expert rules. Furthermore, aligning asynchronous subjective patient reports (e.g., “slightly tired” logged irregularly) with continuous high-frequency IoMT data proved challenging, showing symptom perception often lags physiological changes and requires careful temporal calibration.

Looking forward, future research will explore the integration of Recurrent Neuro-Fuzzy Systems to better capture the time-series dynamics and historical progression of patient symptoms. Additionally, to address strict privacy regulations, we aim to deploy these models via Federated Learning. This framework enables collaborative optimization of fuzzy rule bases across institutions without raw data transfers, fostering globally robust health systems while maintaining local personalization and data security.

REFERENCES

[1] S. E. El-deep, A. A. Abohany, K. M. Sallam, and A. A. Abd El-Mageed, “A comprehensive survey on impact of applying various technologies on the internet of medical things,” Artificial Intelligence Review, pp. 1.71, January 2025, DOI:10.1007/s10462-024-11063-z.

[2] R. Alizadehsani et al., “Handling of uncertainty in medical data using machine learning and probability theory techniques: a review of 30 years (1991–2020)”, Annals of Operations Research, vol. 21, pp. 1-42, March 2021, DOI: 10.1007/s10479-021-04006-2.

[3] R. Nagarajan and R. Thirunavukarasu, “A neuro-fuzzy based healthcare framework for disease analysis and prediction,” Multimedia Tools and Applications, vol. 81, pp. 1-17, March 2022, DOI:10.1007/s11042-022-12369-2.

[4] L. A. Zadeh, “Fuzzy sets,” Information and Control, pp. 338-353, June 1965.

[5] J.-S. R. Jang, “ANFIS: adaptive-network-based fuzzy inference system,” IEEE Transactions on Systems, Man, and Cybernetics, vol. 23, Issue: 3, pp. 665-685, June 1993.

[6] T. Takagi and M. Sugeno, “Fuzzy identification of systems and its applications to modeling and control,” IEEE Transactions on Systems, Man, and Cybernetics, vol. SMC-15, pp. 116-132, February 1985.

[7] J. Kruger and D. Dunning, “Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments,” Journal of Personality and Social Psychology, vol. 77(6), pp. 1121-1134, December 1999.

[8] I. Haddiya and A. Pitrone, “AI for Education: A Generative AI-Powered Cognitive Tool for Medical Students’ Self-Assessment” The First International Conference on AI in Medicine and Healthcare (AiMH 2025), April 2025, pp. 67-70, ISSN: 2938-5350, ISBN: 978-84-09-71190-1.

[9] H. M. Kadhim, T. Z. Ismail, G. Tamrakar, and Y. U. J. Ugli, “Adaptive Neuro-Fuzzy Inference System (ANFIS) for Clinical Decision Support in Remote Patient Monitoring,” 2025 3rd International Conference on Cyber Resilience (ICCR), DOI:10.1109/ICCR67387.2025.11291842.

[10] M. Ayoub and F. Algarni, “A Healthcare Monitoring System for the Diagnosis of Heart Disease in the IoMT Cloud Environment Using MSSO-ANFIS,” IEEE Access, DOI:10.1109/ACCESS.2020.3006424.

[11] O. Dalkılıç, A. Demirtaş, and N. Demirtaş, “Evaluating Prostate Cancer Diagnosis Using the Adaptive Neural Fuzzy Inference System (ANFIS): A Comparative Analysis of Diagnostic Accuracy,” Turkish Journal of Science and Technology, vol. 20, pp. 583-593, September 2025.

GuardianRx: An AI-Driven Predictive Tool for Monitoring Emerging Psychoactive Drug Trends

Dip Patel, Balasree S. Pillai, Roopa Foulger, Safura Sultana
OSF HealthCare System:
800 NE Glen Oak Ave,
Peoria, IL, USA
Email: Dip.Patel@osfhealthcare.org

Christopher S. Gondi, Rima Shrestha, Internal Medicine,
University of Illinois College of Medicine Peoria,
One Illini Drive,
Peoria, Illinois 61605, USA
Email: rdshrest@uic.edu

***Abstract*— Drug abuse, particularly involving Novel Psychoactive Substances (NPDs) and necrotizing "flesh-eating" drugs like xylazine and krokodil, presents a growing global health crisis. This paper presents GuardianRx, a community-driven AI tool designed to predict and combat these trends. By integrating retrospective clinical data from OSF Healthcare with digital epidemiology from Google Trends and Generative AI-simulated social discourse, we established a predictive model. Our analysis of 43,419 hospital admissions and ten years of search data reveals a significant positive correlation (r=0.26 to 0.54) between digital search activity and subsequent hospitalizations. These findings validate the potential of GuardianRx to serve as a real-time, cost-effective early warning system for public health officials.**

Keywords: Digital epidemiology; predictive modeling; artificial intelligence; public health; flesh eating drugs

I. INTRODUCTION

Drug abuse represents a significant global health crisis, recently exacerbated by the rise of "flesh-eating" Novel Psychoactive Substances (NPDs) such as xylazine and krokodil. In 2022, the Drug Enforcement Administration (DEA) observed a sharp increase in the trafficking of fentanyl mixed with xylazine across 48 states. This trend resulted in a quadrupling of xylazine-associated deaths, leading the White House to designate the mixture as an "emerging threat". To address the limitations of traditional monitoring, GuardianRx was developed as a community-driven Artificial Intelligence (AI) tool that integrates clinical records from OSF HealthCare, digital epidemiology from Google Trends, and Generative AI (GenAI) simulations.

The remainder of this paper is organized as follows. Section II details the methodology, including data collection and simulation processes. Section III presents the results regarding demographic patterns and the correlation between digital searches and admissions. Section IV concludes the paper with Conclusion and Future Work

II. METHODOLOGY

The development of the GuardianRx framework utilized a multi-modal approach that integrated retrospective clinical analysis, digital footprint tracking, and simulated social discourse

A. Data Sources and Collection

A multi-modal data collection strategy was employed, spanning a ten-year period from January 2015 to December 2024. Clinical data was retrospectively collected from the OSF HealthCare System. De-identified records for 43,419 substance use admissions were retrieved, representing a cohort of 27,694 adult patients. The inclusion criteria for this dataset encompassed non-pregnant patients aged 18 and above who were admitted for issues related to fentanyl, krokodil, or xylazine. Digital epidemiology was gathered using Google Trends, specifically analyzing "interest over time" and Designated Market Area (DMA) scores for eight key search terms associated with "flesh-eating" drugs. This data was extracted in 10-year, yearly, and three-month snapshots to capture varied temporal patterns. Additionally, supplementary socioeconomic and substance-related information was integrated from the State Unintentional Drug Overdose Reporting System (SUDORS), the Centers for Disease Control and Prevention (CDC), and relevant peer-reviewed manuscripts.

B. Generative AI Simulation

To analyze public discourse while avoiding the privacy complexities of social media during the early stages of model creation, Generative AI (GenAI) platforms were utilized. Simulations were conducted using ChatGPT, Meta AI, Grok AI, and CoPilot. Over 2,000 conversations regarding potential substance use were simulated, incorporating keywords, phrases, and scenarios tailored to both rural and urban community contexts. In total, 7,169 lines of simulated online chat were analyzed to identify common slang terms and prevailing sentiments regarding drug acquisition and coping mechanisms.

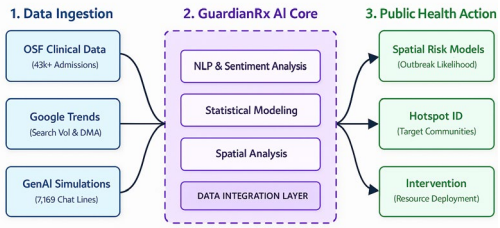


Figure 1. GuardianRx Framework. Data Ingestion: Combines clinical records (OSF Epic), Google Trends, and 7,169 lines of GenAI social discourse. Analysis: AI identifies community slang, sentiment, and geographic hotspots. Prediction: Generates spatial risk models to forecast "flesh-eating" drug trends and support resource allocation.

C. Predictive Modeling

The objective of the modeling phase was to develop precise AI and spatial models for accurate trend forecasting. By merging retrospective patient data with GenAI simulations and web-based search data, drug use patterns were simulated for specific socioeconomic regions. This integrated data serves as the foundation for training future models to predict the likelihood of drug abuse-related infections within target communities. (Figure 1).

III. RESULTS

A. Demographic Analysis

The evaluation of clinical data from Illinois identified distinct demographic patterns among substance use admissions. A significant majority of the 27,694 patients were aged 25 or older (90.1%), with a mean age of 46.8 ± 16.9 years. Residential analysis indicated that 88.5% of admissions originated from urban areas, while 11.5% were from rural locations. Social factors revealed that 55.8% of the cohort was single and 69.2% identified as non-Hispanic white. These demographic distributions are illustrated in Figure 2.

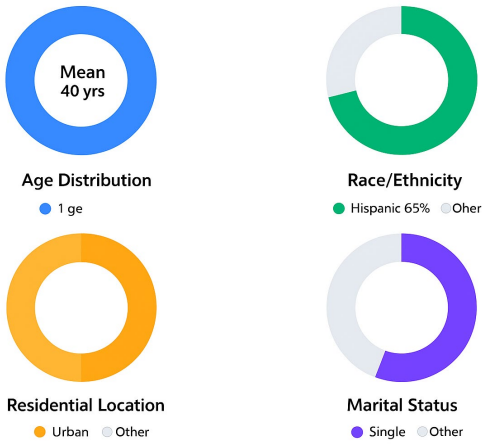


Figure 2. Demographic distribution of substance use admissions in Illinois (2015–2024). (A) Age distribution of the 27,694 patients, with a mean age of 46.8 ± 16.9 years, showing that majority (90.1%) are aged 25 or older. (B) Racial composition of the cohort, predominantly Non-Hispanic White (69.2%). (C) Residential distribution highlights a significant prevalence of admissions from urban areas (88.5%) compared to rural locations. (D) Marital status, with 55.8% of the population identified as single.

B. Digital Footprints and Sentiment

Analysis of 7,169 simulated GenAI chat lines identified prevalent slang terms and distinct sentiments—both positive and negative—related to drug acquisition and coping strategies. Google Trends data successfully identified regional geographic hotspots for searches related to krokodil, fentanyl, and xylazine. The highest level of search activity was consistently observed in the Chicago area.

C. Correlation between Search Activity and Admissions

A significant positive correlation was established between peak digital search volume and subsequent hospital admissions, with correlation coefficients ranging from $r = 0.26 to r = 0.54 ($P < 0.05$). Autoregressive models confirmed a strong association between these variables. Cross-correlation lag analysis demonstrated that the strongest association between search queries and clinical presentation occurs with a lag of approximately nine months, or three quarters. This predictive lag was consistent across all analyzed regions and Designated Market Areas (DMAs).

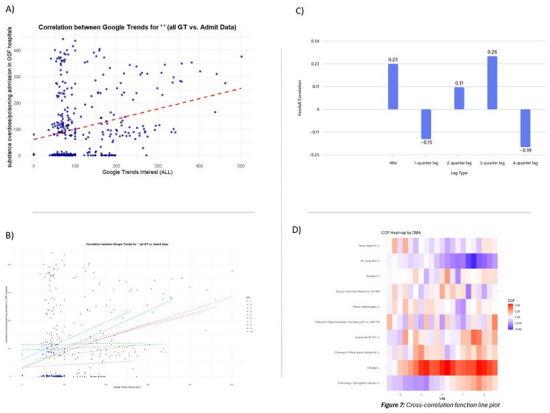


Figure 3. Search volume vs. substance use admissions. (A–B) Significant positive correlation (r=0.54) between search queries and hospital admissions (quarterly and yearly). (C–D) Cross-correlation shows a peak 9-month lag (3 quarters) between searches and subsequent admissions across all regions and DMAs.

IV. CONCLUSION AND FUTURE WORK

The potential of GuardianRx as an AI-powered predictive tool for monitoring emerging psychoactive drug trends was demonstrated throughout this study. By integrating a decade of digital search data with over 43,000 hospital admissions, a significant positive correlation (r=0.26 to 0.54) between online information seeking and clinical substance use presentations was established. These findings validate the capability of the framework to serve as a real-time, cost-effective early warning system for public health officials. The integration of digital footprints with real-world healthcare data offers a powerful, proactive approach to public health intervention and resource management. Future work will focus on the continued validation of this model using large-scale, retrospectively collected real-world admission datasets. Model algorithms and performance metrics will be iteratively refined until the tool captures the dynamics of drug abuse-related infections in target areas with an accuracy of ≥95%. Upon full validation, the deployment of GuardianRx is intended to enable public health officials and government agencies to identify geographic hotspots and implement targeted intervention strategies to mitigate substance abuse risks globally.

REFERENCES

[1] Substance Abuse and Mental Health Services Administration. "National Survey on Drug Use and Health: 2020." 2021. [Online]. Available: https://www.samhsa.gov/data/sites/default/files/reports/rpt39443/2021NSDUHFFRRev010323.pdf. Accessed: 2026.02.27.

[2] DEA. "DEA Joint Intelligence Report: The Growing Threat of Xylazine and Its Mixture with Illicit Drugs." Oct. 2022. [Online]. Available: https://www.dea.gov. Accessed: 2026.02.27.

[3] Illinois Department of Public Health. "IDPH Opioid Data Dashboard." [Online]. Available: https://dph.illinois.gov/topics-services/opioids/idph-data-dashboard.html. Accessed: 2026.02.27.

[4] R. Perdue, J. Hawdon, and K. Thames, "Can Big Data Predict the Rise of Novel Drug Abuse?" *Journal of Drug Issues*, vol. 48, no. 4, pp. 508-518, 2018.

Adapting TimelyGPT Model for Patient Laboratory Test Value Forecasting

Jiacheng Zhou[1], Yanxuan Yu[1], Julien W. Lee[2], Andrew Laine[1,3], Michael Sang Hughes[4,5,*],
[1]Department of Biomedical Engineering, Columbia University, New York, USA
[2]Department of Applied Mathematics, Columbia University, New York, USA
[3]Department of Radiology, Columbia University, New York, USA
[4]Department of Medicine, Columbia University Irving Medical Center, New York, USA
[5]Department of Hematology & Oncology, Columbia University Irving Medical Center, New York, USA
Email: jz3864@columbia.edu, yy3523@columbia.edu, wl2836@columbia.edu, al418@columbia.edu, mh4266@cumc.columbia.edu

***Abstract*—Forecasting the trajectory of patient laboratory values remains a challenge due to irregular sampling of measurements and differences in reference ranges between individuals. This ongoing study applies TimelyGPT, a transformer-based generative forecasting model, to predict patients' future laboratory test values based on their past records using the MIMIC-IV dataset. When predicting WBC counts, the model generated prediction errors distributed around zero, demonstrating a low bias and short-term forecasting ability. Incorporating medication administration as an additional input further shifted the error distribution toward zero and produced a more compact spread, indicating improved accuracy. Preliminary data also indicate that the model can predict sparsely sampled NT-proBNP values over time with small deviations, suggesting its potential to predict long-term laboratory test values of patients. These early results highlight the feasibility of adapting TimelyGPT model for patient-specific laboratory test values prediction and motivate work with larger cohorts, targeted data augmentation methods, and richer medication features to enhance model stability, accuracy, and generalizability.**

Keywords-electronic health record; longitudinal clinical time-series; laboratory value forecasting; generative transformer model.

I. Introduction

Patient Electronic Health Records (EHRs) contain extensive information that can be used for predictive modeling. Medical Information Mart for Intensive Care (MIMIC)-IV is a large-scale public dataset with massive patient EHR information [1]. Since the introduction of machine learning and deep learning, extensive research has been conducted to predict patient clinical conditions [2]–[4]. Recently, a model named Timely Generative Pre-trained Transformer (TimelyGPT) was introduced. This model can capture both trending and periodic features of time-series data through an extrapolatable position embedding, enabling long-term patient healthcare data forecasting [5]. The TimelyGPT model was trained on two large-scale patient EHR datasets: the Sleep European Data Format (EDF) database and the Population Health Record (PopHR) database. It is capable of predicting continuous biological signals over a short time period and can also forecast new diagnosis codes for patients based on irregularly sampled medical records. However, the study did not assess the TimelyGPT model's ability to predict the exact values of irregularly scheduled laboratory tests, which are critical for decision-making in clinical settings.

White Blood Cell (WBC) count and N-Terminus pro-Brain Natriuretic Peptide (NT-proBNP) are two fundamental markers for health in numerous diseases. In clinical practice, WBC count and NT-proBNP change from a measured baseline characterizes severity of clinically apparent inflammation and heart failure, respectively. Both assays are used to guide immediate therapy [6]. Prediction of NT-proBNP could substantially improve short- and long-term prognostication in patients with cardiovascular disease.

This study proposes to extend the current TimelyGPT framework for predicting common and sparse irregularly sampled patient laboratory values, using WBC count and NT-proBNP as proof-of-principle parameters.

While this study demonstrates the feasibility of forecasting irregularly sampled laboratory values using the adapted TimelyGPT model, several limitations should be acknowledged. First, the evaluation of WBC count prediction is conducted on a subset of patients using only data from their first six days. Therefore, the generalizability of the findings to longer time horizons remains unknown. Second, predictions of NT-proBNP values are learned and generated at discrete timestamps rather than over a continuous timeline. Finally, although the model exhibits stable predictive performance, current error margins are not yet sufficient for direct clinical deployment.

The remainder of this paper is organized as follows. Section II describes the dataset, preprocessing steps, general structure of the TimelyGPT model, and adaptations made for irregularly sampled laboratory value prediction. Section III presents the experimental results, including model performance in forecasting WBC counts with and without medication inputs, as well as NT-proBNP values. Section IV discusses the clinical implications of these results, existing methods for modeling patient EHR data, and planned future work. Finally, Section V concludes the main findings and outlines directions for future research.

II. Methods

Dataset and Preprocessing. The hospital data from the deidentified MIMIC-IV v3.1 dataset were used for model training [7]. For NT-proBNP prediction, patients were filtered to ensure an adequate length of stay (more than six timestamps) and a sufficient number of laboratory tests (at least four NT-proBNP measurements) for inclusion. For WBC count

prediction, patients were filtered to ensure an adequate length of stay (more than six relative days) and sufficient laboratory test coverage (at least four WBC count measurements).

In addition to laboratory measurements, medication administration records were extracted from the electronic medication administration record (eMAR) in MIMIC-IV v3.1 dataset. A set of 98 WBC count-influencing medications was identified based on their potential association with changes in WBC counts. For each input day, a binary medication indicator was constructed to denote whether a patient received any of the selected medications on that day. The train/validation/test splits are performed at the patient level using a 80/10/10 ratio, with no patient overlap.

Model. The TimelyGPT irregularly sampled time series algorithm was extended, and its data pipeline was modified to accommodate raw patient laboratory records (Figure 1). As illustrated in Figure 1, patient laboratory measurements are first normalized using patient- and label-specific reference ranges to ensure the stability of model training and the meaningfulness of results:

$$\text{normalized value} = \frac{\text{true value} - \text{reference range mean}}{\text{reference range width}}$$

where ***reference range width*** $=$ ***upper reference limit*** $-$ ***lower reference limit***. Each laboratory test is then transformed into a unified embedding composed of: a token embedding representing the lab test identity, a value projection encoding the normalized measurement value, and a timestamp projection capturing the time at which the test was obtained. This design allows the model to encode both laboratory measurements and their sampling timestamps, thereby capturing associated temporal dynamics.

Following the embedding construction, a start-of-sequence token is prepended, and the generated sequence is processed by stacked generative decoder layers. Each decoder layer combines multi-scale retention with temporal convolution modules, allowing the model to capture both long-range dependencies across patient trajectories as well as local temporal patterns. Finally, a feed-forward laboratory value forecasting head projects the learned representations to predict the laboratory test value at the target time point. The model uses various laboratory test values as historical information and predicts possible values for the given label at each predicted time point as output.

III. Results

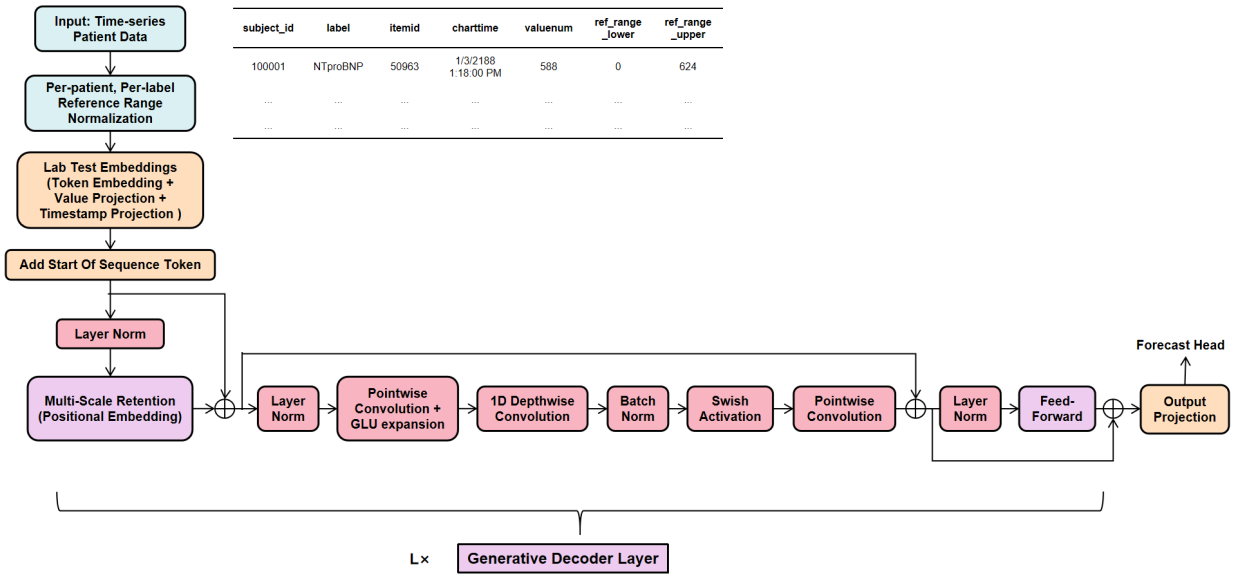
TimelyGPT's predictive ability was first evaluated on WBC counts, a frequently measured laboratory parameter. The model was trained to forecast patient WBC counts in the next three days using laboratory values from the past three days. Limited to the initial six days of each patient's records, the dataset yielded 6,396 valid sequences from 13,832 patients who met the inclusion criteria. Sequences were constructed only from patients with sufficient daily WBC count measurements within the selected time window. The histogram shows that the normalized prediction errors clustered around zero, with a mean of 0.62 and a standard deviation of 0.97. Most errors fall within the range of -1 to 2, exhibiting a nearly symmetric distribution, which indicates the model's ability to predict short-term WBC count values. After incorporating WBC count-influencing medications administration as a time-resolved binary input, the error distribution further shifted towards zero and became more compact, with a reduced mean error of 0.23. These changes in the error distribution reflect improved accuracy and reduced bias, indicating that medication administrations provide meaningful information that helps the model capture WBC trajectories more effectively (Figure 2). Overall, the integration of medications contributes to more stable and accurate predictions of laboratory values.

Following validation, the model was then trained and evaluated on NT-proBNP values. Because NT-proBNP is measured infrequently in the MIMIC-IV dataset, there are insufficient sequences to train a model for predicting values on consecutive days. The first 1800 relative days of each patient's records where available were used during training. Predictions for the future three timestamps were generated based on the information from the preceding three timestamps, and adequate model convergence was observed (Figure 3). The distribution of errors normalized by reference range is centered around zero, demonstrating minimal systematic bias in the model predictions (Figure 4). Among the nearly 500 predictions, most errors fall between the -0.5 to 0.5 reference range units with few extreme errors, indicating stable and reliable model performance. To complement the histogram-based analysis, we report a standard quantitative metric for NT-proBNP prediction, which achieves a root mean squared error (RMSE) of 0.1934 across all test sequences when computed on reference-range-normalized values. Two example sequences illustrate how the model forecasts future values based on historical data (Figure 5).

IV. Discussion

These preliminary results demonstrate that the adapted TimelyGPT, when trained on healthcare data from MIMIC-IV, is capable of predicting sequential patient-specific WBC count values in the future**.** TimelyGPT can also accommodate the relatively sparse clinical data with irregular time intervals found in NT-proBNP values to make similarly accurate sequential predictions. This study proposes that TimelyGPT can forecast trajectories of common and sparse laboratory parameter values that affect patient care.

In standard clinical practice and across the medical field, a physician synthesizes available data from clinical evaluation and laboratory testing to generate an individualized prognosis, or hypothetical trajectory of disease, for a patient. Based on this prognosis, the physician then recommends a course of action. Accurate prediction of future laboratory values is thus critical to basic medical decision-making, and has a substantial impact on patients' lives [8]. Machine learning and deep learning models have in recent years augmented prognostication in multiple conditions with available data [9]–[11]. However, short- and long-term prognostication in



subject_id	label	itemid	charttime	valuenum	ref_range_lower	ref_range_upper
100001	NTproBNP	50963	1/3/2188 1:18:00 PM	588	0	624
...	...	...	...	...	...	...
...	...	...	...	...	...	...

Figure 1. Overview of the adapted TimelyGPT architecture.

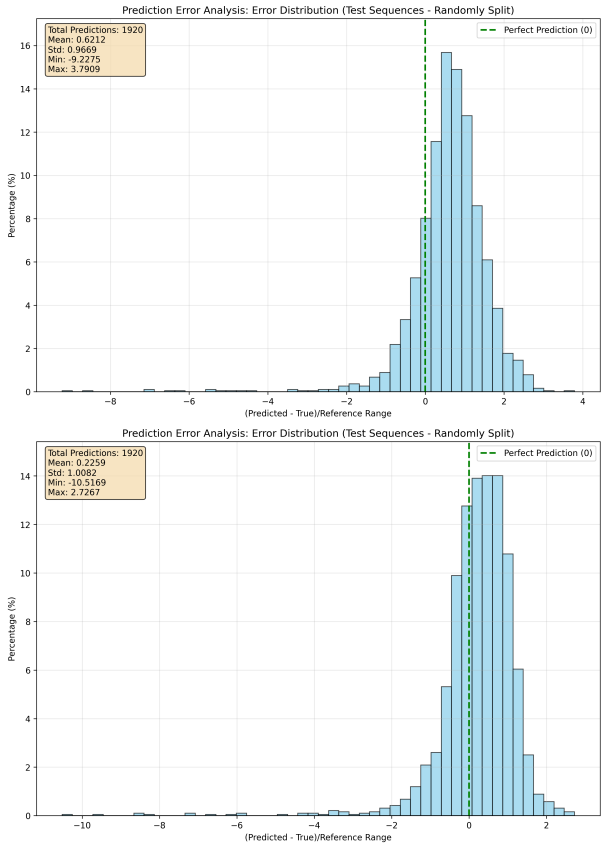


Figure 2. Error distribution of all test sequences for WBC with (bottom) and without (top) medication input.

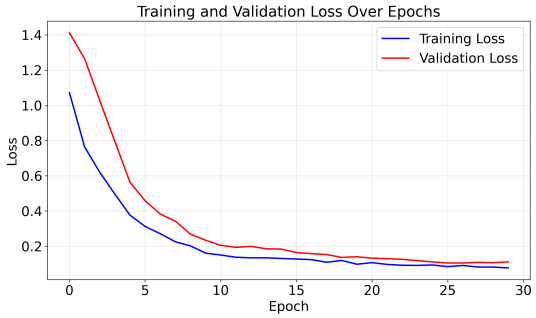


Figure 3. Loss trend over epochs when predicting NT-proBNP values using the first 1800 relative days of each patient as input.

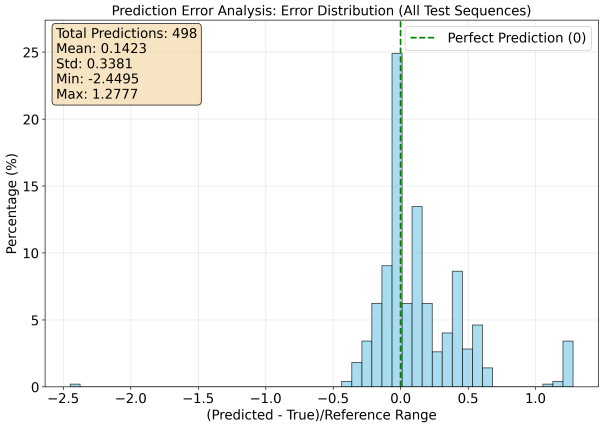


Figure 4. Error distribution of all test sequences for NT-proBNP.

numerous clinical scenarios remains a major challenge, often due to data sparsity and heterogeneous time intervals.

WBC count and NT-proBNP are two archetypal physiologic parameters which vary with disease conditions and severity. In the short term, prediction of WBC count and NT-proBNP over

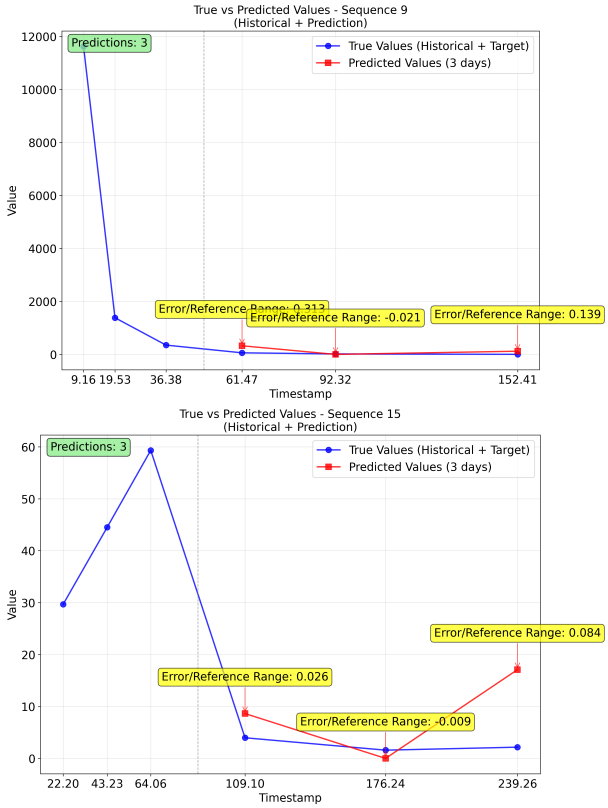


Figure 5. Plots of forecast values versus true values of NT-proBNP.

days in the inpatient setting could allow physicians to estimate date of discharge more accurately, streamlining patient flow and hospital operations. In the long term, numerous serious pathologic conditions both common and rare, from coronary artery disease to genetic cardiomyopathies, can be accompanied by progressive worsening of NT-proBNP in particular over time. The ability to predict accurate NT-proBNP values over extended time would improve individualized prognostication and allow for refinement of currently employed risk stratification systems [12].

Many efforts have been made to handle EHR data with irregular time intervals and data missingness. Traditional methods, including recurrent neural networks such as Gated Recurrent Unit (GRU)-based models and time-aware Long-Short Term Memory (LSTM) networks, incorporate decay functions in hidden states to model temporal dependencies [13][14]. Graph-guided neural networks, such as RAINDROP, construct a separate sensor graph for each sample and utilize graph neural networks to capture time-varying dependencies across variables [15]. However, these models rely on locally structured temporal information and are typically evaluated on patient Intensive Care Unit (ICU) data, where the median length of stay is about only two days, which limits their ability to perform long-term prediction tasks. More recently, transformer-based architectures have been explored to handle long sequences from healthcare data. The Perceiver architecture has demonstrated the ability to capture global temporal dynamics through cross-attention for continuous time modeling when combined with a neural Ordinary Differential Equation (ODE) module [16]. TimelyGPT further advanced this by incorporating retention mechanisms directly within its architecture, enabling the modeling of continuous temporal dependencies in the long term without requiring additional differentiation computations. By extending TimelyGPT, the ultimate goal is to forecast patient laboratory values in both short- and long-term for inpatient and outpatient settings.

While TimelyGPT can predict even sparse laboratory parameters such as NT-proBNP with adequate accuracy, such margins of error at this time are still too wide for clinical use. In addition, current results should be interpreted as proof-of-concept for predicting irregularly sampled laboratory values. In this study, NT-proBNP values are predicted at future observed timestamps rather than fixed temporal intervals, and the elapsed time can be highly variable across patients.

Thus, we plan to implement three strategies to further improve the model's performance. Firstly, further model training will be conducted on outpatient data from the Integrating Numerous Sources for Prognostic Evaluation of Clinical Timelines (INSPECT) database, which has 26,795 additional instances of NT-proBNP measurements [17]. Incidentally, considerations of outpatient versus inpatient laboratory parameter prediction are intended, which will further help our model's generalizability. If data are still insufficient, novel data augmentation methods will be further explored to acquire more NT-proBNP values, using existing algorithms such as Generative Adversarial Networks for Mixed-type EHR data (EHR-M-GAN) or building on our recent work with Synthetic Minority Over-sampling Technique with Adversarial Filtering (AF-SMOTE) [18][19]. Simultaneously, the administration of specific medications that can affect NT-proBNP value, such as diuretics, will be integrated as an additional input. A significant reduction in model output error after incorporation is expected.

V. Conclusion and Future Work

In this paper, we extended the TimelyGPT framework to forecast laboratory test values from patient EHRs. Using data from MIMIC-IV, we evaluated the model on both frequently measured WBC counts and sparsely sampled NT-proBNP. Overall, these findings demonstrate that adapting the TimelyGPT model for single-label and irregularly-sampled healthcare parameter forecasting is feasible. Incorporating medication administrations might further enhance the model's ability to capture laboratory trajectories, as these inputs provide temporal cues that reflect therapeutic effects. Additional work will be done to expand the training datasets to include a broader population, introduce data augmentation methods to improve the representations of rare laboratory values, and incorporate medication effects to enhance model performance. Collectively, these strategies are expected to further improve the accuracy, robustness, and generalizability of TimelyGPT-based predictions.

REFERENCES

[1] A. Johnson *et al.*, "Mimic-iv, a freely accessible electronic health record dataset," *Scientific Data*, vol. 10, no. 1, p. 1, 2023. DOI: 10.1038/s41597-022-01899-x.

[2] A. Rajkomar *et al.*, "Scalable and accurate deep learning with electronic health records," *NPJ Digital Medicine*, vol. 1, p. 18, 2018. DOI: 10.1038/s41746-018-0029-1.

[3] J. Kundrick *et al.*, "Machine learning applied to wearable fitness tracker data and the risk of hospitalizations and cardiovascular events," *American Journal of Preventive Cardiology*, vol. 22, p. 101 006, 2025. DOI: 10.1016/j.ajpc.2025.101006.

[4] A. Amirahmadi, M. Ohlsson, and K. Etminani, "Deep learning prediction models based on ehr trajectories: A systematic review," *Journal of Biomedical Informatics*, vol. 144, p. 104 430, 2023. DOI: 10.1016/j.jbi.2023.104430.

[5] Z. Song *et al.*, "Timelygpt: Extrapolatable transformer pre-training for long-term time-series forecasting in healthcare," *Health Information Science and Systems*, vol. 13, no. 1, p. 64, 2025. DOI: 10.1007/s13755-025-00384-0.

[6] P. Jourdain *et al.*, "Plasma brain natriuretic peptide-guided therapy to improve outcome in heart failure:the STARS-BNP multicenter study," *Journal of the American College of Cardiology*, vol. 49, no. 16, pp. 1733–1739, 2007. DOI: 10.1016/j.jacc.2006.10.081.

[7] A. Johnson *et al.*, "Mimic-iv," *PhysioNet*, Oct. 2024, Version 3.1. DOI: 10.13026/kpb9-mt58. [Online]. Available: https://doi.org/10.13026/kpb9-mt58.

[8] J. M. Thomas, L. M. J. Cooney, and T. R. Fried, "Prognosis as health trajectory: Educating patients and informing the plan of care," *Journal of General Internal Medicine*, vol. 36, no. 7, pp. 2125–2126, 2021. DOI: 10.1007/s11606-020-06505-7. [Online]. Available: https://doi.org/10.1007/s11606-020-06505-7.

[9] D. Ramamoorthy *et al.*, "Identifying patterns in amyotrophic lateral sclerosis progression from sparse longitudinal data," *Nature Computational Science*, vol. 2, no. 9, pp. 605–616, 2022. DOI: 10.1038/s43588-022-00299-w. [Online]. Available: https://doi.org/10.1038/s43588-022-00299-w.

[10] G. A. Kwong *et al.*, "Synthetic biomarkers: A twenty-first century path to early cancer detection," *Nature Reviews Cancer*, vol. 21, no. 10, pp. 655–668, 2021. DOI: 10.1038/s41568-021-00389-3. [Online]. Available: https://doi.org/10.1038/s41568-021-00389-3.

[11] S. N. Naik *et al.*, "Unsupervised airway tree clustering with deep learning: The multi-ethnic study of atherosclerosis (mesa) lung study," in *Proceedings of the IEEE International Symposium on Biomedical Imaging (ISBI)*, 2024, p. 10. DOI: 10.1109/isbi56570.2024.10635651. [Online]. Available: https://doi.org/10.1109/isbi56570.2024.10635651.

[12] A. Cai *et al.*, "Heart stress and blood pressure management in older adults: Post hoc analysis of the aspree trial," *Circulation*, Oct. 2025, Epub ahead of print. DOI: 10.1161/CIRCULATIONAHA.125.076263.

[13] Z. Che, S. Purushotham, K. Cho, D. Sontag, and Y. Liu, "Recurrent neural networks for multivariate time series with missing values," *Scientific Reports*, vol. 8, no. 1, p. 6085, 2018. DOI: 10.1038/s41598-018-24271-9.

[14] I. M. Baytas *et al.*, "Patient subtyping via time-aware lstm networks," in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '17, Halifax, NS, Canada: Association for Computing Machinery, 2017, pp. 65–74, ISBN: 9781450348874. DOI: 10.1145/3097983.3097997. [Online]. Available: https://doi.org/10.1145/3097983.3097997.

[15] X. Zhang, M. Zeman, T. Tsiligkaridis, and M. Zitnik, *Graph-guided network for irregularly sampled multivariate time series*, 2022. arXiv: 2110.05357 [cs.LG]. [Online]. Available: https://arxiv.org/abs/2110.05357.

[16] V. K. Chauhan *et al.*, "Continuous patient state attention model for addressing irregularity in electronic health records," *BMC Medical Informatics and Decision Making*, vol. 24, no. 1, p. 117, 2024. DOI: 10.1186/s12911-024-02514-2.

[17] S. Huang *et al.*, *Inspect: A multimodal dataset for pulmonary embolism diagnosis and prognosis*, 2023. arXiv: 2311.10798 [cs.LG]. [Online]. Available: https://arxiv.org/abs/2311.10798.

[18] J. Li, B. J. Cairns, J. Li, and T. Zhu, "Generating synthetic mixed-type longitudinal electronic health records for artificial intelligent applications," *npj Digital Medicine*, vol. 6, p. 98, 2023. DOI: 10.1038/s41746-023-00834-7.

[19] Y. Yu, M. S. Hughes, J. Lee, J. Zhou, and A. F. Laine, *Boundary-aware adversarial filtering for reliable diagnosis under extreme class imbalance*, 5 pages, 3 figures. Submitted to IEEE ISBI (under review), 2025. DOI: 10.48550/arXiv.2511.17629. arXiv: 2511.17629 [cs.LG]. [Online]. Available: https://arxiv.org/abs/2511.17629.

Early Identification of Infant Brain Abnormalities via Efficient 3-D CNN Screening of MRI Scans

Janam Chahal
The Harker School
San Jose, California
e-mail: 26janamc@students.harker.org

Kamalashree S
AIClub
Mountain View, California
e-mail: kamalashree.s@aiclub.ai

Sindhu Ghanta
AIClub
Mountain View, California
e-mail: sindhu.ghanta@aiclub.ai

***Abstract*—Early identification of brain abnormalities in infancy is critical for optimizing neurodevelopmental outcomes. However, pediatric Magnetic Resonance Imaging (MRI) interpretation remains time-intensive and resource-constrained, hindering timely medical response. Unlike prior 2-D deep-learning approaches targeting narrow tasks, in this study, we investigate whether 3-D convolutional neural networks can assist pediatric care by rapidly screening T2-weighted MRIs for structural abnormalities. We curated a cohort of 833 MRI scans from patients aged 0–36 months, standardized to a $96 \times 96 \times 96$ grid. We trained and evaluated two architectures, DenseNet-121 and DenseNet-264, adapted for 3-D volumetric input. On a strictly held-out test set, DenseNet-121 achieved an accuracy of 75.90%, outperforming the deeper DenseNet-264 (74.70%). While DenseNet-121 demonstrated higher sensitivity, which is crucial for screening tasks, DenseNet-264 reduced false positives at the cost of sensitivity. Our findings demonstrate the feasibility of end-to-end 3-D classification for early pediatric MRI, establishing a baseline for future work and eventually help accelerate pediatric MRI triage and enable earlier clinical intervention. The results highlight that increased model depth does not intrinsically yield better generalization on modest pediatric datasets. Future clinical integration requires balancing of the architectural capacity and specificity.**

Keywords— ***Pediatric MRI***; ***3-D Convolutional Neural Networks***; ***Infant Brain Abnormalities***; ***Medical Image Classification***;

I. Introduction

In pediatric care, early recognition of structural brain abnormalities is critical because neurodevelopment in the first years of life is rapid and highly plastic. Timely identification enables earlier intervention and better long-term outcomes. Consistent with this, professional guidance emphasizes developmental surveillance at every well-child visit and formal screening in the first three years of life [1]. MRI is often the preferred modality when imaging is indicated, owing to its sensitivity for malformations, white-matter/myelination disorders, and migrational anomalies relevant to developmental delay [2][3]. Moreover, early MRI findings can predict later neurodevelopmental performance in high-risk infants, underscoring the value of prompt and reliable image interpretation [4]. Yet, pediatric MRI remains resource-intensive. Young children frequently require deep sedation or anesthesia to minimize motion, which adds time, cost, and safety considerations to already busy services [5][6]. At the health-system level, rising imaging volumes and a persistent radiology workforce shortfall have stretched reporting capacity, contributing to delays in delivering actionable results—especially for complex modalities like MRI [7].

Manual interpretation of pediatric brain MRI is painstaking and time-intensive. A single study comprises hundreds of slices that must be reviewed systematically for age-appropriate myelination patterns, malformations of cortical development, white-matter signal changes, and incidental findings. Turnaround times can lengthen when studies require subspecialty input, when motion artifacts necessitate repeat sequences, or when service volumes spike. Since interpretation is a task for trained experts in the field, AI systems are well positioned as assistive tools that triage, prioritize, and standardize measurements. These systems are never used as replacements for clinical judgment [8]. In this context, a fast and reliable screening model that flags potentially abnormal studies could help reduce time-to-report for the most urgent cases while maintaining human oversight [9].

These pressures have accelerated interest in Artificial Intelligence (AI) as an assistive tool in imaging. Meta-analyses indicate that deep learning systems can achieve diagnostic performance comparable to clinicians under study conditions, while also highlighting the need for rigorous external validation and careful study design [10]. Despite growing interest in AI for pediatric neuroimaging, key gaps remain. Most prior studies focus on narrow clinical tasks and rely on 2-D slice-based models that inadequately capture 3-D anatomical context, particularly important in rapidly developing infant brains. Very few works evaluate volumetric deep learning approaches in the 0–36-month age range, where rapid maturation and heterogeneous acquisition protocols pose unique challenges. Dataset sizes are often small, performance is rarely assessed on rigorously held-out cohorts, and the impact of model depth under limited pediatric data is largely unexplored. For volumetric data such as brain MRI, 3-D convolutional networks explicitly model inter-slice anatomical context and can outperform 2-D, slice-based approaches across several applications [11].

In this work, we formulate early detection of pediatric brain abnormalities as a supervised 3-D volumetric classification problem on T2-weighted MRI. We assemble 833 scans from patients aged 0–36 months acquired on a single scanner (under varying protocols), binarize labels into Normal vs. Not healthy from radiology reports, and standardize each volume to a 96×96×96 grid for efficient batching. We then compare two densely connected architectures, (a) DenseNet-121 and (b) DenseNet-264, adapted to 3-D convolutions. Our study is designed to (i) quantify the feasibility and headroom of

 ISBN: 978-1-68558-362-0

end-to-end 3-D classification in this age range, (ii) assess depth/capacity trade-offs under limited pediatric data, and (iii) report clinically meaningful metrics such as accuracy and confusion matrix on a strictly held-out test set. The broader goal is to evaluate whether such a model can pragmatically support pediatric neuroradiology workflows by accelerating case prioritization while preserving expert oversight [12].

The remainder of this paper is organized as follows. Section II reviews prior work in medical image analysis and pediatric neuroimaging. Section III details the dataset, preprocessing steps, model architectures, and training procedures, while Section IV presents the experimental results. Section V provides a critical evaluation of model performance and clinical relevance, and Section VI concludes with future directions for improving pediatric MRI classification.

II. Related work

Deep learning has reshaped medical image analysis across modalities and tasks such as classification, detection, segmentation, registration, and even reconstruction. This is accomplished by replacing hand-engineered features with end-to-end representation learning documented in comprehensive surveys of the field [13]. Meta-analytic evidence further suggests that, under study conditions, diagnostic performance of deep models can match that of health-care professionals, while also underscoring persistent gaps around external validation and study quality [10].

Within neuroimaging, volumetric MRI has catalyzed 3-D convolutional approaches that explicitly model inter-slice context; seminal architectures such as 3D U-Net and multiscale 3-D Convolutional Neural Networks (CNN) (with CRF post-processing) established strong baselines for brain lesion segmentation [14]. Subsequent reviews report robust performance of 3-D CNNs across segmentation and classification problems in MRI and other modalities, including neurodegenerative disease classification [15]. Community benchmarks like the Brain Tumor Segmentation (BraTS) challenge have further accelerated progress by standardizing datasets and evaluation protocols and by tracking gains across successive competition years [16].

Beyond general surveys of medical imaging AI, pediatric neuroimaging presents distinct challenges. There are rapid, non-linear changes in tissue contrast due to myelination, smaller brain size, and a higher prevalence of congenital or developmental anomalies relative to adult cohorts. Much of the mature literature and public benchmarks, such as BraTS centers on adult pathologies and segmentation tasks, leaving pediatric classification comparatively underexplored. Prior pediatric efforts have often targeted specific conditions (e.g., age estimation/myelination staging, neonatal white-matter injury, posterior fossa tumors) are relied on 2-D slice-based models with ensembling across slices, which can dilute cross-slice anatomical context [17]–[19]. Although effective in narrow settings, these approaches cannot fully capture the subtle, spatially distributed patterns present in early-life T2 MRI and are further limited by the rapid anatomical changes and small brain structures characteristic of the 0–36 month age range.

Methodologically, 3-D CNNs capture volumetric dependencies that are critical in pediatric brains where subtle, spatially distributed patterns carry diagnostic signal. Dense connectivity (as in DenseNet) promotes feature reuse and stabilizes optimization for deeper networks—properties desirable when labeled pediatric datasets are modest in size [20]. While 3-D DenseNet variants have shown promise in adult neuroimaging and other modalities, there remain relatively few end-to-end studies on full-volume 3-D classification of pediatric T2 MRI in the first three years of life. Our work contributes to this gap by training and evaluating 3-D DenseNet-121 and DenseNet-264 on a curated 0–36-month cohort with standardized preprocessing and a class-stratified held-out test set, providing a reproducible baseline for future pediatric studies.

Unlike prior pediatric studies that rely primarily on 2-D slice-based aggregation models, this work implements full-volume 3-D DenseNet architectures for binary abnormality screening in infants aged 0–36 months. To our knowledge, few studies have systematically evaluated depth-related generalization trade-offs in 3-D CNNs under limited pediatric data conditions using a strictly held-out cohort.

III. Methodology

This section describes the dataset, preprocessing pipeline, model architectures, and evaluation protocol used to assess three-dimensional (3-D) DenseNet classifiers for early pediatric MRI abnormality screening.

A. Dataset and Data Source

We curated a cohort of 833 pediatric T2-weighted brain Magnetic Resonance Imaging (MRI) volumes from patients aged 0–36 months. All scans were acquired at a single institution using the same scanner, though imaging protocols varied across examinations. Of the 833 cases, 565 were labeled as Normal and 268 as Not healthy based on radiologist reports.

Each MRI volume was paired with structured metadata (`meta.csv`) containing an `image_id` and corresponding radiologist-derived diagnosis. Records lacking a corresponding T2-weighted volume were excluded following a file-system audit keyed by `image_id`. For supervised learning, diagnoses were binarized as *Normal* when no pathology was reported and *Not healthy* otherwise.

All labels were derived from clinical reports rather than manual re-annotation. Demographic variables such as gender distribution were not available in the metadata and therefore were not incorporated into modeling.

B. Data Pre-processing

All MRIs were originally stored in NIfTI (`.nii.gz`) format. For compatibility with the training framework, volumes were converted to PyTorch tensor (`.pt`) format.

Preprocessing consisted of the following standardized steps applied uniformly across all scans:

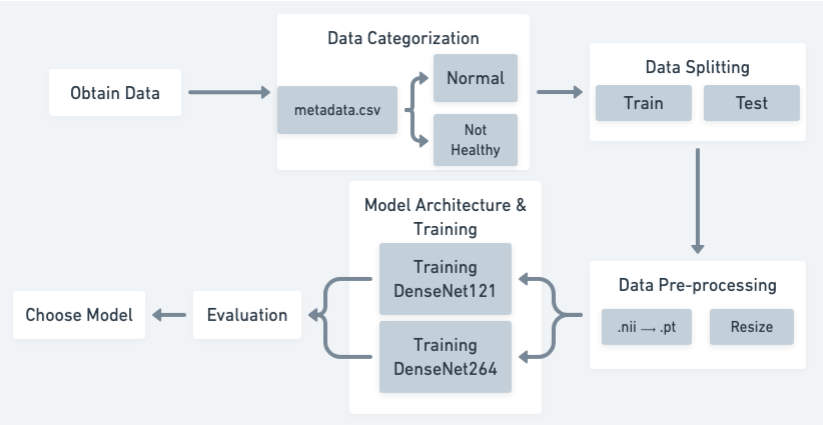


Figure 1. Flowchart of Data Processing and Machine Learning Stages

- Canonical reorientation to ensure consistent anatomical alignment,
- Intensity clipping to robust percentile ranges to reduce the influence of outliers,
- Per-volume z-score normalization,
- Uniform resampling to a fixed grid of $96 \times 96 \times 96$ voxels.

Resampling ensured consistent input dimensions across subjects and enabled efficient batched training of 3-D convolutional models.

The overall data preparation and modeling workflow is summarized in Figure 1. Notably, preprocessing was performed prior to dataset splitting to ensure consistent transformations across all samples.

C. Model Architectures

Two three-dimensional convolutional neural network (CNN) classifiers were implemented based on DenseNet-121 and DenseNet-264 architectures. All two-dimensional convolutional, pooling, and normalization operations were replaced with their three-dimensional counterparts. The initial convolution layer was adapted to accept single-channel T2-weighted input volumes.

Dense connectivity promotes feature reuse and improves gradient flow across depth, while transition layers with compression constrain parameter growth. These properties are advantageous when training deep networks on modest-sized pediatric datasets characterized by anatomical heterogeneity.

D. Model Evaluation

To obtain an unbiased estimate of generalization, we employed a stratified hold-out validation strategy. Twenty percent of cases from each class (Normal and Not healthy) were reserved as a strictly held-out test set. This cohort remained completely unseen during model training and hyperparameter tuning.

The remaining 80% of the data were used for model development, with a validation subset carved from the training portion for hyperparameter selection.

Models produced a single sigmoid output and were trained using weighted binary cross-entropy loss to mitigate class imbalance. Training employed mini-batches with on-the-fly three-dimensional data augmentation.

A coarse hyperparameter sweep explored learning rates in the range $[10^{-6}, 5 \times 10^{-2}]$ and training durations between 10 and 50 epochs. Three-dimensional DenseNet variants are known to be sensitive to optimizer step size; excessively large learning rates can cause unstable optimization, whereas excessively small values may hinder convergence. The selected range therefore spans conservative to moderately aggressive update regimes.

The upper bound of 50 training epochs reflects the tendency of high-capacity 3-D models to overfit under limited data conditions. Early stopping was guided by validation performance. The final configuration for each architecture was selected based on validation accuracy, and all performance metrics reported in Section IV were computed exclusively on the untouched test set.

Evaluation metrics included accuracy, precision, recall (sensitivity), F1-score, and confusion matrix counts.

IV. Results

A. Hyperparameter Tuning

A coarse grid search was performed over learning rates $\{10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, 5 \times 10^{-3}, 10^{-2}, 5 \times 10^{-2}\}$ and training durations of 10, 20, 30, 40, and 50 epochs for both architectures. For DenseNet-121, validation accuracy ranged roughly from 60% to 80% across the grid. Performance improved consistently as training progressed from 10 to 50 epochs at moderate learning rates, with the best validation accuracy of 80.60% obtained at 50 epochs and a learning rate of 10^{-4}, as shown in Table I.

For the deeper 3-D DenseNet-264, validation accuracy were slightly lower on average, and there was a strong dependence on conservative learning rates. Table II displays that the best validation accuracy obtained was 78.11% at 40 epochs and a learning rate of 10^{-5}. Similar to DenseNet-121, intermediate learning rates and moderate training durations (20–40 epochs)

TABLE I. VALIDATION ACCURACIES OBTAINED DURING HYPERPARAMETER TUNING FOR DENSENET121

Epochs / Learning Rate	0.000001	0.00001	**0.0001**	0.001	0.005	0.01	0.05
10	64.68%	74.13%	71.64%	76.12%	78.11%	66.67%	71.64%
20	71.64%	75.12%	67.16%	74.63%	76.12%	71.64%	68.66%
30	73.13%	72.14%	70.65%	77.61%	74.13%	75.12%	68.66%
40	75.12%	75.62%	71.64%	74.13%	77.11%	60.70%	66.67%
50	71.64%	80.10%	**80.60%**	73.13%	74.63%	66.17%	76.12%

tended to perform well, while large learning rates (10^{-2}, 5×10^{-2}) led to earlier saturation or degradation, showing that the higher capacity model is more susceptible to overfitting and unstable updates when learning rates are too high.

Overall, the grid search indicates that (i) both architectures benefit from training beyond 20 epochs, (ii) DenseNet-121 is slightly more forgiving to a range of learning rates, and (iii) DenseNet-264 yields its best performance under more conservative optimization settings.

TABLE II. VALIDATION ACCURACIES OBTAINED DURING HYPERPARAMETER TUNING FOR DENSENET264

Epochs / Learning Rate	0.000001	**0.00001**	0.0001	0.001	0.005	0.01	0.05
10	71.14%	73.13%	73.63%	71.14%	71.64%	56.72%	65.67%
20	73.63%	76.62%	73.63%	75.12%	74.13%	69.15%	63.68%
30	72.64%	66.67%	69.65%	72.64%	76.12%	73.63%	72.64%
40	70.15%	**78.11%**	77.11%	72.64%	65.17%	67.16%	68.66%
50	74.13%	77.61%	75.12%	67.66%	67.66%	73.13%	58.21%

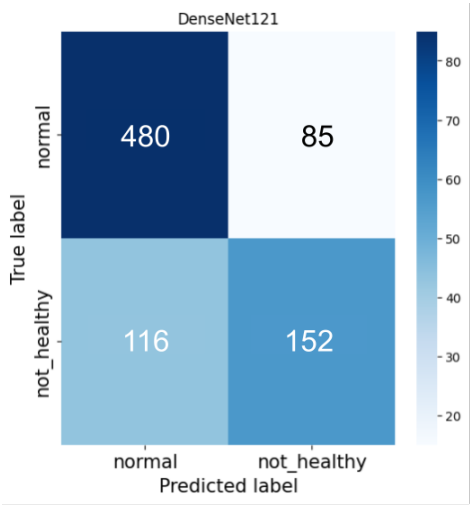
B. Test Set Performance

Using the best configuration selected on the validation split, final evaluation was conducted on a strictly held-out test set. The 3-D DenseNet-121 achieved an overall test accuracy of 75.90%, whereas the 3-D DenseNet-264 achieved 74.70% as shown in Table III. Thus, despite its greater depth and slightly higher peak validation performance in some regions of the search space, DenseNet-264 did not translate this advantage into generalization to the unseen test cohort.

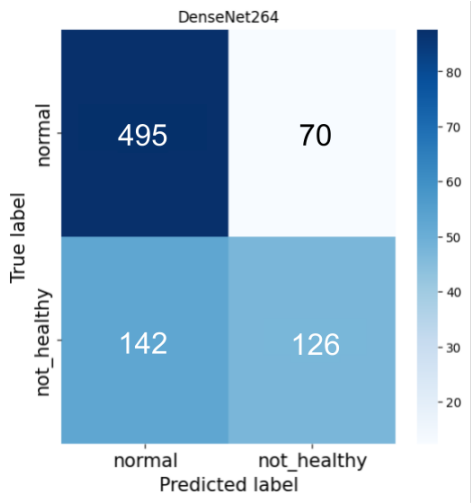
TABLE III. CLASSIFICATION REPORT ON MODEL PERFORMANCES

Models	Accuracy	Precision	Recall	F1 Score
DenseNet121	75.90%	0.75	0.76	0.75
DenseNet264	74.70%	0.74	0.75	0.74

Figure 2 summarizes the class-wise behavior via confusion matrices. Visual inspection shows that DenseNet-264 attains a lower true positive rate (sensitivity) for the Not healthy class but at the cost of a decreased false positive rate, i.e., more Normal scans incorrectly flagged as Not healthy. Similarly, DenseNet-121 attains a higher true positive rate but at the cost of an increased false positive rate.



(a) DenseNet121



(b) DenseNet264

Figure 2. Normalized confusion matrices (in %) for DenseNet121 and DenseNet264.

V. DISCUSSION

The experimental results highlight several important trade-offs between model capacity, optimization, and clinical utility in pediatric brain MRI classification.

First, the accuracy difference between the two architectures on the held-out test set is modest (75.90% for DenseNet-121 vs. 74.70% for DenseNet-264), but it consistently favors the shallower model. Given that both networks were trained on the same curated 0–36 month cohort, this suggests that additional depth and parameters do not automatically translate into better generalization under limited pediatric data. Dense connectivity facilitates feature reuse in both variants, but the larger DenseNet-264 may be more prone to overfitting subtle, cohort-specific patterns unless regularization, data augmentation, or larger training sets further constrain the solution space.

Second, the confusion matrices in Figure 2 show a clinically relevant trade-off. DenseNet-121 attains higher sensitivity for Not healthy scans, which is attractive for a screening or triage role where missing a true abnormality is costly. However, this comes with an increased false positive rate, which could generate more cases for radiologist review that ultimately prove normal. DenseNet-264 attains lower sensitivity for Not healthy

scans.

Third, the hyperparameter sweeps underscore that both architectures are sensitive to learning rate and training duration. DenseNet-121 exhibited robust performance across a band of moderate learning rates (10^{-5} to 10^{-3}) when trained for 30–50 epochs, while DenseNet-264 required a more conservative setting (e.g., 10^{-5}, 40 epochs) to reach its best validation performance. This pattern suggests that future work could benefit from more aggressive regularization (such as stronger augmentation, weight decay, or stochastic depth).

From a workflow perspective, these results support the feasibility of end-to-end 3-D classification of pediatric T2-weighted MRI in the first three years of life. A model with mid-70% test accuracy is not intended to replace expert interpretation but can serve as an assistive tool flagging likely abnormal studies for prioritized review, highlighting challenging cases for second reads, or providing a consistent baseline against which future improvements can be measured.

Future extensions include expanding the dataset to improve generalization, incorporating additional sequences or metadata (e.g., age or protocol information), and validating the models on external cohorts from other scanners and institutions. Integrating interpretability techniques (e.g., 3-D saliency or attribution maps) could further support clinician trust by localizing regions that drive model predictions. Ultimately, the present study establishes a reproducible 3-D DenseNet baseline for pediatric T2 MRI classification and illustrates how architecture depth and hyperparameter choices influence both numerical performance and clinically relevant error profiles.

While direct comparison to radiologist performance was beyond the scope of this study, the model is positioned as an assistive triage mechanism rather than a diagnostic replacement. Future work should incorporate reader studies comparing model output against initial radiologist screening decisions.

A. Limitations

1) Single-Scanner Limitation: All MRIs were acquired from a single scanner and institution. This represents the primary threat to external validity. Scanner-specific acquisition parameters, coil configurations, and reconstruction pipelines may introduce site-dependent biases. Performance may decline when applied to external institutions with different hardware or protocols. Multi-center validation is therefore a critical next step.

2) Heterogeneous "Not Healthy" Label: The "Not healthy" category aggregates multiple pathological conditions, including structural malformations, white-matter abnormalities, and other developmental findings. This coarse binarization simplifies the classification task but obscures subtype-specific performance variation. Future studies should explore multi-class labeling to quantify differential sensitivity across abnormality types.

3) Clinical Utility Threshold: Although mid-70% accuracy demonstrates feasibility, this level of performance is insufficient for autonomous clinical deployment. In real-world triage settings, higher sensitivity—potentially above 90%—would likely be required to ensure minimal missed pathology. Accordingly, the present study should be interpreted as a proof-of-concept baseline rather than a deployable system.

VI. Conclusion and Future Work

This study validates the application of DenseNet architecture variants for classifying pediatric brain MRI, demonstrating the promise of 3-D convolutional networks in enhancing diagnostic accuracy. DenseNet-121 outperformed the deeper DenseNet-264 on the strictly held-out test set, suggesting that increased architectural depth did not improve generalization under limited pediatric data. However, the necessity to address increased false positives remains paramount. These findings contribute an openly documented 3-D pediatric MRI classification baseline that future studies can benchmark against when evaluating larger multi-institutional datasets.

Future investigations will aim to expand our dataset, thereby enhancing model generalization and benchmark comparison. Additionally, exploring hybrid models with attention mechanisms could provide better interpretability and accuracy by focusing computational resources on critical brain regions. Deployment within clinical routines should involve an iterative feedback mechanism from radiologists to facilitate continual learning and system refinement, ultimately supporting efficient and accurate pediatric neurodiagnostic processes.

References

[1] J. M. Zubler et al., "Evidence-informed milestones for developmental surveillance tools", *Pediatrics*, vol. 149, no. 3, Mar. 2022.

[2] S. A. Bélanger and J. Caron, "Evaluation of the child with global developmental delay and intellectual disability", *Paediatr. Child Health*, vol. 23, no. 6, pp. 403–419, Sep. 2018.

[3] A. S. Ali et al., "Magnetic resonance imaging (MRI) evaluation of developmental delay in pediatric patients", *J. Clin. Diagn. Res.*, vol. 9, no. 1, TC21–4, Jan. 2015.

[4] A. M. Pagnozzi et al., "Brain mri before and at term equivalent age predicts motor and cognitive outcomes in very preterm infants", *NeuroImage: Reports*, vol. 5, no. 2, p. 100 262, 2025.

[5] S. M. Jung, "Drug selection for sedation and general anesthesia in children undergoing ambulatory magnetic resonance imaging", *Yeungnam University Journal of Medicine*, vol. 37, no. 3, pp. 159–168, 2020.

[6] F. P. Beaulieu et al., "Sedation and anesthesia for imaging of the infant and neonate—a brief review", *Pediatric Radiology*, vol. 54, no. 10, pp. 1579–1588, 2024.

[7] E. Rula, "Radiology workforce shortage and growing demand something has to give", *Am Coll Radiol*, 2024.

[8] European Society of Radiology (ESR), "What the radiologist should know about artificial intelligence—an esr white paper", *Insights into Imaging*, vol. 10, no. 44, 2019. DOI: 10.1186/s13244-019-0738-2.

[9] M. Annarumma et al., "Automated triaging of adult chest radiographs with deep artificial neural networks", *Radiology*, vol. 291, no. 1, pp. 196–202, 2019. DOI: 10.1148/radiol.2018180921.

[10] X. Liu et al., "A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: A systematic review and meta-analysis", *The lancet digital health*, vol. 1, no. 6, e271–e297, 2019.

[11] S. P. Singh et al., "3d deep learning on medical images: A review", *Sensors*, vol. 20, no. 18, p. 5097, 2020.

[12] A. Bhatia, F. Khalvati, and B. B. Ertl-Wagner, "Artificial intelligence in the future landscape of pediatric neuroradiology: Opportunities and challenges", *AJNR American Journal of Neuroradiology*, vol. 45, no. 5, pp. 549–553, 2024. DOI: 10.3174/ajnr.A8086.

[13] G. Litjens et al., "A survey on deep learning in medical image analysis", *Medical image analysis*, vol. 42, pp. 60–88, 2017.

[14] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3d u-net: Learning dense volumetric segmentation from sparse annotation", in *International conference on medical image computing and computer-assisted intervention*, Springer, 2016, pp. 424–432.

[15] A. E. Ilesanmi, T. O. Ilesanmi, and B. O. Ajayi, "Reviewing 3d convolutional neural network approaches for medical image segmentation", *Heliyon*, vol. 10, no. 6, 2024.

[16] B. H. Menze et al., "The multimodal brain tumor image segmentation benchmark (BRATS)", *IEEE Trans. Med. Imaging*, vol. 34, no. 10, pp. 1993–2024, Oct. 2015.

[17] S. P. Singh et al., "3d deep learning on medical images: A review", *Sensors*, vol. 20, no. 18, p. 5097, 2020. DOI: 10.3390/s20185097.

[18] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3d u-net: Learning dense volumetric segmentation from sparse annotation", in *MICCAI*, ser. Lecture Notes in Computer Science, vol. 9901, Springer, 2016, pp. 424–432. DOI: 10.1007/978-3-319-46723-8_49.

[19] K. Kamnitsas et al., "Efficient multi-scale 3d cnn with fully connected crf for accurate brain lesion segmentation", *Medical Image Analysis*, vol. 36, pp. 61–78, 2017. DOI: 10.1016/j.media.2016.10.004.

[20] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks", in *CVPR*, 2017, pp. 2261–2269. DOI: 10.1109/CVPR.2017.243.

Enhancing MRI Analysis in Temporal Lobe Epilepsy: Sequential Classification and Segmentation of Hippocampal Structures

Vihan Bhattacharjee
Bellarmine College Prep
San Jose, USA
e-mail: v.bhattacharjee27@bcp.org

Kamalasree Sudhakar
AI Club
Mountain View, USA
e-mail: kamalasree.s@aiclub.ai

Sindhu Ghanta
AI Club
Mountain View, USA
e-mail: sindhu.ghanta@aiclub.ai

***Abstract*—Temporal Lobe Epilepsy is a chronic neurological disorder affecting 50 million people worldwide, with diagnosis often complicated by overlapping symptoms with other conditions. Traditional diagnostic methods, like electroencephalography and manual Magnetic Resonance Imaging analysis, face challenges of subjectivity and inaccuracies. This study explores machine learning approaches to address these limitations, utilizing MobileNet and U-Net models to improve diagnostic accuracy and efficiency in Temporal Lobe Epilepsy. MobileNet, a lightweight classification model, successfully distinguishes between normal and abnormal hippocampi, achieving a classification accuracy of 94%. This is further enhanced by the U-Net segmentation network, which achieves an Intersection over Union score of 0.902 in identifying hippocampal abnormalities associated with Temporal Lobe Epilepsy. The integration of these models significantly improves diagnostic precision and efficiency, offering substantial support to radiologists in clinical settings. The promising results underscore the potential of advanced machine learning techniques in enhancing diagnostic processes. Future research will aim to increase model robustness by diversifying training datasets and exploring alternative algorithmic approaches, thereby improving the models' applicability and reliability across various clinical environments.**

Keywords-Temporal Lobe Epilepsy; Machine Learning; MobileNet; U-Net; Hippocampal Abnormalities.

I. Introduction

Temporal Lobe Epilepsy (TLE) is a chronic neurological disease caused by the abnormal discharge of neurons in the brain [1]. TLE affects about 50 million people worldwide [2]. It is the fourth most common neurological disorder, and it impacts people of all ages, including adults, children, and newborns [3].

Seizures caused by TLE can cause brain injury, physical harm, and even a shorter lifespan. Specifically, TLE can lead to significant memory loss due to damage in the hippocampus, as well as impacting attention, learning, and language processing, all of which are vital for daily tasks [4]. Additionally, recent research has indicated that individuals with TLE are at higher risk of depression, anxiety, and mood instability, stemming from dysfunction in the limbic system [5]. The most alarming consequence of TLE is the increased risk of mortality from sudden unexpected death in epilepsy (SUDEP), vascular diseases, suicide, pneumonia, and underlying causes like brain tumors: all of these factors cause individuals with epilepsy to have a mortality rate as much as 11.6% higher than expected [6]. That is why it is essential to ensure early and accurate diagnosis of TLE to increase the chance for patients to receive proactive treatment and restore their quality of life. Currently, epilepsy diagnosis involves multiple steps that combine clinical observation with testing to identify both the seizure type and its underlying cause. Electroencephalography (EEG) serves as the primary diagnostic tool for epilepsy, capturing the brain's electrical signals to detect the irregular patterns characteristic of the condition. However, interpretations of EEG scans are subjective, and radiologists may misdiagnose TLE as other conditions, such as a primary psychiatric disorder (i.e., schizophrenia) [7], which has many similar symptoms as TLE. Additionally, diagnosing epilepsy presents significant challenges due to its non-specific symptoms that sometimes look similar to those of other medical conditions. The diagnostic process primarily depends on reviewing the patient's medical history and documenting seizure episodes from the witnesses, which can be incomplete or inaccurate. That is why previous studies have demonstrated that prior seizures experienced by patients are often not diagnosed as epilepsy, impacting as many as 38% of epilepsy patients [8]. To overcome these limitations in diagnosing TLE, researchers have explored alternative diagnostic methods, particularly by examining the state of the hippocampus [9], a structure located deep within the temporal lobe. The hippocampus plays a crucial role in memory processing, emotion regulation, spatial navigation, and learning [10]. In TLE, the hippocampus often shows damage or structural changes, such as hippocampal sclerosis, in which nerve cells are lost and scarring develops. Hippocampal sclerosis is present in between 30.5% and 45% of all epilepsy syndromes and in 56% of cases of TLE [11]. Therefore, it is valuable for radiologists to detect hippocampal abnormalities in potential TLE patients.

Magnetic Resonance Imaging (MRI) with efficient spatial navigation capabilities is key for evaluating the state of the hippocampus [12]. The process of identifying and marking the hippocampus in MRI scans relies heavily on manual techniques, which require a significant time investment and are subject to inter-observer differences. For example, in temporal lobe epilepsy, hippocampal sclerosis can sometimes appear as very mild shrinkage or subtle signal changes, which may be overlooked even by experienced radiologists [13]. Additionally, processing MRI images presents challenges due to the substantial computational resources needed and the extensive time required for completion [1].

Automation can assist radiologists in enhancing their detection process. Machine learning and deep learning algorithms can extract complex features from MRI images that are often

 ISBN: 978-1-68558-362-0

difficult for the human eye to distinguish, enabling precise detection of hippocampal sclerosis and other hippocampal abnormalities. Models trained on MRI data have immense potential to improve the accuracy of detecting hippocampal abnormalities; for instance, deep learning models can perform automated segmentation of the hippocampus, measure its volume precisely, and detect microstructural changes [14]. However, there are many considerations researchers must make when constructing these models, including data quality, data generalizability, and preprocessing methods. This study explores the implementation of the U-Net segmentation algorithm to localize hippocampal abnormalities, as well as the MobileNet classification model to serve as a prior step in distinguishing healthy from non-healthy hippocampal tissues, providing a comprehensive approach to automated TLE diagnosis.

The paper is organized and begins with an introduction in Section I and exploration of related work in Section II. An explanation of the various stages of methodology was in Section III, followed by results in Section IV. Section V contains the discussion. Finally, a conclusion about the implications of the study is present in Section VI.

II. Related work

Several researchers have explored the potential of utilizing computer vision deep learning algorithms to detect minute or microstructural abnormalities in hippocampal MRIs to assist in TLE diagnosis. Segmentation models have shown especially promising results, as they are able to precisely delineate the hippocampal region with minimal human intervention. Experiments with segmentation are often conducted by incorporating transfer learning to pre-trained Convolutional Neural Networks (CNN) to minimize computational resources while maximizing accuracy. However, some common limitations in these studies are insufficient hyperparameter optimization, high computational resource requirements for training complex models, and class imbalance due to limited availability of hippocampal-specific slices.

For example, Chaouch et al. [15] studied hippocampal atrophy in mesial temporal lobe epilepsy via CNN deep learning segmentation. While these segmentation models show success in isolating the particular volume and location of hippocampus atrophy, as seen through a Dice score of 0.86 and sensitivity of 93%, the datasets used to train these models have a vast imbalance of relevant shots of the hippocampus: out of 3050 slices used in the model, only 459 actually contain hippocampal regions, making approximately 80% of the data obsolete and rendering the model liable to class imbalance issues and overfitting.

Additionally, Chang et al. [16] utilized a custom CNN architecture with 3 main layers to diagnose TLE and Alzheimer's disease from MRI brain scans by analyzing both the hippocampal region, as well as the broader structure of the brain. Their model achieves an accuracy of 90.45% in correctly identifying the presence of TLE, and an F1 score of 0.85. However, a lack of exhaustive hyperparameter tuning, as acknowledged by the researchers themselves, likely constrained the model's accuracy and stability, as the most optimal epochs and learning rates have the potential to significantly improve model performance. Similarly, Jiang et al. [17] created a novel CNN architecture, HS-Net, to accurately detect the presence of hippocampal sclerosis characteristic of TLE, achieving results like an Area Under the Curve (AUC) of 0.894 and an accuracy of 82.88 . Nonetheless, the model fails to precisely delineate the location or structure of the impacted hippocampus, and thus only serves to classify the presence of hippocampal sclerosis without actually automating the diagnostic process.

Most notably, Widodo et al. [12] utilizes the U-Net model to segment the hippocampus in MRI images. They utilized transfer learning to train a 3D U-Net model with data from a range of patients with various cognitive states (normal, mild cognitive impairment, and Alzheimer's). They achieved an Intersection Over Union (IoU) of 0.75 and a sensitivity of 0.80. While this work demonstrates the effectiveness of U-Net architectures for hippocampal segmentation, it does not address class imbalance within the dataset or test how hyperparameter selection influences model performance. Similar to the approach in Widodo's study, this study utilizes the U-Net architecture for segmenting the hippocampus in MRI images. Furthermore, to overcome the flaws present in previous studies, a wider range of hyperparameters is used to test the architecture's efficacy and the dataset is balanced in preprocessing to ensure the model is trained upon an equal number of healthy and unhealthy scans. Most significantly, this study also conducts experiments with the CNN MobileNet to conduct classification on the MRI images and classify them as "healthy" vs "unhealthy" before segmentation; this creates a more real-world workflow that saves computational resources and only utilizes segmentation in necessary cases. Despite conducting experiments with a limited number of models, the proposed framework has the potential to localize hippocampal abnormalities more effectively and efficiently than traditional methods and other experimental approaches.

III. Methods

Dataset Analysis: This study utilized publicly available data from Jafari-Khouzani et al., a 2.58 GB dataset containing 50 brain T1-weighted MRI volumes with hippocampus labels [18]. The decision to use a dataset of T1-weighted MRI images rather than alternatives like T2 was driven by several key factors. T1-weighted MRI images provide high anatomical detail and clearly distinguish between gray and white matter, with white matter appearing lighter and gray matter darker [19]. This makes them ideal for detecting structural abnormalities within the hippocampus. Additionally, T1 scans produce sharper, higher-resolution images and are employed in procedures that require highly precise data [20].

Specifically, 40 of the images belong to patients who have temporal lobe epilepsy. Several of the patients had atrophic hippocampi, allowing for the model to train with real-world data with more challenging segmentation, while the remaining 10 images were from subjects without epilepsy and displayed normal hippocampi. The dataset also contained

 ISBN: 978-1-68558-362-0

labels that detailed the precise anatomical boundaries of the hippocampus within each volumetric scan. Due to the dataset's class imbalance, there may have been a risk of overfitting the model to become biased towards the unhealthy class. This data imbalance was mitigated by random undersampling, where the number of scans within the healthy class were made the same as the scans in the unhealthy class, equating both data subsets.

Data Preprocessing: The preprocessing phase of this study involved several systematic steps to prepare the dataset for model training. The first step began with doing a detailed dataset analysis to understand the distribution between the healthy and non-healthy image cases, and separating the images per class distribution. Figure 1 shows an unprocessed MRI scan. Image adjustments were done to load the MRI image and its corresponding label file to the model.

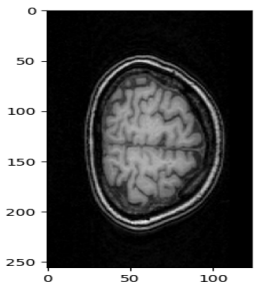


Figure 1. Sample 3D MRI scan from the dataset

Segmentation: To preprocess the images for segmentation, the first step was to convert the 3D MRI volumes into 2D image slices to simplify processing and model training. To convert the 3D volumetric data to 2D, individual image slices were extracted, which were cross-sectional images of the 3D images. The 2D slices were extracted in three orientations: axial (horizontal) slices, coronal (frontal) slices, and sagittal (lateral) slices. The MRI scans were divided into slices to increase the size of the dataset, facilitating more robust training; lower computational resources necessary for the models; and to leverage pre-trained segmentation models, which are traditionally trained on 2D slices.

After the slices were isolated, all of them were normalized to convert the wide-ranging MRI voxels into pixels with a standard range of intensity. Min-Max normalization was used to ensure that all the pixels contained a standard intensity from 0 to 255 [21]. The segmentation labels were also converted to binary masks, such as 0 for the background and 255 for the hippocampus region. The normalized images and the corresponding masks were saved in two separate folders as .png files. Figure 2 illustrates the methodology used for conducting the experiments of this research.

Classification: After resizing the previously gathered MRI slices to 224 pixels, each image was converted to a 3-channel RGB format to adhere to the requirements of MobileNet. Then, pixel values were rescaled from [0,255] to [-1,1], to be assigned to the labels of "unhealthy" and "healthy." Finally, class labels were converted to vectors (e.g., [0,1]) to be assigned to either health category. This ensured that every slice could be easily classified as either healthy or unhealthy during subsequent training.

Data Splitting: Then the preprocessed dataset was split into sets of training, validation, and test by using a random sampling approach. Approximately 10% of the dataset was allocated to validation, and 10% to testing. The remaining 80% was used for training. This ensured the proper evaluation of model performance. Figure 3 illustrates the completed MRI scan and the corresponding mask after the preprocessing stage.

Model Training and Validation: The study was structured around three sequential experiments designed to overcome the limitations inherent in using solely segmentation or classification models. Using only classification would lack the necessary anatomical localization for clinical treatment, while applying segmentation to every MRI slice is computationally inefficient, particularly for large datasets. This integrated approach minimizes computational resources and energy consumption while retaining critical diagnostic detail. Transfer learning was utilized across all models, leveraging pre-trained weights to accelerate convergence and improve performance given the dataset size.

The first experiment was Hippocampus segmentation using U-Net. This experiment focused on detecting the precise anatomical delineation of the hippocampus, a key step for volumetric analysis in TLE diagnosis. The U-Net CNN was selected due to its proven efficacy in biomedical image segmentation. U-Net models provide optimal results even with small tumor lesions or relatively unclear organ boundaries, making it especially valuable for medical image segmentation, which often faces these challenges [22]. U-Net is the optimal choice for researchers due to its ability to obtain accurate results, even when trained on a limited dataset [23], as the model automatically incorporate alterations, such as flips and rotations to expand the size of datasets, and incorporates end-to-end training, which enables researchers to train U-Net models directly on broad medical images, rather than handcrafted features, reducing the need for excessive data. Additionally, Lu et al. [24] finds that U-Net's divide-and-conquer encoder strategy, which divides an input image into five different feature maps, facilitates efficient segmentation and localization; this feature is crucial for clinical environments, where rapid decisions need to be made with accurate data. While other models may exhibit strong performances, Turk et al. [25] shows how U-Net achieves a comparable accuracy to other experimental models; for instance, in Lu et.al's study, the traditional U-Net model achieved a 97.49% accuracy, compared to 98.00% of a modified model. Furthermore, the U-Net model required the least training time and computational cost, compared to other deep learning models like V-Net and SegNet, making it more practical in healthcare settings.

During the training process, hyperparameters were meticulously tuned; learning rates ranged from 0.000001 to 0.05, while epochs were assessed from 10 to 50. The results of model training were analyzed with loss and the IoU metric, defined as Area of Overlap/Area of Union.

The second experiment was classification using MobileNet,

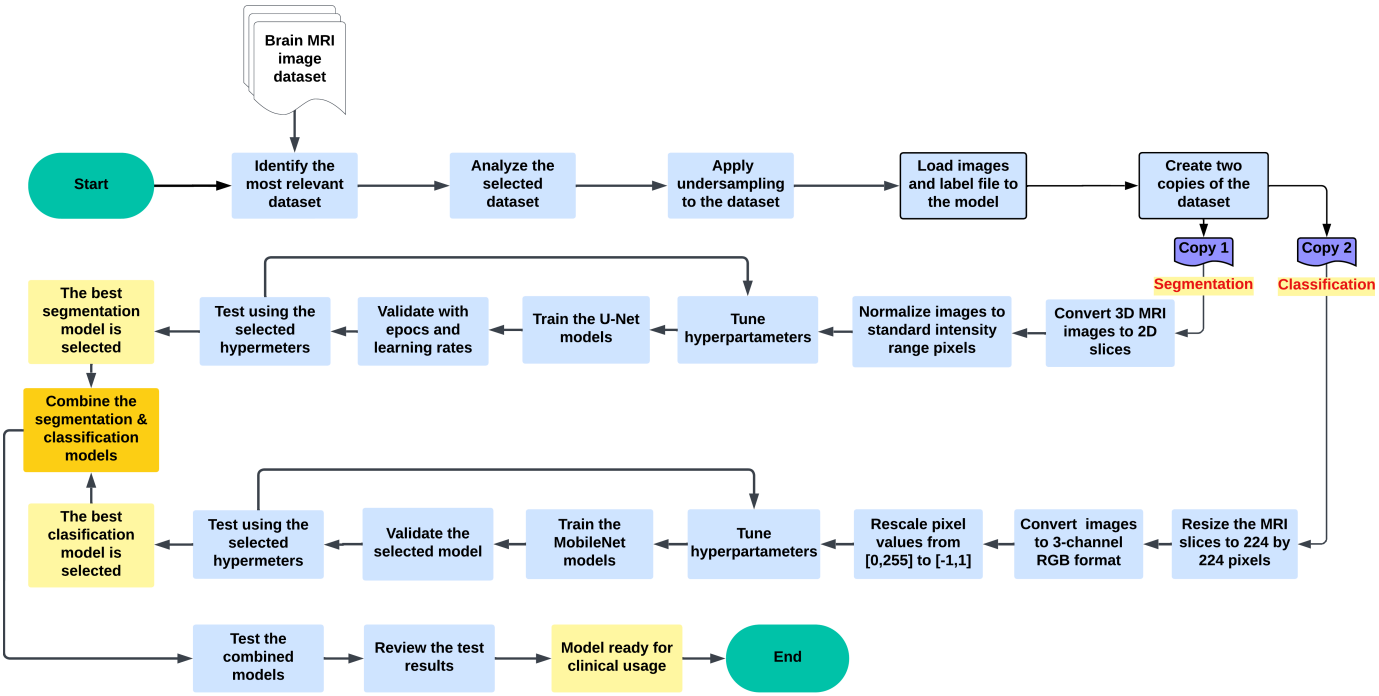
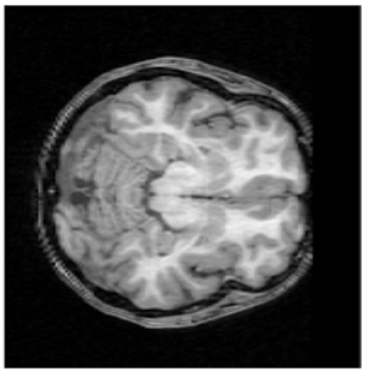


Figure 2. Flowchart of image Data Processing and Machine Learning stages



(a) Preprocessed 2D MRI scan

(b) Preprocessed mask

Figure 3. Preprocessed data for segmentation

a Convolutional Neural Network commonly used in image classification. The objective of this experiment was to create an efficient binary screening tool capable of differentiating between healthy and unhealthy hippocampus slices. The MobileNet architecture was chosen for its lightweight design and computational efficiency, making it ideal for the high-volume classification required in the initial screening phase.

Despite requiring fewer parameters and having significantly fewer computational needs than other similar CNNs, Ochoa-Ornelas et al.'s study [26] demonstrated how the model could achieve an accuracy of 98.77% on a dataset containing medical imagery. Additionally, MobileNet produces similar accuracies to other heavier models like Xception on medical imagery classification, as evidenced by Prakash et al. [27].

The model was trained to classify MRI slices into "healthy" or "unhealthy" categories. Similar to the segmentation experiment, hyperparameters were consistently monitored with epochs and learning rates in the same ranges. Final results were quantified by analyzing accuracy metrics and detailed confusion matrices to evaluate the rate of misclassification.

The final experiment simulated the real-world application, integrating the optimal classification and segmentation models sequentially. The MobileNet model (Experiment 2) first processed the entire testing set to identify slices likely exhibiting TLE pathology. Subsequently, the high-precision U-Net model (Experiment 1) was applied only to the slices flagged as "unhealthy" by the MobileNet model. This combined method validated the system's ability to minimize computational resources while providing accurate localization of the hippocampus, establishing a methodology with enhanced efficiency and accuracy. The final performance of this integrated

system was tested using the testing dataset. Final analysis of accuracy via sample images revealed high confidence scores as well as accurate segmentation.

Results Analysis: After validation and testing were completed for both the U-net and MobileNet models, the results were analyzed using IoU metrics, loss, and confusion matrices. In segmentation, the IoU value is a valuable metric that evaluates how much the predicted mask overlaps with the actual location of the hippocampus. For classification, the IoU compares the overlap between predicted and true positive classes. A higher IoU value indicates a more accurate model performance. Additionally, evaluation of loss for segmentation or confusion matrices for classification was conducted for each model to gain a greater understanding of specific types of errors or inaccuracies.

IV. Results

This section outlines findings from the segmentation and classification experiments. The results include analyses of IoU metrics, loss metrics, confusion matrices, and graphical and tabular comparisons of model performance.

In the segmentation experiment, the model was trained with the U-Net algorithm. There were 35 experiments using a combination of seven learning rates (0.000001-0.05) and epochs (10-50), as shown in Table I and II. IoU quantifies the degree of overlap between the pixels within the model's predicted mask and the ground truth images; a higher IoU indicates superior segmentation accuracy, reflecting the model's robustness across a variety of scans. IoU was used to measure the performance of the model because it provides a rigorous and interpretable metric that captures both false positives and false negatives, ensuring that the evaluation reflects the true spatial accuracy of the segmentation. Additionally, IoU also captures both over-segmentation and under-segmentation errors, providing a more comprehensive assessment than simple pixel accuracy.

TABLE I. Hippocampus Segmentation U-Net IoU Results

Epochs / Learning Rate	0.000001	0.00001	0.0001	0.001	0.005	0.01	0.05
10	0.0064	0.0193	0.5159	0.8028	0.7991	0.8242	0.0175
20	0.0074	0.024	0.7015	0.865	0.2977	0.8486	0.8401
30	0.6543	0.6802	0.73	0.8233	0.8709	0.8794	0.8588
40	0.8755	0.879	0.8886	**0.8961**	0.8944	0.8944	0.8776
50	0.0121	0.373	0.876	0.874	0.8707	0.8763	0.8642

TABLE II. Hippocampus Segmentation U-Net Loss Results

Epochs / Learning Rate	0.000001	0.00001	0.0001	0.001	0.005	0.01	0.05
10	0.3911	0.1297	0.0055	0.0022	0.0025	0.0023	0.0972
20	0.289	0.123	0.0031	0.002	0.0312	0.0024	0.0024
30	0.0039	0.0034	0.0028	0.0025	0.0031	0.003	0.0031
40	0.0023	0.0023	0.0027	**0.0064**	0.0072	0.0057	0.0036
50	0.195	0.009	0.0023	0.0029	0.0027	0.0031	0.0024

In the U-Net experiment, the highest IoU during validation of the model was achieved at epoch 40 with a learning rate of 0.01, yielding an IoU of 0.8961. When solely this combination of learning rate and epoch was used upon an unseen test dataset, the model yielded an IoU of 0.902, demonstrating the model's consistent yet accurate performance across a wide range of data. However, these same hyperparameters also yielded a relatively high loss of 0.0064, as shown in Table II, indicating that while the model correctly captured the shape of the target regions, it did so with uncertain probability outputs that were penalized by the loss function. This loss analysis is crucial to gauge the accuracy of the U-Net model, because it reveals how confidently the model delineates the location of the hippocampus. The loss function directly measures the model's probabilistic errors, demonstrating instances when confidence is low despite the model making a correct prediction.

(a) Sample image from test dataset (b) Ground truth hippocampus mask (c) U-Net predicted Mask

Figure 4. Successful Segmentation of Hippocampal Abnormalities

The image shown in Figure 4 highlights an example of how the segmentation done for the sample hippocampus image is accurate. The model successfully identifies both hippocampal regions in the correct locations. The accurate results suggest that the model can determine the anatomical structure due to its training.

In the classification experiment, the model was trained with the MobileNet algorithm. Similarly, 35 experiments were also conducted with the same range of learning rates and epochs, as shown in Table III. For these sets of experiments, accuracy was evaluated, along with results from confusion matrices, as shown in Figure 4. While accuracy measures what percent of the model's prediction accurately matches each image's corresponding class label, confusion matrices give further insight into the false positives and negatives the model may exhibit. These trends demonstrate the model's weaknesses and potential class imbalances, ultimately helping to identify areas where architectural adjustments or additional data are needed. Validation revealed that at 20 epochs and a learning rate of 0.01, the model reached its highest accuracy of 0.9483, reflecting that in over 94% of instances, the model was able to accurately identify the class of the image as "healthy" or "unhealthy". When these specific hyperparameters were tested on an unseen test dataset, the model yielded an accuracy of 0.94. Another metric that was analyzed was confusion matrices, which demonstrated that the model had significantly more true positives than true negatives or false positives. The model detected true positives at a rate of 87.36% and true negatives at 96.55%, as shown in Figure 5, indicating the model's ability to distinguish accurately between different images of the hippocampi.

TABLE III. CLASSIFICATION MOBILENET ACCURACY RESULTS

Epochs / Learning Rate	0.000001	0.00001	0.0001	0.001	0.005	0.01	0.05
10	0.5345	0.7874	0.8908	0.9195	0.9138	0.9195	0.8851
20	0.7701	0.8506	0.9138	0.9368	0.9425	**0.9483**	0.9195
30	0.7701	0.8678	0.9195	0.9368	0.9368	0.9425	0.5
40	0.7089	0.8678	0.931	0.9368	0.9425	0.931	0.9195
50	0.7759	0.8793	0.9368	0.9253	0.9425	0.9368	0.9253

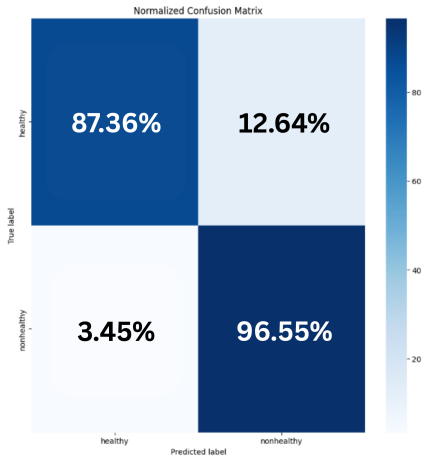


Figure 5. Confusion Matrix for MobileNet Classification on Test Dataset

V. DISCUSSION | EVALUTION

Segmentation Performance: Results from hyperparameter tuning demonstrate that the U-Net segmentation model accurately localizes the precise location of the hippocampus. Table I shows that learning rates between 0.0001 and 0.01 consistently yielded high IoU scores. At 10 epochs, only a few learning rates (0.001-0.01) achieved acceptable segmentation quality, with lower rates failing to converge, and higher rates showing instability. The results at 20 epochs showed some improvement, though the 0.005 learning rate experienced a dramatic collapse with IoU of 0.2977, demonstrating the model's instability in this range. The transition started at 30 epochs where learning rates 0.000001 and 0.00001 started showing improvement in IoU, with values above 0.6543. These results indicate the model requires comprehensive training, even for low learning rates, to gather sufficient information for accurate predictions.

A more accurate performance occurred at 40 epochs, with all IoU being greater than 0.8755. Similar to the other epochs, the learning rates at the lower end of the range contain the lowest IoU scores, demonstrating the model's lack of convergence at those rates. The best configuration, with the highest IoU, occurs when the optimal learning rate is 0.001 and the epoch is 40, leading to an IoU of 0.8961. However, at 50 epochs, the model's failure to converge at lower learning rates is apparent, with the lowest IoU being 0.0121. Results from higher learning rates do not show any improvement from 40 epochs, suggesting that higher levels of epochs do not increase performance and may introduce mild overfitting or training instability. Figure 6 represents the hyperparameter tuning process in a graphical format, plotting the learning rate and epoch versus IoU. This graph also corroborates the previous IoU analysis.

Additionally, loss results from this experiment also support the IoU findings, reflecting strong underlying prediction quality despite the differences in probability calibration. The loss metric captures the model's confidence while making predictions. At 10 epochs, there is generally very high loss, especially for very low learning rates (0.000001) and very high learning rates (0.05), demonstrating that the model is underfitting and is unadjusted to the dataset. The results at 20 epochs show a decrease in loss, demonstrating the model's growing accuracy. At 30 and 40 epochs, the model exhibits a stability in loss, with most loss values from 0.02 to 0.03. This suggests 30 and 40 epochs is sufficient for the model to achieve optimal convergence without overfitting. At 50 epochs, improvements in loss are marginal and are not offset by the additional computational resources necessary. At the hyperparameters with the high IoU output, the model yielded a loss of 0.0064, a relatively high loss, especially when compared to the model's performance with the same epoch and different learning rates. A relatively high loss occurring despite a high IoU may indicate the model's uncertainty while making pixel-level predictions. This lack of confidence may occur if the dataset contains ambiguous or noisy scans, making certain predictions difficult. Although augmenting the dataset would increase the model's confidence in prediction and decrease loss across the board, ambiguous data reflects real-world conditions present in clinical facilities, which may contain various scans with different qualities. However, clinicial validation by radiologists is necessary to fully validate the efficacy of the model in real-world settings. Overall, this selection of learning rates and epochs helped to balance convergence, stability, and overfitting, helping to create a model that could robustly and accurately localize abnormal hippocampal tissue.

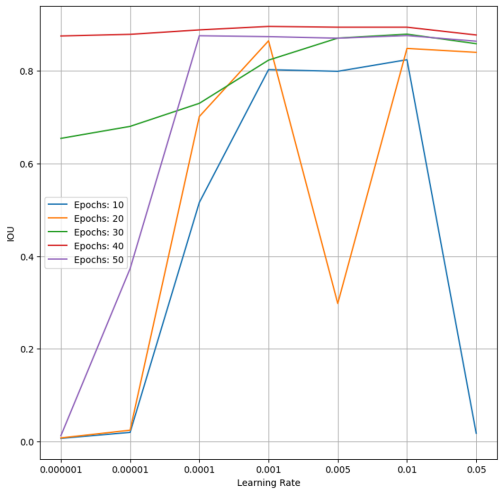


Figure 6. Graph of Segmentation IoU Results

The most optimal combination of hyperparameters yielded

an IoU of 0.902 on an unseen test dataset, highlighting the model's ability to generalize patterns accurately beyond the test dataset; the relative consistency between the validation and test IoU results demonstrates the model's stability and lack of overfitting.

Classification Performance: For the second experiment of classification, the MobileNet model demonstrated excellent balanced performance in distinguishing healthy from non-healthy hippocampal tissues. The performance of the model was recorded according to a simple accuracy metric, where the higher the score, the more accurate the model was in identifying healthy or unhealthy hippocampal scans. Similar to the segmentation experiment, hyperparameters were closely analyzed with the same ranges in learning rate and epochs. Table II demonstrates how all combinations of hyperparameters achieved relatively high accuracy scores, with the majority being greater than 0.77. Across all epochs, the lowest learning rate of 0.000001 had the lowest accuracy score, conveying how this small learning rate hindered gradient updates, preventing the model from capturing meaningful updates in each epoch. Lower-end learning rates (0.00001 and 0.0001) had varied results, but larger epochs achieved higher scores due to the model's ability to create more gradient updates with more epochs. The model achieved the best results with learning rates ranging from 0.001 to 0.01, indicating that moderately high learning rates allowed MobileNet to converge efficiently and extract discriminative features reliably for classification. The highest accuracy of 0.9483 was achieved at 20 epochs and a learning rate of 0.01, indicating how despite containing a lower number of epochs, the larger learning rate allowed the model to effectively adjust its gradients and learn patterns effectively. The largest learning rate of 0.05 fared slightly poorly compared to the previous epochs most likely due to training instability, with a sudden decrease in performance at 30 epochs. This model's accuracy results indicate how the MobileNet model fairs better with larger learning rates as they enable faster convergence and more effective gradient updates. The graph in Figure 7 highlights the accuracy results of the classification experiment, by showing the trends across the hyperparameters.

The normalized confusion matrix, as shown in Figure 5, corroborates the accuracy results and shows the classification model's performance in distinguishing between "healthy" and "non-healthy" cases. The model exhibits a high accuracy in identifying true positive cases, with 87.36% accuracy, as well as true negative cases, with a 96.55% accuracy. This indicates that the model is able to accurately discern if the MRI scan is healthy or unhealthy, with minimal error. The model fares slightly better when detecting healthy scans, as evidenced by its higher accuracy for true negatives. Conversely, the slightly lower true positive accuracy indicates that a small proportion of unhealthy scans may be misclassified as healthy, highlighting areas where further refinement during training could enhance sensitivity without compromising overall reliability.

These results can be analyzed in terms of precision and recall metrics. Precision measures the proportion of true positive predictions to total positive predictions, focusing on decreasing the occurrence of false positive predictions. Recall, however, measures the proportion of true positive predictions to total predictions and focuses on minimizing false negatives. The precision score for "healthy" cases was 0.96, but for "unhealthy" cases, it was 0.88. However, the recall score for "healthy" cases was 0.87, but for "unhealthy" cases, it was 0.97. These results illustrate how the model prioritized minimizing false negative predictions over isolating false positive cases. A false positive diagnosis may lead to unnecessary medical intervention, leading to additional procedures and tests that contain certain risks. In clinical settings, however, a false negative diagnosis may pose a more significant danger to patients, leading to delayed treatment and intervention, reducing the chances of optimal patient recovery. These findings demonstrate how the MobileNet algorithm has the potential to be deployed in a clinical setting to detect structural abnormalities in hippocampal MRIs, as it is able to accurately classify unhealthy from healthy scans. Further experiments will be conducted under the purview of radiologists to further validate the results.

Combination of Both Experiments: When a dual approach of the models was employed, with classification occurring before segmentation, the overall architecture yielded success. When tested on an unseen test image, the architecture exhibited a confidence score of 1 and was able to accurately segment the location of the impacted hippocampus on a prediction mask. The overall results from all three experiments demonstrate how a dual deep learning approach with classification and segmentation is able to successfully detect hippocampal abnormalities, and ultimately detect the presence of TLE. This approach has the potential to assist physicians with diagnosing TLE, as these complementary models produce clinically meaningful outputs, which guide further evaluation and treatment planning.

Despite the model's potential for success, there were some limitations within the methodology of the study which may have constrained the model's reliability. The analysis was conducted using a single-source dataset, which only contained 50 scans; this limited amount of data may have limited the model's exposure to the diversity of data found in typical

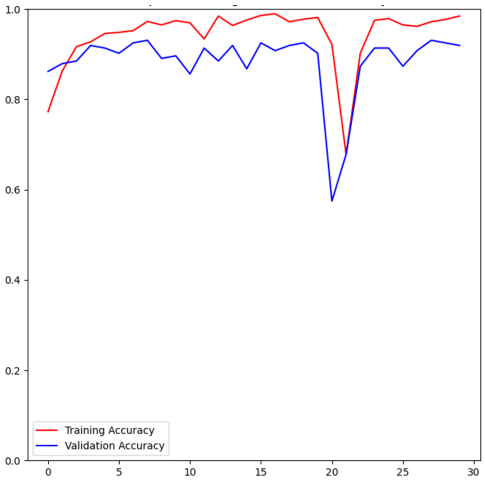


Figure 7. Graph of Classification Accuracy Results

clinical settings. Additionally, a more detailed analysis of other varieties of classification and segmentation models is necessary to determine the best possible architectures to achieve the highest accuracies and greatest generalizability across varied datasets.

For the second experiment of classification, the MobileNet model demonstrated excellent balanced performance in distinguishing healthy from non-healthy hippocampal tissues. The performance of the model was recorded according to a simple accuracy metric, where the higher the score, the more accurate the model was in identifying healthy or unhealthy hippocampal scans. Similar to the segmentation experiment, hyperparameters were closely analyzed with the same ranges in learning rate and epochs. Table II demonstrates how all combinations of hyperparameters achieved relatively high accuracy scores, with the majority being greater than 0.77. Across all epochs, the lowest learning rate of 0.000001 had the lowest accuracy score, conveying how this small learning rate hindered gradient updates, preventing the model from capturing meaningful updates in each epoch. Lower-end learning rates (0.00001 and 0.0001) had varied results, but larger epochs achieved higher scores due to the model's ability to create more gradient updates with more epochs. The model achieved the best results with learning rates ranging from 0.001 to 0.01, indicating that moderately high learning rates allowed MobileNet to converge efficiently and extract discriminative features reliably for classification. The highest accuracy of 0.9483 was achieved at 20 epochs and a learning rate of 0.01, indicating how despite containing a lower number of epochs, the larger learning rate allowed the model to effectively adjust its gradients and learn patterns effectively. The largest learning rate of 0.05 fared slightly poorly compared to the previous epochs most likely due to training instability, with a sudden decrease in performance at 30 epochs. This model's accuracy results indicate how the MobileNet model fairs better with larger learning rates as they enable faster convergence and more effective gradient updates.

The normalized confusion matrix as shown in Figure 5 exhibits a high accuracy in identifying true positive cases, with 87.36% accuracy, as well as true negative cases, with a 3.45% accuracy. These results can be analyzed in terms of precision and recall metrics. Precision measures the proportion of true positive predictions to total positive predictions, focusing on decreasing the occurrence of false positive predictions. Recall, however, measures the proportion of true positive predictions to total predictions and focuses on minimizing false negatives. The precision score for "healthy" cases was 0.96, but for "unhealthy" cases, it was 0.88. However, the recall score for "healthy" cases was 0.87, but for "unhealthy" cases, it was 0.97. These results illustrate how the model prioritized minimizing false negative predictions over isolating false positive cases. In clinical settings, a false negative diagnosis may pose a more significant danger to patients, leading to delayed treatment and intervention. These findings demonstrate how the MobileNet algorithm has the potential to be deployed in a clinical setting to detect structural abnormalities in hippocampal MRIs.

VI. Conclusion and Future Work

This study underscores the potential of integrated deep learning models as powerful tools for identifying hippocampal abnormalities associated with TLE. By achieving an IoU of 0.902 with the U-Net segmentation model and a test accuracy of 0.94 with the MobileNet classifier, the study highlights the strong performance of both architectures in accurately localizing unhealthy hippocampal tissue. These results, supported by IoU, accuracy and confusion matrix metrics, reveal the capability of the combined architecture to minimize misdiagnoses and provide consistent and unbiased analysis suitable for clinical decision support. However, several opportunities remain for future research. The study utilized a relatively limited data set with slice-level labels, which can constrain the generalizability of the model across diverse patient populations and imaging conditions. A dataset containing volumetric scans from a variety of centers could reduce the risk of bias. Furthermore, a broader exploration of hyperparameters, such as optimizer selection, learning rate schedules, and augmentation strategies, as well as the incorporation of 3D volumetric data, could significantly enhance model robustness and diagnostic precision. Additionally, expanding the number of algorithms that were trained and tested upon would further enhance understanding of which models are most effective for segmentation and classification. Expanding the framework to encompass additional epilepsy subtypes or other neurological disorders may further improve versatility.

Overall, this study presents a promising approach to advancing TLE diagnostics through the integration of deep learning into neuroimaging workflows. By enabling efficient, accurate, and scalable detection of hippocampal abnormalities, the proposed models can assist clinicians in early identification and assessment of epilepsy-related pathology. Future advancements, like training the models from a larger variety of data sources, validation by industry professionals, and rigorous evaluation in real-world clinical environments, will be pivotal in shaping the future of deep learning-driven diagnostic imaging.

References

[1] A. Chaouch, A. Ben Abdallah, R. Ayari, M. H. Bedoui, and M. Aissi, "Dl segmentation and 3d assessment of hippocampal atrophy in mesial temporal lobe epilepsy", in *2022 6th International Conference on Advanced Technologies for Signal and Image Processing (ATSIP)*, 2022, pp. 1–6. DOI: 10.1109/ATSIP55956.2022.9805895.

[2] E. Kadhim, S. Al-Jumaili, and O. N. Uçan, "Diagnosis of epileptic seizures and hypoxic-ischemic encephalopathy using artificial intelligence based on eeg signal: A review", in *2024 8th International Symposium on Innovative Approaches in Smart Technologies (ISAS)*, 2024, pp. 1–6. DOI: 10.1109/ISAS64331.2024.10845267.

[3] C. R. Arias, R. M. Durón, and A. V. Delgado-Escueta, "Identification of new epilepsy syndromes using machine learning", in *2019 IEEE 39th Central America and Panama Convention (CONCAPAN XXXIX)*, 2019, pp. 1–4. DOI: 10.1109/CONCAPANXXXIX47272.2019.8977043.

[4] E. Tramoni-Negre, I. Lambert, F. Bartolomei, and O. Felician, "Long-term memory deficits in temporal lobe epilepsy", *Revue Neurologique*, vol. 173, no. 7-8, pp. 490–497, Jul. 2017, Epub 2017 Aug 31. DOI: 10.1016/j.neurol.2017.06.011.

[5] D. Sone et al., "Contribution of the μ-opioid receptor system to affective disorders in temporal lobe epilepsy: A bidirectional relationship?", *Epilepsia*, vol. 64, no. 2, pp. 420–429, 2022. DOI: 10.1111/epi.17463.

[6] K. D. Laxer et al., "The consequences of refractory epilepsy and its treatment", *Epilepsy & Behavior*, vol. 37, pp. 59–70, Aug. 2014, Open Access article under CC BY-NC-ND license. Review article., ISSN: 1525-5050. DOI: 10.1016/j.yebeh.2014.05.031.

[7] R. K. Sadangi, B. Patra, R. Maity, N. P. Maity, and B. Charan Sutar, "A review on the diagnosis of epilepsy using ml", in *2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, 2023, pp. 1–6. DOI: 10.1109/ICCCNT56998.2023.10306863.

[8] N. Alessi, P. Perucca, and A. M. McIntosh, "Missed, mistaken, stalled: Identifying components of delay to diagnosis in epilepsy", *Epilepsia*, vol. 62, no. 7, pp. 1494–1504, 2021. DOI: 10.1111/epi.16929.

[9] S. J. Czuczwar, Ed., *Epilepsy*. Brisbane, Australia: Exon Publications, Apr. 2022, Open Access. Available from NCBI Bookshelf, ISBN: 978-0-6453320-4-9. DOI: 10.36255/exon-publications-epilepsy.

[10] D. Villamizar-Torres, A. C. Cepeda Trillos, and A. Vargas-Moreno, "Mesial temporal sclerosis and epilepsy: A narrative review", *Acta Epileptologica*, vol. 6, no. 1, p. 28, Sep. 2024, Open Access, ISSN: 2524-5333. DOI: 10.1186/s42494-024-00172-5.

[11] M. Thom, "Review: Hippocampal sclerosis in epilepsy: A neuropathology review", *Neuropathology and Applied Neurobiology*, vol. 40, no. 5, pp. 520–543, Aug. 2014, Open Access. Highly cited: 428+ citations, ISSN: 0305-1846. DOI: 10.1111/nan.12150.

[12] R. S. S. Widodo, I. K. E. Purnama, and R. F. Rachmadi, "Volumetric hippocampus segmentation using 3d u-net based on transfer learning", in *2024 IEEE International Conference on Computational Intelligence and Virtual Environments for Measurement Systems and Applications (CIVEMSA)*, 2024, pp. 1–6. DOI: 10.1109/CIVEMSA58715.2024.10586572.

[13] E. H. Middlebrooks et al., "Radiologic classification of hippocampal sclerosis in epilepsy", *American Journal of Neuroradiology*, Feb. 2024, Published online ahead of print. DOI: 10.3174/ajnr.A8214.

[14] J. Jiang et al., "Automated detection of hippocampal sclerosis using real-world clinical mri images", *Computers in Biology and Medicine*, vol. 162, p. 107 457, 2023. DOI: 10.1016/j.compbiomed.2023.107457.

[15] A. Chaouch et al., "Deep learning brain mri segmentation and 3d reconstruction: Evaluation of hippocampal atrophy in mesial temporal lobe epilepsy", in *Good Practices and New Perspectives in Information Systems and Technologies (WorldCIST 2024)*, ser. Lecture Notes in Networks and Systems, vol. 986, Springer, Cham, 2024, pp. 243–253. DOI: 10.1007/978-3-031-60218-4_22.

[16] A. J. Chang, R. Roth, E. Bougioukli, et al., "Mri-based deep learning can discriminate between temporal lobe epilepsy, alzheimer's disease and healthy controls", *Communications Medicine*, vol. 3, no. 1, pp. 1–13, 2023. DOI: 10.1038/s43856-023-00262-4.

[17] J. Jiang et al., "Automated detection of hippocampal sclerosis using real-world clinical mri images", *Frontiers in Neuroscience*, vol. 17, p. 1 180 679, 2023. DOI: 10.3389/fnins.2023.1180679.

[18] K. Jafari-Khouzani, K. Elisevich, S. Patel, and H. Soltanian-Zadeh, *Mri dataset for hippocampus segmentation (hippseg_2011)*, https://www.nitrc.org/projects/hippseg_2011/, Accessed: 2025-11-29, 2017.

[19] O. F. Kaneko, N. J. Fischbein, J. Rosenberg, M. Wintermark, and M. M. Zeineh, "The "white gray sign" identifies the central sulcus on 3t high-resolution t1-weighted images", *American Journal of Neuroradiology*, vol. 38, no. 2, pp. 276–280, 2017. DOI: 10.3174/ajnr.A5014.

[20] A. Sudhyadhom, I. U. Haq, K. D. Foote, M. S. Okun, and F. J. Bova, "A high resolution and high contrast mri for differentiation of subcortical structures for dbs targeting: The fast gray matter acquisition t1 inversion recovery (fgatir)", *NeuroImage*, vol. 47, T44–T52, 2009, International Brain Mapping Intraoperative Surgical Planning Society (IBMISPS), ISSN: 1053-8119. DOI: https://doi.org/10.1016/j.neuroimage.2009.04.018.

[21] X. Pei et al., "Robustness of machine learning to color, size change, normalization, and image enhancement on micrograph datasets with large sample differences", *Materials & Design*, vol. 232, p. 112 086, 2023. DOI: 10.1016/j.matdes.2023.112086.

[22] S. Guo and L. Li, "Medical image segmentation based on 3d u-net", in *2020 IEEE 5th International Conference on Signal and Image Processing (ICSIP)*, IEEE, 2020, pp. 279–283. DOI: 10.1109/ICSIP49896.2020.9277322.

[23] N. Siddique, P. Sidike, C. Elkin, and V. Devabhaktuni, "U-net and its variants for medical image segmentation: A review of theory and applications", *IEEE Access*, vol. 9, pp. 82 031–82 057, 2021, arXiv preprint arXiv:2011.01118. DOI: 10.1109/ACCESS.2021.3086020.

[24] H. Lu, Y. She, J. Tie, and S. Xu, "Half-unet: A simplified u-net architecture for medical image segmentation", *Frontiers in Neuroinformatics*, vol. 16, p. 911 679, Jun. 2022. DOI: 10.3389/fninf.2022.911679.

[25] F. Turk and M. Kılıçaslan, "Lung image segmentation with improved u-net, v-net and seg-net techniques", *PeerJ Computer Science*, vol. 11, e2700, Feb. 2025. DOI: 10.7717/peerj-cs.2700.

[26] R. Ochoa-Ornelas, A. Gudiño-Ochoa, J. A. García-Rodríguez, and S. Uribe-Toscano, "Enhancing early lung cancer detection with mobilenet: A comprehensive transfer learning approach", *Franklin Open*, vol. 10, p. 100 222, 2025, Available online 22 January 2025. DOI: 10.1016/j.fraope.2025.100222.

[27] M. Prakash M and K. GN, "A comparative analysis of mobilenet and xception architecture for classification of endoscopy images", *International Journal of Computing and Artificial Intelligence*, vol. 3, no. 2, pp. 60–65, 2022, ISSN: 2707-658X. DOI: 10.33545/27076571.2022.v3.i2a.56.

Heart-Rate–Based Work–Rest Scheduling in Construction

Tala Yunis, Karim Zahed
Dept. of Industrial Engineering and Management
American University of Beirut
Beirut, Lebanon
e-mail: kz28@aub.edu.lb

Issam Srour
Dept. of Civil and Environmental Engineering
American University of Beirut
Beirut, Lebanon
e-mail: is04@aub.edu.lb

***Abstract*—Construction work is physically demanding and may lead to cumulative fatigue and increased injury risk. Wearable devices allow continuous monitoring of workers' physiological strain, but few studies translate such data into concrete work–rest decisions on active sites. This extended abstract summarizes an ongoing field study that uses wrist-based heart-rate monitoring to design and evaluate a physiology-guided work–rest framework for construction workers. Minute-level heart-rate data from 33 workers across three building sites were converted to Heart Rate Reserve (%HRR) and used to trigger candidate rest breaks and define a marginal recovery metric. Then, a Mixed-Integer Linear Programming (MILP) model selected breaks under a daily rest cap. Preliminary results show that targeted, physiology-informed breaks can substantially reduce exposure to high-strain minutes while remaining within realistic rest budgets.**

Keywords-wearable devices; construction workers; work–rest scheduling.

I. INTRODUCTION

Construction remains one of the most labor-intensive sectors, and project outcomes depend strongly on workers' health and productivity. In the United States, construction still accounts for about one-fifth of workplace fatalities [1], and cumulative fatigue is a key day-to-day contributor to this risk. A recent review highlights that fatigue in construction is linked to musculoskeletal disorders, performance errors, loss of motivation, and reduced productivity [2]. Beyond human costs, fatigue has been estimated to generate productivity losses on the order of $1,300–$3,100 per employee per year [3]. Experimental evidence also shows that fatigue degrades hazard recognition and risk perception, increasing the likelihood of accidents and near misses [4].

At the same time, modern sites are increasingly digitized, yet schedules and break policies rarely adapt to workers' physiological state. Wearable devices now offer a practical way to collect in-situ physiological data and support real-time safety monitoring [5]. The value of these data, however, lies in converting measurements into timely recovery opportunities. Recent construction studies have begun using wearable information to inform work–rest allocation and micro-breaks [6], but existing frameworks typically rely on simplified or simulated scenarios and do not fully exploit continuously collected, on-site physiology.

This study addresses that gap by using wrist-based Heart Rate (HR) monitoring on active construction sites to trigger and size breaks based on percentage heart-rate reserve and to allocate them through an optimization model that prioritizes breaks with the greatest marginal physiological recovery. **Figure 1** illustrates the system architecture of the proposed physiology-guided work–rest scheduling framework.

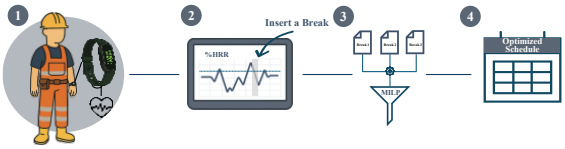


Figure 1. System architecture of the proposed physiology-guided work–rest scheduling framework.

The remainder of this paper is organized as follows. Section II describes the wearable monitoring setup and the proposed scheduling framework. Section III presents the optimization approach and preliminary results. Section IV addresses the discussion and ongoing work and concludes with final remarks.

II. METHODS: WEARABLE MONITORING AND SCHEDULING FRAMEWORK

A. Field Data Collection

Data were collected from 33 male construction workers across three active building sites in Beirut, Lebanon, covering residential and institutional projects with typical concrete and finishing trades. Workers wore a Fitbit Inspire 3 device on their wrist throughout their time on site over multiple weeks, and their heart rate was continuously recorded during working hours. For each worker, resting and age-predicted maximal heart rates were used to compute Heart Rate Reserve (HRR), and minute-level workload was expressed as %HRR [7], [8] using:

$$\%HRR = \frac{HR_{Work} - HR_{Resting}}{HR_{Maximum} - HR_{Resting}} \times 100\% \qquad (1)$$

Building on this, a rest-allowance (RA) formulation that converts time spent at elevated %HRR into required rest is used. Conceptually, a break is modeled as a period during which heart rate decreases from the pre-break level toward resting heart rate; as HR is assumed to relax toward rest over the break, each break minute "pays back" some of the accumulated physiological load. The rest allowance thus specifies how many break minutes are needed, given a

worker's recent %HRR profile, to bring exposure back toward a target strain level.

B. Identifying Candidate Breaks

For each worker-day, RA was evaluated on a 60-minute sliding window. Whenever RA is greater than 0, a candidate break was created with a duration equal to the RA output, then rounded to 5-minute increments. During each simulated break, %HRR was modeled to drop toward resting levels, and its benefit was summarized as Δ%HRR, defined as the average difference between the no-break and with-break %HRR over the break and its immediate recovery period.

C. Optimization of Break Selection

As it is not practical to allocate unlimited rest, we imposed a daily cap on non-lunch breaks (e.g., 30 minutes per worker per day). A Mixed-Integer Linear Programming (MILP) model then selects a subset and duration of candidate breaks subject to:

- the daily rest budget per worker,
- separation between breaks (minimum spacing), and
- alignment with the actual working window.

Break benefit is quantified using a marginal recovery metric, Δ%HRR. The MILP objective is to maximize the minute-weighted sum of Δ%HRR over all selected breaks, effectively prioritizing breaks that yield the greatest physiological relief per minute of rest.

III. Results: Impact on Physiological Strain

Across the three sites, the mean age was 29 years. At the individual level, worker mean %HRR ranged from about 17% to 41%, with most standard deviations <5% HRR, indicating modest within-worker variability but substantial differences between workers. Approximately 92% of worker-days triggered at least one candidate break, with the modal required rest lying between 5 and 15 minutes per day and an average of about 40 minutes of rest per worker-day when no cap was imposed. Under a 30-minute daily limit, the MILP retained roughly 80–85% of the total candidate rest time by shortening rather than discarding many breaks. This produced optimized break schedules that were more clustered and better aligned with sustained high-strain periods.

As per **Figure 2**, applying the optimized breaks reduced the share of workday minutes spent at or above 40% HRR while keeping the whole-day mean %HRR nearly unchanged, indicating that recovery was concentrated in high-strain windows rather than simply lowering overall workload.

Figure 2. Share of workday minutes with %HRR ≥40% before vs. after MILP breaks.

At the worker level, most workers experienced a reduction in minute-weighted %HRR in the break windows compared with the corresponding no-break trajectories. These results suggest that relatively small amounts of targeted rest, informed by wearable data and simple optimization, can meaningfully reshape the distribution of physiological strain across the workday.

IV. Discussion and Ongoing Work

The findings demonstrate that minute-level wrist-based heart-rate data from active construction sites can be transformed into actionable, physiology-guided work–rest schedules. By converting %HRR time series into rest-allowance–based candidate breaks and then selecting a feasible subset via an optimization model that maximizes marginal recovery under a daily rest cap, the framework provides a concrete path toward data-driven rest policies that reduce high-strain exposure without imposing excessive downtime. Among the limitations of the study are that the current implementation relies on a single physiological signal (heart rate) and site-specific data from three projects in one city. Future work should expand the dataset, explore richer sensing (environmental and activity measures), incorporate subjective recovery indicators, refine the recovery model, and test on-site implementation of the optimized schedules.

References

[1] U.S. Bureau of Labor Statistics, "Fatal falls in the construction industry in 2023," The Economics Daily. [Online]. Available: https://www.bls.gov/opub/ted/2025/fatal-falls-in-the-construction-industry-in-2023.htm [retrieved: January 2026].

[2] M. Tafazzoli, M. Namian, and A. Al-Bayati, "Workers' Fatigue in Construction Projects, Assessment, Detection, and Mitigation, A Review," presented at the Proceedings of 59th Annual Associated Schools of Construction International Conference, 2023, pp. 478–488. doi: 10.29007/rc7t.

[3] M. R. Rosekind et al., "The cost of poor sleep: workplace productivity loss and associated costs," *J Occup Environ Med*, vol. 52, no. 1, pp. 91–98, Jan. 2010, doi: 10.1097/JOM.0b013e3181c78c30.

[4] M. Namian, F. Taherpour, E. Ghiasvand, and Y. Turkan, "Insidious Safety Threat of Fatigue: Investigating Construction Workers' Risk of Accident Due to Fatigue," *J. Constr. Eng. Manag.*, vol. 147, no. 12, p. 04021162, Dec. 2021, doi: 10.1061/(ASCE)CO.1943-7862.0002180.

[5] W. Lee, K.-Y. Lin, E. Seto, and G. C. Migliaccio, "Wearable sensors for monitoring on-duty and off-duty worker physiological status and activities in construction," *Automation in Construction*, vol. 83, pp. 341–353, Nov. 2017, doi: 10.1016/j.autcon.2017.06.012.

[6] Z. Abuwarda, K. Mostafa, P. Morita, and T. Hegazy, "Optimizing Construction Work–Rest Schedules and Worker Reassignment Utilizing Wristband Physiological Data," *J. Constr. Eng. Manag.*, vol. 150, no. 11, p. 04024149, Nov. 2024, doi: 10.1061/JCEMD4.COENG-14884.

[7] K. Rodahl, *Physiology of work*. CRC Press, 1989.

[8] A. Vitalis, "The use of heart rate as the main predictor of the cost of work," in *Proceedings of the Inaugural Conference of the NZ Ergonomics Society, Auckland*, 1987, pp. 168–181.

Enhancing CineMRI Clinical Documentation by Detecting and Correcting Ambiguity with Large Language Models

Guillermo Villanueva Benito[1]
Biomedical Data Science Group
[1]Barcelona Institute for Global Health (ISGlobal)
Barcelona, Spain
e-mail: guillermo.villanueva@isglobal.org

Paula Petrone[1,2]
Digital Health Unit
[2]Barcelona Supercomputing Center (BSC-CNS)
Barcelona, Spain
e-mail: paula.petrone@bsc.es

Matias Calandrelli[3], Martín Descalzo[3], Sandra Pujadas[3], Juan Fernandez[3]
Cardiac Imaging Unit, Cardiology Department
[3]Hospital de la Santa Creu i Santa Pau
Barcelona, Spain
e-mail: {mcalandrelli | mdescalzo | sandrapujadas | juanfmav}@gmail.com

***Abstract*— Cardiac cine Magnetic Resonance Imaging (cineMRI) is the gold standard for assessing left-ventricular wall motion, yet interpretation varies and free-text reports often contain ambiguous terminology. We developed CineScribe, an Artificial Intelligence (AI)-assisted framework that structures diagnostic information, detects ambiguous clinical reports, and generates standardized cineMRI documentation. Using a dataset of 982 cineMRI studies, CineScribe achieved state-of-the-art report structuration performance (F1-score = 0.92). Associated confidence scores effectively identified ambiguous cases (F1-score = 0.82), most of which carried misdiagnosis risk. Generated reports from final review findings were rated complete and accurate in 78% of cases, supporting more consistent and reliable cineMRI documentation.**

Keywords- Cardiac cineMRI; LLMs; Documentation; ambiguity.

I. INTRODUCTION

Despite major advances in recent decades, cardiovascular diseases (CVDs) remain the leading cause of mortality worldwide, accounting for more than 17.9 million deaths annually [1]. Many diagnostic decisions in cardiology rely heavily on imaging, and cardiac cine Magnetic Resonance Imaging (cineMRI) is the gold standard for evaluating cardiac structure and function. CineMRI provides a comprehensive assessment of global and regional ventricular motion, primarily determined by the presence and severity of Regional Wall Motion Abnormalities (RWMAs). Yet its interpretation remains inherently variable, even among experienced cardiologists. This intrinsic variability is compounded by the lack of standardized reporting practices which lead to clinical reports being often written using imprecise and inconsistent terminology, increasing the risk of misdiagnosis [2]-[4]. Identifying ambiguous reports that may lead to miscommunication, detecting diagnostically complex cases requiring further expert consensus, and implementing standardized review and documentation processes are essential to ensure accurate diagnosis and appropriate downstream patient management.

The rest of the paper is organized as follows: Section II describes the objectives and provides an overview of the framework. Section III details the methodology, Section IV presents the main results, and Section V concludes with a summary and future research directions.

II. OBJECTIVES

To address the challenges presented in the previous section, we developed and validated an AI-assisted framework designed to facilitate more reliable and standardized cineMRI clinical documentation of left ventricular wall motion. The framework, illustrated in Figure 1, integrates three complementary components:

1) Automatic report structuration to extract and organize diagnostic information from free-text reports.
2) Detection of ambiguous or imprecise reports that might lead to misdiagnosis, thereby enabling targeted expert cineMRI review and consensus.
3) Standardized AI-assisted report generation from structured diagnostic findings to ensure clear, complete, and reproducible clinical documentation.

The model first interprets the original free-text cineMRI report, converts it into a structured representation visualized as a bullseye diagram, and assigns a global confidence score reflecting uncertainty in the extracted interpretation. Subsequently, experts reassess the corresponding cineMRI scans for flagged ambiguous reports and provide the final structured diagnostic bullseye. CineScribe then generates a standardized clinical report using the confirmed structured diagnosis and Left Ventricular Ejection Fraction (LVEF).

III. METHODS

We conducted a retrospective study including 982 cineMRI examinations from the Hospital de la Santa Creu i Sant Pau (Barcelona, Spain). Three board-certified cardiologists, with certification in cardiac Magnetic

Resonance Imaging (MRI), manually annotated the clinical reports, assigning a RWMA label to each myocardial segment [5] based on the findings described in the original text. For multi-expert (x3) annotated cases, report ambiguity was defined as inter-expert disagreement, with ambiguous reports corresponding to those that yielded more than one valid structured interpretation from medical experts.

We developed CineScribe, a fine-tuned lightweight Large Language Model (LLM) based on Llama3 [6], trained for both report-structuration and clinical report-generation tasks. The model's confidence score during the structuration task was validated as a quantitative proxy for report ambiguity. Reports were re-evaluated by expert cardiologists, who reviewed the original cineMRI videos to reassess the corresponding diagnostic findings. CineScribe was quantitatively benchmarked for report-structuration performance and prospectively evaluated on a stressed evaluation dataset composed exclusively of abnormal cases by expert cardiologists for its ability to detect ambiguous reports, flag cases at risk of misdiagnosis that require expert review, and for the clinical quality of its generated reports, following the QUEST framework [7].

IV. RESULTS

CineScribe achieved an F1-score of 0.92 (95% Confidence Interval (CI): 0.89 -- 0.94) in the report structuration task, demonstrating state-of-the-art performance comparable to GPT-5 (F1-score = 0.89; 95% CI: 0.85 -- 0.92). The model's confidence score demonstrated strong discriminative ability for detecting ambiguous reports, achieving an Area Under the Receiver Operating Characteristic Curve (ROC–AUC) of 0.76 on the evaluation dataset.

Subsequent expert cineMRI review showed that 71% of non-ambiguous reports were safe-to-follow, whereas 84% of ambiguous reports carried potential misdiagnosis risk, underscoring CineScribe's utility in flagging diagnostically complex cases that warrant targeted expert review.

Finally, standardized clinical reports generated by CineScribe from structured reviewed findings were rated both complete and accurate in 78% of cases by expert cardiologists.

V. CONCLUSIONS AND FUTURE WORK

CineScribe improves cineMRI interpretation by accurately structuring diagnostic information, flagging ambiguous reports that signal diagnostically complex cases, and producing clear and consistent clinical reports. Our findings suggest that ambiguous report language often arises in clinically challenging cases. From a clinical perspective, this has important implications as ambiguous reports can mask diagnostically difficult cases that would benefit from consensus expert review.

Among the limitations of this work is its reliance on data from a single institution, which may restrict the generalizability of the findings, as well as its motion-focused scope, which represents only a subset of full cineMRI clinical reporting. Future work should further explore human-in-the-loop approaches, such as the one proposed here, to improve report quality and support continued fine-tuning and enhancement of the model's capabilities. Further directions also include the integration of image-derived cineMRI features, as well as prospective deployment and validation in multi-center settings. Finally, continued research is needed to better characterize and address the intrinsic sources of variability in cineMRI assessment, particularly in diagnostically complex cases.

REFERENCES

[1] G. A. Mensah, G. A. Roth, and V. Fuster, "The global burden of cardiovascular diseases and risk factors: 2020 and beyond", Journal of the American College of Cardiology, vol. 74, no. 20, pp. 2529-2532, 2019.

[2] A. B. Rosenkrantz, M. Kiritsy, and S. Kim, "How "consistent" is "consistent"? A clinician-based assessment of the reliability of expressions used by radiologists to communicate diagnostic confidence.", Clinical Radiology, vol. 69, no. 7, pp. 745–749, 2014.

[3] J. M. Bosmans, J. J. Weyler, A. M. De Schepper and P. M. Parizel, "The radiology report as seen by radiologists and referring clinicians: results of the COVER and ROVER surveys.", Radiology, vol. 259, no. 1, pp. 184–195, 2011.

[4] W. Levinson, "Physician-patient communication: a key to malpractice prevention.", JAMA, vol. 272, no. 20, pp. 1619–1620, 1994.

[5] M. D. Cerqueira, et al., "Standardized myocardial segmentation and nomenclature for tomographic imaging of the heart: A statement for healthcare professionals from the Cardiac Imaging Committee of the Council on Clinical Cardiology of the American Heart Association", Circulation, vol. 105, no. 4, pp. 539–542, 2002.

[6] A. Grattafiori, et al., "The llama 3 herd of models.", arXiv preprint arXiv:2407.21783, 2024.

[7] K. K. Y. Ng, I. Matsuba, P. C. Zhang, "RAG in Health Care: A Novel Framework for Improving Communication and Decision-Making by Addressing LLM Limitations.", NEJM AI, vol. 2, no. 1, AIra2400380, 2024.

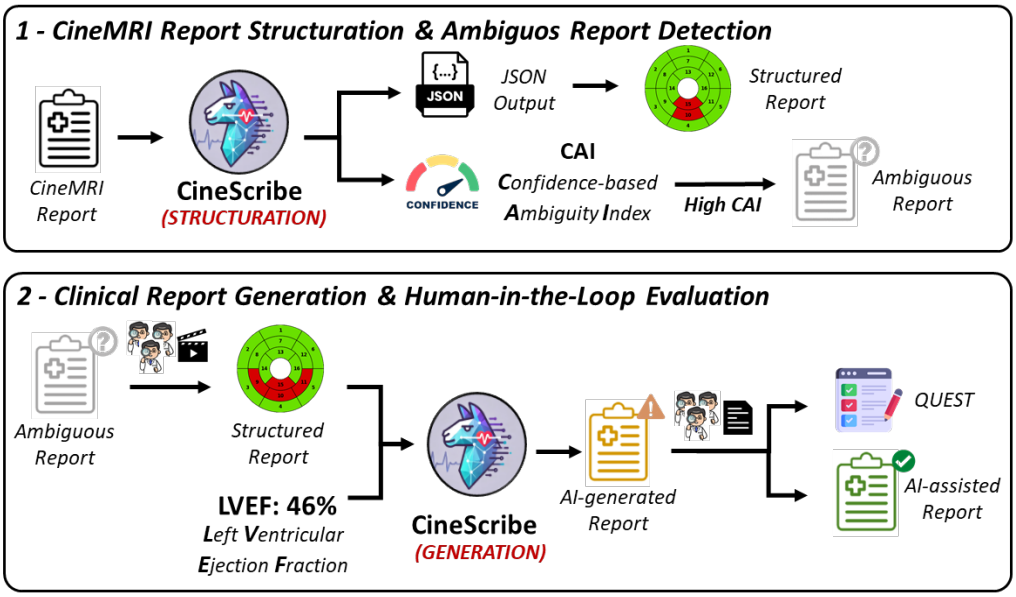


Figure 1. AI-assisted CineMRI Reporting Framework Overview: In the first stage, CineScribe structures clinical reports, flagging potentially ambiguous cases. In the second stage, experts reassess the corresponding cineMRI scans, providing the final structured diagnosis, which is then used by CineScribe to generate a standardized clinical report. Generated reports undergo expert review and are used to produce the final AI-assisted report.

Toward a Unified Public X-ray Dataset Integrating Multiple Databases to Advance Complex Fracture Analysis

Fatma Atitallah[1], Assem Thabet[1], Johannes C. Ayena[2,3], Neila Mezghani[2,3]

[1]Research Laboratory MACS, University of Gabès, Gabès, Tunisia
e-mail: fatma.atitallah@enig.u-gabes.tn, assem.thabet@issig.rnu.tn

[2]Applied Artificial Intelligence Institute (I2A), TELUQ University, Montreal, Quebec, Canada
[3]Centre de Recherche du Centre Hospitalier de l'Université de Montréal, Montreal, Quebec, Canada
e-mail: johannes.ayena@teluq.ca, neila.mezghani@teluq.ca

***Abstract*—Detecting and segmenting complex upper-limb fractures in X-ray imaging remains a persistent challenge due to subtle visual patterns, anatomical overlap, and heterogeneous acquisition conditions. While deep learning continues to advance fracture analysis, its progress is limited by fragmented public datasets that differ in format, annotation standards, clinical focus, and image quality. This paper addresses this limitation by unifying three publicly accessible X-ray datasets (FracAtlas, Bone Fracture Detection, and GRAZPEDWRI-DX) into a single, standardized, high-quality resource dedicated to upper-limb fracture research. The proposed integration pipeline includes rigorous dataset selection, annotation harmonization, preprocessing, normalization, and multi-stage quality control to ensure consistency and clinical reliability. The resulting dataset provides extensive anatomical diversity, a wide spectrum of fracture types, and verified segmentation masks suitable for training and benchmarking deep learning models. Although class imbalance and uneven fracture representation persist, the unified dataset establishes a valuable foundation for developing robust, generalizable, and clinically meaningful Artificial Intelligence (AI) systems for fracture detection, localization, and segmentation. Future extensions will focus on balancing fracture categories and evaluating state-of-the-art architectures on the proposed dataset.**

***Keywords*-Upper-limb fractures; X-ray imaging; Dataset harmonization; Medical image preprocessing; Fracture detection; Fracture segmentation; Artificial Intelligence; Dataset integration.**

I. Introduction

Precisely detecting and segmenting complex bone fractures in X-ray images remains an essential yet highly challenging task in medical imaging, due to the subtle nature of fractures, the frequent overlap of anatomical structures, and the considerable variation in image quality [1]. Upper-limb fractures, including those of the fingers, hands, wrists, forearms, and elbows, can be particularly challenging due to the complex anatomy of these areas, the small size of some bones, and the wide variety of fracture types, such as displaced, comminuted, or hairline fractures [2]. Errors or delays in diagnosing such fractures can lead to prolonged recovery periods, reduced functional mobility, and increased healthcare costs [3]. X-ray imaging continues to be the primary method for evaluating fractures due to its speed and accessibility [4][5]. Despite these benefits, interpreting X-rays accurately remains a research challenge because it depends largely on the clinician's expertise [1]. Studies have shown significant variability between observers, especially when evaluating subtle or rare fractures, highlighting the need for objective and standardized tools to support clinical decision-making [4]. Artificial Intelligence (AI) and deep learning have shown considerable promise in medical imaging, helping clinicians more reliably detect, locate, and segment fractures [6][7]. However, building effective AI models depends on access to large, diverse, and high-quality datasets [8][9]. At present, existing X-ray datasets are often fragmented, vary in format, and lack consistent annotations, which limits their practical use for model development [10]. This study proposes to address these limitations by integrating, standardizing, and harmonizing multiple publicly available X-ray datasets into a unified resource. In doing so, we create a comprehensive, clinically meaningful, and fully standardized X-ray dataset dedicated to upper-limb fracture analysis.

The novelty of this work lies in the systematic integration of three major publicly available datasets FracAtlas [10][16], Bone Fracture Detection (Kaggle) [17], and GRAZPEDWRI-DX [18] into a single, coherent resource. Publicly available datasets that are fully annotated for fracture segmentation are rare, and identifying datasets that simultaneously provide X-ray imaging, upper-limb coverage, pixel-level segmentation annotations, and comprehensive metadata remains a significant challenge. Among the publicly accessible resources, only these three datasets currently meet our selection criteria. While other fracture related datasets do exist, they are often limited to Computed Tomography (CT) imaging modalities or do not cover the upper limb, such as RSNA [11] and 5K+ CT Images on Fractured Limbs [12], or are primarily designed for classification tasks, such as the MURA dataset [0], rather than segmentation, making them unsuitable for the objectives of this work.

This scarcity highlights the uniqueness of such datasets and directly motivates the core objective of this study: the construction of a unified, balanced, and high-quality data distribution from heterogeneous public sources. To achieve this objective, the proposed approach harmonizes annotation formats, imaging resolutions, metadata structures, and preprocessing pipelines, while implementing a multi-stage quality control procedure to guarantee data consistency and reliability.

By creating a reliable and comprehensive dataset, this work establishes a strong foundation for the development of AI systems capable of supporting clinicians in accurate and efficient fracture diagnosis, improving patient care, and advancing medical imaging research for the detection and segmentation of complex bone fractures.

The paper is structured as follows: Section II details the methodology, introduces the datasets with their key strengths and limitations, and describes their preparation before fusion. Fusion, Standardization, and Normalization of Datasets covers the merging and preprocessing steps. Dataset Quality Assurance presents post-fusion evaluation, including automated and visual checks. Section III reports and discusses the results, key observations, and trends. Finally, Section IV concludes and outlines future research directions.

II. METHODS

Selecting an appropriate imaging modality is essential for analyzing complex or atypical bone fractures. Public datasets enable the integration of heterogeneous sources into a unified database for automated fracture detection and segmentation [13], ensuring clinical relevance and compatibility with machine learning workflows [14][15]. This study used three publicly accessible datasets, FracAtlas, the Kaggle Bone Fracture Detection dataset, and GRAZPEDWRI-DX (Table I), chosen for their availability, rich annotations, and strong focus on upper-limb fractures. All datasets underwent fusion, standardization, normalization, and quality control to ensure consistency.

The datasets FracAtlas, Bone Fracture Detection, and GRAZPEDWRI-DX were selected primarily based on their public accessibility and their suitability for fracture detection and segmentation tasks in X-ray imaging. The selection focused on datasets covering the upper-limb, which remains underrepresented in publicly available fracture datasets.

To develop robust and generalizable AI models, datasets representing diverse populations and clinical contexts were prioritized. FracAtlas, Kaggle Bone Fracture Detection, and GRAZPEDWRI-DX offer complementary anatomical coverage, detailed annotations, and demographic variability. Together, they form a cohesive resource for training and evaluating fracture detection and segmentation algorithms. Table I summarizes their characteristics, including population type, anatomical coverage, size, annotation formats, metadata, and geographic origin.

A. Datasets Preparation

Prior to dataset integration, the three selected datasets were manually examined. The FracAtlas dataset comprises only 717 fracture images (mixed upper and lower limp), while the other two datasets contain substantially more (as illustrated in Table I). To mitigate this imbalance, data augmentation was applied to the FracAtlas images after selecting only upper-limb samples tripling their number.

Additionally, 3,000 images were randomly selected from each of the GRAZPEDWRI-DX and Bone Fracture Detection datasets before preprocessing to reduce their size. This procedure ensured a more balanced selection across the three datasets prior to fusion.

1) **Data Preparation for FracAtlas:**
The FracAtlas dataset contains X-ray images with pixel-level segmentation masks for precise fracture detection. The dataset is structured into *Images* (Fractured and Non-fractured), *Annotations* (COCO JSON utilized in this study), and *Metadata* recorded in a CSV file. Let N_{init} denote the total number of annotated fractured images before filtering. We first filter the dataset to retain only upper-limb fractures (hand and shoulder) and exclude lower-limb regions (leg and hip). Formally, the filtering condition can be expressed as:

$$\text{UpperLimbFra} = \Big\{ i \;\Big|\; (\text{hand}_i = 1 \;\vee\; \text{shoulder}_i = 1) \Big\} \wedge (\text{leg}_i = 0 \;\wedge\; \text{hip}_i = 0) \wedge (\text{fractured}_i = 1) \tag{1}$$

where:

- i indexes each image in the dataset,
- $\vee$ (logical OR) indicates that at least one of the conditions is true,
- $\wedge$ (logical AND) indicates that all conditions must be true simultaneously.

This defines *UpperLimbFra* as the set of images showing fractures in the upper limbs (hand or shoulder) while excluding any fractures in the legs or hips.
Each image I_i is associated with a binary mask M_i, where fracture pixels are assigned a value of 255 and background pixels are set to 0 [19], as formally defined by:

$$M_i(x,y) = \begin{cases} 255, & \text{if the pixel } (x,y) \\ & \text{belongs to a fracture,} \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

Applying the filter defined in (1) reduced the dataset to N_{fil} (number of filtered upper-limb fracture images). This filtered set restricts that only relevant images containing fractures in the upper extremities (hand or shoulder) and excluding lower-limb involvement are retained for subsequent analysis and augmentation.
Moreover, images without valid masks ($\sum M_i = 0$) were removed. Data augmentation, as described in [20], was applied to each filtered image. Let A denote the number of augmented images per original sample. In this study, the augmentation factor was fixed to $A = 3$, a choice that provides additional variability without disproportionately amplifying any particular subset of the dataset. This specific choice helps maintain a comparable number of images in the filtered FracAtlas subset relative to the other two datasets, preventing any single dataset from dominating the combined database and supporting a balanced representation for training. The augmentation techniques applied include:

TABLE I. COMPARATIVE SUMMARY OF THE THREE PUBLICLY AVAILABLE FRACTURE DATASETS, INCLUDING DATASET ORIGIN.

Parameter	FracAtlas	Kaggle Bone Fracture Detection	GRAZPEDWRI-DX
Population	Mixed adults & children	Unspecified	Pediatric only (0–19 yrs)
Dataset origin / Country	Bangladesh (3 hospitals)	N/A	Austria (University Hospital Graz)
Body regions	Upper & lower limbs	Mainly upper limbs	Wrist only
Dataset size	4,083 images; 717 with fractures	8296 images	20,327 images; 6,091 patients
Annotations	Classification labels, masks, bounding boxes	Bounding boxes, some masks	Masks, polygons, bounding boxes
Metadata	Available	None	Rich; includes clinical descriptors
Strengths	• Clinically reviewed • Subtle fractures included • Multi-hospital collection	• Wide upper-limb variety • Useful for detection models (YOLO)	• Large-scale pediatric dataset • Detailed annotations • Masks, polygons, bounding boxes
Limitations	• Underrepresented regions • Class imbalance	• Missing metadata • Unclear labeling • Integration harder	• Single-center dataset • Uneven class distribution

Note. yrs: years; N/A: Not Available

- Horizontal and vertical flips applied with probabilities $p_{\text{flipH}} = 0.5$ and $p_{\text{flipV}} = 0.3$.
- Random rotation by an angle $\theta \in \{-15°, -10°, -5°, 5°, 10°, 15°\}$.
- Brightness and contrast modification following $I_{\text{aug}} = \alpha I_{\text{orig}} + \beta$, with brightness factor $\alpha \in [0.8, 1.2]$ and contrast shift $\beta \in [-20, 20]$.
- Additive Gaussian noise: $I_{\text{aug}} = I_{\text{aug}} + \mathcal{N}(0, \sigma^2)$ with noise level $\sigma = 10$.

Here, I_{orig} denotes the original image, and I_{aug} its augmented version. The parameters α and β control brightness and contrast, respectively. The variable θ represents the randomly selected rotation angle, while p_{flipH} and p_{flipV} denote the probabilities of horizontal and vertical flipping. Finally, $\mathcal{N}(0, \sigma^2)$ refers to zero-mean Gaussian noise with variance σ^2.

2) Data Preparation for Bone Fracture Detection from Kaggle:

The Kaggle Bone Fracture dataset, provided in YOLOv8 format, was preprocessed to support both fracture detection and segmentation tasks. This dataset contains seven classes: 'elbow positive', 'fingers positive', 'forearm fracture', 'humerus fracture', 'humerus', 'shoulder fracture', and 'wrist positive', as reported in [17]. The preprocessing pipeline included generating binary masks, computing image-level statistics, and creating metadata for each image.

- **Binary Masks for Segmentation**

  For each image I_i and its corresponding annotations L_i, a binary mask M_i was generated via (2).

  Bounding boxes in YOLO format, with normalized coordinates (x_c, y_c, w, h), were converted to pixel coordinates as follows:

$$\begin{aligned} x_{\min} &= \max(0, x_c \cdot W - \frac{w \cdot W}{2}) \\ x_{\max} &= \min(W - 1, x_c \cdot W + \frac{w \cdot W}{2}) \\ y_{\min} &= \max(0, y_c \cdot H - \frac{h \cdot H}{2}) \\ y_{\max} &= \min(H - 1, y_c \cdot H + \frac{h \cdot H}{2}) \end{aligned} \tag{3}$$

  where W and H denote the image width and height. When available, polygon annotations were rasterized into the mask for more precise fracture localization as represented by:

$$M_i(x, y) = \begin{cases} 255, & (x, y) \in \text{polygon interior} \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

  All masks were then binarized as:

$$M_i(x, y) = \begin{cases} 255, & M_i(x, y) > 127 \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

  Images without valid annotations or with empty masks were excluded.

- **Metadata Creation**

  For each retained image, a metadata CSV file was generated containing:

  – image name I_i,
  – fracture class for each annotation c_j,
  – annotation coordinates (x_j, y_j),
  – approximate fracture area formally defined as:

$$A_j = (x_{\max} - x_{\min}) \times (y_{\max} - y_{\min}) \tag{6}$$

  Image-level statistics were computed:

  – Number of fractures: $n_i = \text{count}(c_j)$
  – Mean coordinates: $\bar{x}_i = \frac{1}{n_i} \sum_j x_j, \quad \bar{y}_i = \frac{1}{n_i} \sum_j y_j$
  – Mean area: $\bar{A}_i = \frac{1}{n_i} \sum_j A_j$

– Severity category:

$$\text{severity}(I_i) = \begin{cases} \text{mild}, & n_i = 1 \\ \text{multiple}, & n_i > 1 \end{cases} \quad (7)$$

- **Visualization and Statistics**
  The percentage distribution of fracture types was computed as:

$$p_k = \frac{\text{count}(c_j = k)}{N_{\text{final}}} \times 100 \quad (8)$$

where k denotes a possible fracture class value, and N_{final} is the final number of image–mask pairs after preprocessing.

3) **Data Preparation for GRAZPEDWRI-DX:** For preprocessing, a subset of images N_{init} was selected from the GRAZPEDWRI-DX dataset part 1 to maintain balance when merging with other datasets. A fixed random seed allowed reproducible results [4][18].
   The preprocessing pipeline included the following steps:
   - Defining input paths for raw images and Pascal VOC annotations, and creating output folders for preprocessed images and masks.
   - Parsing each annotation to extract polygons or bounding boxes, which were used to generate precise binary masks (0 = background, 255 = fracture).
   - Clamping and resizing masks when necessary to match the image dimensions.
   - Saving the masks alongside images with harmonized filenames.
   - Recording metadata, including filenames, mask names, and classes, in a CSV file.

B. Fusion, Standardization, and Normalization of Datasets

In order to build a comprehensive and consistent dataset, multiple fracture datasets were fused into a single collection while preserving demographic labels (Adult, Child, or Mixed—Adult and Child). Each image I_i and its corresponding mask M_i were resized (resampled) to a standard resolution of 512×512 pixels [21], resulting in I_i^{res} and M_i^{res}, the resized image and mask, respectively. This process is formally defined by:

$$\begin{aligned} I_i^{\text{res}} &= \text{Res}(I_i, 512, 512, \text{interp} = \text{cubic}), \\ M_i^{\text{res}} &= \text{Res}(M_i, 512, 512, \text{interp} = \text{nearest}) \end{aligned} \quad (9)$$

Cubic interpolation preserves image details [22], while nearest-neighbor interpolation maintains mask accuracy [23]. X-ray images often exhibit uneven brightness or shading. To correct this, the N4ITK Bias Field Correction was applied to the luminance channel of each image as formally defined as:

$$\hat{I}_i = I_i^{\text{res}} \cdot B_i^{-1}, \quad B_i = \text{BiasField}(I_i^{\text{res}}) \quad (10)$$

where B_i is the estimated bias field. After bias correction, intensity normalization was performed to scale the pixel values to a consistent range suitable for machine learning, as represented by:

$$I_i^{\text{norm}} = \frac{\hat{I}_i - \mu_{\hat{I}_i}}{\sigma_{\hat{I}_i}} \cdot 255 \quad (11)$$

Here, $\mu_{\hat{I}_i}$ and $\sigma_{\hat{I}_i}$ are the mean and standard deviation of the bias-corrected image $\hat{I}_i$. This normalization ensures consistent brightness and contrast across all images. Preprocessing was performed in batches with parallel processing for efficiency. Metadata, such as dataset source, demographic category, and mask information, were saved alongside the images.

The resulting fused $\mathcal{D}_{\text{fused}}$ and normalized dataset I_i^{norm} is formally defined as:

$$\mathcal{D}_{\text{fused}} = \{(I_i^{\text{norm}}, M_i^{\text{bin}}, \text{metadata}_i)\}_{i=1}^{N_{\text{total}}} \quad (12)$$

where N_{total} denotes the total number of image–mask pairs in the fused dataset, computed as the sum of all filtered and augmented images across the three constituent datasets.

C. Dataset Quality Assurance

After merging the datasets, we ensured consistency and reliability through automated verification, visual inspection, and quantitative analysis. We implemented a standardized integration pipeline that included annotation harmonization, intensity normalization, and multi-stage quality control. These steps were specifically designed to minimize potential data distortions during the merging process and to preserve all clinically relevant fracture features, despite variations in imaging protocols, anatomical coverage, and acquisition quality across the original datasets. This approach guarantees that the resulting unified dataset maintains both anatomical fidelity and suitability for training robust and generalizable deep learning models. Automated checks confirmed that each image had a readable, correctly sized, and non-empty mask, with any anomalies logged for review.

Additionally, a random subset of images was manually inspected by overlaying the fracture regions on the corresponding X-rays to verify accurate mask delineation. To quantify fracture coverage, we calculated the proportion of each image occupied by fractures using (13):

$$\text{Fracture\_\%}_i = \frac{\sum_{x,y} \mathbb{1}[M_i(x,y) > 127]}{\text{width}_i \times \text{height}_i} \times 100, \quad (13)$$

Here, M_i denotes the standardized binary mask of image I_i from its original dataset (before integration into the unified database), and $\mathbb{1}$ is the indicator function counting fracture-labeled pixels. Using these values, we computed standard descriptive statistics for each dataset, including minimum, maximum, mean, and median fracture coverage. To assess fracture complexity, we analyzed the morphology of each connected component in the mask. For a component C_j in M_i with area A_j, width w_j, height h_j, and perimeter P_j, its complexity was computed using (14):

TABLE II. UNIFIED SUMMARY OF THE THREE DATASETS: SIZES, FILTERING, AUGMENTATION, SURFACE STATISTICS, AND COMPLEXITY.

Dataset	N_{init}	**Filtering & Resulting Outputs**			**Surface Statistics (%)**				**Complexity**
		N_{fil}	A	N_{final}	**Min**	**Max**	**Mean**	**Median**	**Mean / Max**
FracAtlas	717	439	3	1,756	0.007	2.241	0.352	0.312	11.48 / 28.66
Kaggle	3,000	1,491	–	1,491	0.123	18.583	2.256	1.651	11.20 / 24.97
GRAZPEDWRI-DX	3,000	2,017	–	2,017	0.320	14.986	2.622	2.288	11.56 / 24.63

Notes: N_{init}: initial number of X-ray images selected in each dataset; N_{fil}: number of valid image–mask pairs remaining after filtering; A: number of augmentation operations applied (only FracAtlas was augmented with $A = 3$); N_{final}: final number of image–mask pairs after filtering and augmentation; *Min*, *Max*, *Mean*, *Median* (%): proportion of fracture pixels; *Mean / Max of Complexity*: average and maximum fracture complexity based on morphological descriptors (connected components, aspect ratio, compactness).

$$\text{Complexity}(C_j) = 0.5 \cdot \frac{\max(w_j, h_j)}{\min(w_j, h_j)} + 0.5 \cdot \frac{P_j^2}{A_j}. \quad (14)$$

The overall complexity of image I_i was then defined as the mean complexity across its n_i components through (15):

$$\text{Complexity}(I_i) = \frac{1}{n_i} \sum_{j=1}^{n_i} \text{Complexity}(C_j). \quad (15)$$

Finally, for each dataset, we computed the mean and maximum complexity across all images retained after preprocessing, using (16):

$$\begin{aligned} \text{Mean Complexity} &= \frac{1}{N_{\text{final}}} \sum_{i=1}^{N_{\text{final}}} \text{Complexity}(I_i), \\ \text{Max Complexity} &= \max_{i \le N_{\text{final}}} \text{Complexity}(I_i). \end{aligned} \quad (16)$$

These metrics provide a comprehensive description of fracture morphology by capturing size and shape irregularities, offering a clear overview of dataset properties for downstream analysis and model development.

III. RESULTS AND DISCUSSION

The filtering procedure defined in (1) reduced the initial FracAtlas dataset from $N_{\text{init}} = 717$ to $N_{\text{fil}} = 439$ upper-limb fracture images. To preserve class balance and dataset diversity, each retained image I_i and its corresponding mask M_i underwent three augmentation operations ($A = 3$), resulting in a final total of $N_{\text{final}} = 1{,}756$ aligned image-mask pairs, as summarized in Table II.

Figure 1 shows the distribution of fracture types after filtering and augmentation, highlighting the predominance of hand fractures and confirming the effective exclusion of leg and hip cases. This ensures the dataset remains representative of upper-limb fractures and suitable for subsequent training and evaluation.

Similarly, the Kaggle Bone Fracture dataset underwent the same preprocessing. Binary masks were generated via (5) and only valid image–mask pairs were retained. Out of 3,000 initially selected images, 1,491 pairs remained. Metadata were then created, and fracture type percentages calculated using (8) to support statistical analysis and task integration. Figure 2 illustrates class imbalance, particularly for wrist-positive and humerus cases.

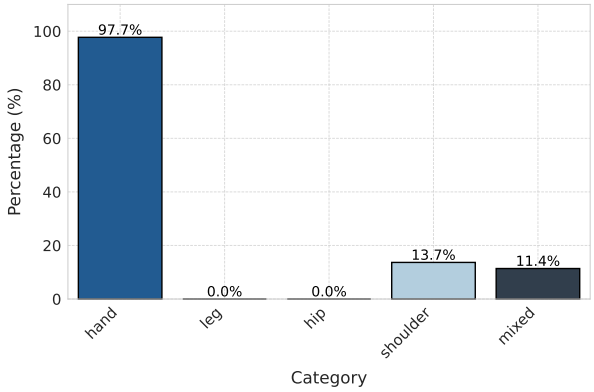


Figure 1. Class distribution in the selected FracAtlas subset.

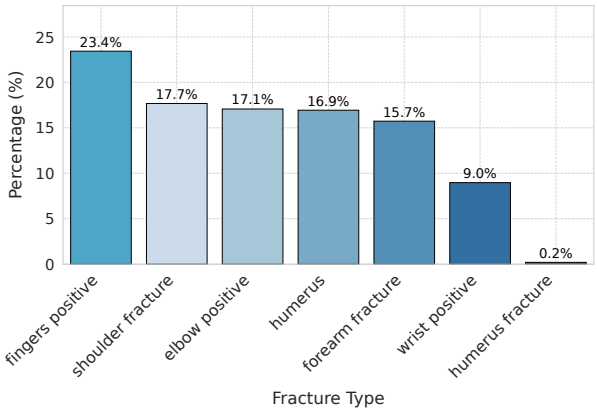


Figure 2. Class distribution in the selected Bone Fracture Detection subset.

The GRAZPEDWRI-DX dataset followed the same pipeline. From $N_{\text{init}} = 3,000$ images, 2,017 valid image–mask pairs were retained (Table II). The resulting subset, with harmonized images, masks, and metadata, is ready for integration with the other datasets, providing a robust training set for fracture detection and segmentation [4][7][18].

Merging these three datasets produced a unified, coherent X-ray collection for upper-limb fracture analysis. Despite differences in acquisition protocols, demographics, and annotation styles, preprocessing, normalization, and validation ensured alignment in image quality, mask formats, and metadata.

After resizing and normalization, summary statistics (Table II) and histograms (Figure 3) highlight variability and

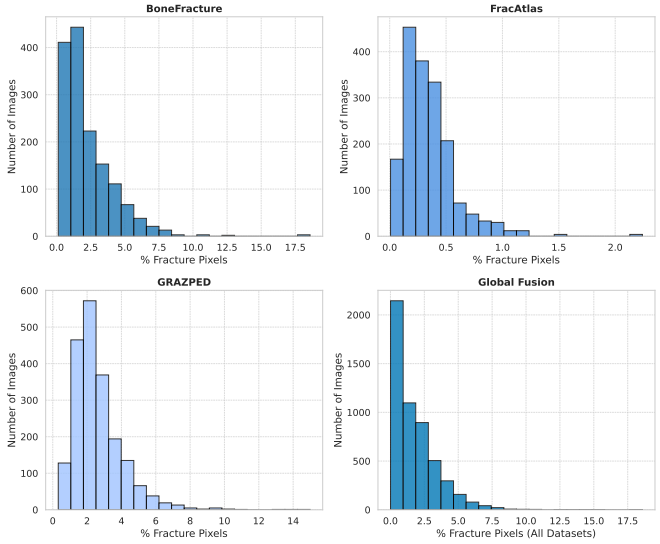


Figure 3. Distribution of fractured pixels across datasets.

potential outliers. Fractures were categorized by size: subtle (0–2% of pixels), medium (2–5%), and large (>5%). Bone fracture detection (Kaggle mainly contains small fractures, FracAtlas mostly subtle ones, and GRAZPED shows a balanced distribution of small to medium fractures, including a few extreme case (Figure 3). Overall, most images feature very small fractures, with larger ones increasingly rare. This distribution suggests a bias toward detecting minor cracks, emphasizing the importance of data augmentation, oversampling, or class-weighted training.

Fracture complexity was quantified using connected components, aspect ratio, and compactness, with the mean complexity across components defining each image's overall complexity. Complexity scores were computed for all images to enable cross-dataset comparisons, and Table II reports mean and maximum values. Notably, small fractures are not always simple; some remain regular while others, though subtle, show intricate shapes, multiple fragments, or irregular boundaries. A full assessment should therefore include fine, medium, and large cases to capture patterns important for robust detection.

The choice of $A = 3$ augmentations per image was subjective, as no data-driven method determined the optimal number. This is a limitation, and future work could test different augmentation levels to evaluate their impact on performance. Remaining limitations also include class imbalance, with subtle fractures being overrepresented, potentially biasing models toward minor fractures. Strategies, such as additional augmentation, synthetic image generation, or class-weighted training are recommended to mitigate this issue.

Moreover, it is important to emphasize that this work is part of an ongoing improvement process. At present, the study relies on publicly available datasets that are annotated for segmentation, contain upper limb X-ray images, and meet our quality standards. The over-representation of certain fracture types, such as wrist fractures, is partly due to their higher prevalence in the population and their greater availability in clinical records, as highlighted by Ye et al. [24]. In the future, if additional datasets meeting these requirements become accessible, they will be incorporated to further increase data diversity, reduce potential bias, and enhance the robustness and generalization capability of the proposed module.

Finally, both the datasets (before and after fusion) and the associated preprocessing, augmentation, and mask generation code are publicly available on the Synapse website, ensuring full reproducibility and enabling task-specific customization. This is made possible through Synapse, a secure platform for sharing, storing, and collaboratively managing biomedical and clinical research data, which allows researchers to access and reproduce datasets reliably [25].

IV. Conclusion and Future Work

This paper introduces a unified and standardized X-ray dataset focused on upper-limb fractures, created by merging FracAtlas, Bone Fracture Detection (Kaggle), and GRAZPEDWRI-DX. The integration pipeline harmonizes image formats, annotations, preprocessing steps, and quality assurance measures, resulting in a consistent and clinically relevant resource suitable for training and evaluating deep learning models. The dataset offers broad anatomical diversity and detailed segmentation masks, supporting research in fracture detection, localization, and segmentation. Remaining challenges include class imbalance and limited representation of complex fractures, which will be addressed through dataset expansion and targeted sampling strategies. Future work will benchmark state-of-the-art architectures and explore clinical deployment potential. Overall, this dataset provides a solid foundation for developing more reliable and generalizable AI-based fracture assessment systems.

Acknowledgment

The authors gratefully acknowledge the financial support provided by Mitacs through the Mitacs Globalink program, which enabled this research internship, as well as by the Fondation Docteur Sadok Besrour.

Data Availability Statement

The unified dataset generated in this study are publicly available in synapse at the following DOI: https://doi.org/10.7303/syn71834100

References

[1] Z. Su, A. Adam, M. F. Nasrudin, M. Ayob, and G. Punganan, "Skeletal fracture detection with deep learning: A comprehensive review," *Diagnostics*, vol. 13, no. 20, p. 3245, 2023. [Online]. Available: https://www.mdpi.com/2075-4418/13/20/3245.

[2] O. Khalilzadeh, C. Canella, and L. M. Fayad, "Wrist and hand," in *Musculoskeletal Diseases 2021–2024: Diagnostic Imaging*, ser. IDKD Springer Series, J. Hodler, R. A. Kubik-Huch, and G. K. von Schulthess, Eds., Springer, Cham, Switzerland, 2021, pp. 41–55. DOI: 10.1007/978-3-030-71281-5_4. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-030-71281-5_4.

[3] K. Liu et al., "Time to surgical management of distal radius fractures: Effects on health care utilization and functional outcomes," *Canadian Journal of Surgery*, vol. 67, no. 4, E286–E294, 2024. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC11233172/.

[4] A. Hassan, I. Afzaal, N. Muneeb, A. Batool, and H. Noor, "Ai-based applied innovation for fracture detection in x-rays using custom cnn and transfer learning models," *arXiv e-prints*, arXiv:2509, 2025. [Online]. Available: hhttps://arxiv.org/html/2509.06228v1.

[5] X. Zhai et al., "High-speed x-ray visualization of dynamic crack initiation and propagation in bone," *Acta Biomaterialia*, vol. 90, pp. 278–286, 2019. [Online]. Available: https://doi.org/10.1016/j.actbio.2019.03.045.

[6] J. Ju, Z. Qu, H. Qing, Y. Ding, and L. Peng, "Evaluation of artificial intelligence-based diagnosis for facial fractures, advantages compared with conventional imaging diagnosis: A systematic review and meta-analysis," *BMC Musculoskeletal Disorders*, vol. 26, no. 1, pp. 1–19, 2025. [Online]. Available: https://link.springer.com/article/10.1186/s12891-025-08842-2.

[7] T. Aldhyani et al., "Diagnosis and detection of bone fracture in radiographic images using deep learning approaches," *Frontiers in Medicine*, vol. 11, p. 1 506 686, 2025. [Online]. Available: https://doi.org/10.3389/fmed.2024.1506686.

[8] J. Li et al., "A systematic collection of medical image datasets for deep learning," *ACM Computing Surveys*, vol. 56, no. 5, pp. 1–51, 2023. [Online]. Available: https://dl.acm.org/doi/full/10.1145/3615862.

[9] B. Boecking et al., "Making the most of text semantics to improve biomedical vision–language processing," in *European conference on computer vision*, Springer, 2022, pp. 1–21. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-031-20059-5_1.

[10] I. Abedeen et al., "Fracatlas: A dataset for fracture classification, localization and segmentation of musculoskeletal radiographs," *Scientific data*, vol. 10, no. 1, p. 521, 2023. [Online]. Available: https://www.nature.com/articles/s41597-023-02432-4.

[16] M. H. Tasin, *Fracatlas original dataset*, [retrieved: January, 2026], 2023. [Online]. Available: https://www.kaggle.com/datasets/mahmudulhasantasin/fracatlas-original-dataset.

[17] P. K. Darabi, *Bone fracture detection: Computer vision project*, [retrieved: January, 2026], 2024. [Online]. Available: https://www.kaggle.com/datasets/pkdarabi/bone-fracture-detection-computer-vision-project.

[18] E. Nagy, M. Janisch, F. Hržić, E. Sorantin, and S. Tschauner, "A pediatric wrist trauma x-ray dataset (grazpedwri-dx) for machine learning," *Scientific data*, vol. 9, no. 1, p. 222, 2022. [Online]. Available: https://www.nature.com/articles/s41597-022-01328-z.

[11] H. M. Lin et al., "The rsna cervical spine fracture ct dataset," *Radiology: Artificial Intelligence*, vol. 5, no. 5, e230034, 2023.

[12] D. D. Ruikar, K. Santosh, R. S. Hegadi, L. Rupnar, and V. A. Choudhary, "5k+ ct images on fractured limbs: A dataset for medical imaging research," *Journal of Medical Systems*, vol. 45, no. 4, p. 51, 2021.

[0] G. Duan, S. Zhang, Y. Shang, and W. Kong, "Research on x-ray diagnosis model of musculoskeletal diseases based on deep learning," *Applied Sciences*, vol. 14, no. 8, p. 3451, 2024.

[13] P. Rajpurkar et al., "Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning," *arXiv preprint arXiv:1711.05225*, 2017. [Online]. Available: https://doi.org/10.48550/arXiv.1711.05225.

[14] P. Lakhani and B. Sundaram, "Deep learning at chest radiography: Automated classification of pulmonary tuberculosis by using convolutional neural networks," *Radiology*, vol. 284, no. 2, pp. 574–582, 2017. [Online]. Available: https://pubs.rsna.org/doi/abs/10.1148/radiol.2017162326.

[15] I. M. Wani and S. Arora, "Computer-aided diagnosis systems for osteoporosis detection: A comprehensive survey," *Medical & biological engineering & computing*, vol. 58, no. 9, pp. 1873–1917, 2020. [Online]. Available: https://link.springer.com/article/10.1007/s11517-020-02171-3.

[19] A. A. Roque and H. Sebastian, "Chest x-ray pneumothorax segmentation using efficientnet-b4 transfer learning in a u-net architecture," *arXiv preprint arXiv:2509.03950*, 2025. [Online]. Available: https://doi.org/10.48550/arXiv.2509.03950.

[20] E. Goceri, "Medical image data augmentation: Techniques, comparisons and interpretations," *Artificial intelligence review*, vol. 56, no. 11, pp. 12 561–12 605, 2023. [Online]. Available: https://link.springer.com/article/10.1007/s10462-023-10453-z.

[21] M. Kutbi, "Artificial intelligence-based applications for bone fracture detection using medical images: A systematic review," *Diagnostics*, vol. 14, no. 17, p. 1879, 2024. [Online]. Available: https://www.mdpi.com/2075-4418/14/17/1879.

[22] J. A. Parker, R. V. Kenyon, and D. E. Troxel, "Comparison of interpolating methods for image resampling," *IEEE Transactions on medical imaging*, vol. 2, no. 1, pp. 31–39, 2007. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/4307610.

[23] R. Harini and C. Chandrasekar, "Image segmentation using nearest neighbor classifiers based on kernel formation for medical images," in *International conference on pattern recognition, informatics and medical engineering (prime-2012)*, IEEE, 2012, pp. 261–265. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/6208355.

[24] J. Ye, Q. Li, and J. Nie, "Prevalence, characteristics, and associated risk factors of wrist fractures in americans above 50: The cross-sectional nhanes study," *Frontiers in endocrinology*, vol. 13, p. 800 129, 2022.

[25] Sage Bionetworks, *Synapse: A collaborative data sharing and research platform*, https://sagebionetworks.org/platform/synapse, [retrieved: January, 2026], 2026.

Personalized Medicine Meets Artificial Intelligence A Systematic Literature Review

Xin Zhao
Department of Computer Science
Seattle University
Seattle, USA
email: xzhao1@seattleu.edu

Parth Reshamwala
Centene Corporation
Austin, USA
email: Parth.Reshamwala@centene.com

***Abstract*—Personalized medicine represents a paradigm shift from the traditional one-size-fits-all approach to a more customized healthcare model that seeks to provide the right treatment to the right patient at the right time. With the significant increase in the maturity of related technology over the past several years, Artificial Intelligence (AI) is enhancing personalized medicine, making it more precise, efficient, and accessible. To gain a deeper understanding of the interplay between personalized medicine and AI, we conducted a Systematic Literature Review (SLR), aiming to provide AI's pivotal role in advancing tailored healthcare solutions. Synthesizing insights from 36 articles, our investigation explores key factors of AI-driven personalized medicine, including techniques, tools, and effectiveness. Furthermore, we examine the limitations and lay out the potential future work inherent in this field. Our review organizes the state of the art in AI-based personalized medicine and paves the way for better patient outcomes and more effective healthcare delivery in clinical practice.**

Keywords-Personalized medicine; Artificial Intelligence; Systematic Literature Review.

I. Introduction

In recent years, the convergence of personalized medicine and AI has heralded a new era in healthcare. Personalized medicine, which tailors medical treatment to individual patients, marks a significant change from traditional one-size-fits-all approaches. This shift improves diagnostic accuracy, optimizes treatment, and enhances patient outcomes through large-scale data and algorithms. The role of AI in this transformation is pivotal, offering unprecedented capabilities to analyze diverse datasets, identify patterns, and generate insights that drive more effective and personalized healthcare solutions [1] [2].

In the realm of personalized medicine, AI holds immense potential. For instance, recent research has demonstrated that AI-driven approaches can significantly improve the identification of genetic variants and their associations with various diseases [3]. By leveraging AI algorithms, researchers can analyze vast genomic datasets to identify genetic variations associated with diseases and drug responses. This enables the development of targeted therapies tailored to an individual's genetic makeup, enhancing treatment efficacy while minimizing adverse effects [2]. Additionally, AI-driven approaches facilitate the interpretation of complex genetic data, providing clinicians with insight to guide treatment decisions and improve patient outcomes [4]. Despite challenges, such as data quality, interpretability of results, and ethical considerations surrounding genetic privacy, AI is still considered a revolution in human healthcare [5]. Addressing these challenges is essential to fully harness the power of AI in genomic medicine.

Furthermore, AI plays a crucial role in optimizing the selection and dosing regimens of medications for individual patients [1]. By analyzing patient data, including genetic profiles, medical history, and treatment responses, AI algorithms can predict the most effective drugs and dosages for specific individuals. This personalized approach not only improves treatment results but also reduces the likelihood of adverse reactions and side effects. This ongoing progress promises even greater advancements in personalized medicine, making AI a powerful tool to optimize patient care.

Despite the growing body of studies and successful applications, the integration of AI in personalized medicine faces numerous challenges. An outstanding challenge is the lack of a comprehensive synthesis of existing studies. This paper aims to fill this gap by thoroughly exploring AI's role in personalized medicine. We conducted a Systematic Literature Review (SLR) to synthesize existing research, analyzing how AI technology is applied in various aspects of personalized medicine. The overarching goal of our investigation is to delve into the intersection of AI and personalized medicine from three aspects: techniques, tools, and clinical specialties, providing a holistic view of the current state and future directions of AI in personalized medicine. To reach this goal, we propose the following Research Questions (RQs):

1) RQ 1: What techniques are adopted in the area of AI-based personalized medicine?
2) RQ 2: What tools are used when applying AI to personalized medicine?
3) RQ 3: What specialty is benefited from the intersection of AI and personalized medicine?

II. Related Work

Recent literature has explored the application of AI across various domains of personalized medicine. Gallo [3] and Suwinski [2] examined AI-driven approaches for genetic data interpretation and drug development. The integration of electronic health records into AI frameworks was studied by Abul-Husn and Kenny [6], highlighting their relevance for individualized care. Petrovic [7] and Nova [8] discussed advancements in deep learning and generative AI models, particularly in clinical decision-making and patient-specific data handling.

 ISBN: 978-1-68558-362-0

Methodological insights were provided by Dinesh [9], Gifari [10], and Ho [11], who explored algorithmic frameworks, AI infrastructure, and enabling technologies for precision medicine. While these studies contribute meaningful perspectives, many are either domain-centric or emphasize selected AI methods. This review complements prior efforts by synthesizing findings across AI techniques, implementation tools, and clinical specialties, to present a consolidated view of the field.

Overall, prior work shows that AI enables personalized medicine by improving diagnostics, treatment selection, and clinical decision-making using large and diverse healthcare data. However, the literature is fragmented across techniques, tools, and clinical specialties, limiting a unified understanding of research gaps. To address this, we conduct a Systematic Literature Review that consolidates and categorizes existing studies and highlights challenges and future directions for AI-based personalized medicine.

III. Methodology

This paper follows Kitchenham's approach [12] to develop our systematic literature review, to ensure a rigorous and comprehensive review process. Kitchenham's approach has been widely used in conducting literature reviews related to AI and medicine [13][14].

A. Query Search

To thoroughly investigate the role of AI in personalized medicine, we construct a detailed search query that includes key terms and relevant phrases. Below are the search terms and combinations that form the backbone of our search strategy:

- "AI in Personalized Medicine"
- "Machine Learning in Personalized Medicine"
- "Artificial Intelligence and Personalized Healthcare"
- "Machine Learning in Precision Medicine"
- "Big Data Analytics for Individualized Treatment"

To enhance the inclusiveness of our search, we also use Boolean operators in the query to create more search terms:

- "AI AND "Personalized Medicine"
- "Artificial Intelligence" AND "Precision Healthcare"
- "Machine Learning AND "Individualized Treatment"
- "Big Data Analytics" AND "Personalized Medicine"

We also include alternate spellings and synonyms to broaden the scope of our search. For example, we use "Artificial Intelligence" and "AI" interchangeably, and "Precision Medicine" is used as a synonym for "Personalized Medicine."

This systematic approach to formulating our query aims to capture all relevant information, providing a solid foundation for synthesizing insights and drawing meaningful conclusions about the role of AI in advancing personalized medicine.

B. Source Selection

The criteria for selecting sources for this research aim to include high-quality and diverse literature in the field of AI-based personalized medicine. We prioritized the PubMed and SpringerOpen databases due to their extensive collection of research in biomedical and health informatics, with PubMed being a primary source for articles in medicine and life sciences. ScienceDirect, a leading scientific database, was also included to cover a broad spectrum of scholarly articles in various disciplines relevant to health informatics.

We further expanded our selection by including Google Scholar, IEEE Xplore, ResearchGate, PLOS, Hindawi, Cureus, and MDPI to ensure comprehensive coverage of the research landscape. Five additional specialized platforms were included: the British Institute of Radiology, the American Society for Clinical Pharmacology and Therapeutics (ASCPT), the Pacific International Conference on Ecosystem Services (PICES), BioMed Central (BMC), and the Journal of Medical Internet Research (JMIR), to provide a multidisciplinary perspective.

In total, we selected 15 digital databases as sources for identifying relevant papers.

C. Selection Criteria

Inclusion Criteria: The inclusion criteria encompass studies that (1) directly address the integration of AI in personalized medicine. This includes research that focuses on tailoring medical treatment to individual characteristics, such as genomics, diagnostic methods, and treatment optimization. (2) Studies selected for inclusion involve the application of AI techniques, algorithms, and technologies in the healthcare context. Furthermore, preference is given to literature published in peer-reviewed journals and conference proceedings, ensuring the reliability and validity of the findings. (3) Studies involve clinical trials, experiments, or real-world applications, providing practical insights into the implementation of AI in personalized medicine.

Exclusion Criteria: The exclusion criteria were established to ensure the quality and relevance of the selected literature. Studies were excluded if they (1) did not involve the application of AI or machine learning in the context of personalized medicine. (2) lacked clinical relevance, such as those focused solely on theoretical AI models without healthcare applications. (3) were not published in peer-reviewed journals or reputable conference proceedings, including abstracts, editorials, and opinion pieces. (4) were published before the year 2015, as our focus is on recent advancements in the field. (5) demonstrated low methodological quality, including unclear research design, insufficient data transparency, lack of validation, or inadequate explanation of AI model performance. These criteria ensured that only high-quality, clinically applicable, and scientifically rigorous studies were included in the final review.

IV. Review Execution

In this section, we outline the execution of our review process, as discussed in Kitchenham's methodology [12].

A. Initial Search Results

We started with a broad search across 15 academic databases, as discussed in the previous section. This extensive search led to a large initial pool of 31,924 papers. The aim was to cover a wide range of studies discussing technological advancements, applications, challenges, and prospects in personalized medicine related to AI.

B. Preliminary Screening

The preliminary screening involved a preliminary review to eliminate papers that were not directly related to personalized medicine or AI, resulting in the exclusion of 28,450 papers. These exclusions were based on the relevance of the papers to our research scope, despite containing the keywords we defined, did not focus on AI-based approaches. This preliminary screening ensures that the remaining papers address the integration of AI in the context of personalized healthcare. After this preliminary screening, we have 3,474 papers left.

C. Shortlisting

Following the preliminary screening, we examined the titles and abstracts as a more refined set of criteria to further narrow down the selection. We excluded an additional 2,980 papers that, although related to personalized medicine, did not adequately address the application of AI or were not published in peer-reviewed sources.

Another 224 papers were excluded after assessing their introduction sections for alignment with our research focus. After shortlisting, we have 270 papers left.

D. Quality Assessment

At this stage, the remaining 270 papers were subjected to a structured quality assessment to ensure methodological rigor, clinical relevance, and scientific contribution. Each study was evaluated based on specific criteria, including the clarity of research design, transparency of data and methodology, validation of AI models, and reproducibility of results. Studies were excluded if they lacked peer review, presented incomplete or ambiguous findings, or failed to demonstrate a practical application of AI in personalized medicine. This assessment led to the exclusion of 212 papers that did not meet the required standards, resulting in a final set of 58 papers for analysis.

E. Data Extraction

Finally, the data extraction process was carried out on the remaining 58 papers. We focused on studies that provided practical insights into the implementation of AI in personalized medicine, particularly those involving clinical trials, experiments, or real-world applications. This rigorous selection process resulted in a final set of 36 papers. Figure 1 illustrates our paper selection process.

F. Included Studies

Table I lists all the studies included after the review process has been executed.

V. Results and Discussions

A. Techniques

The first research question that we examine is: What techniques are adopted in the area of AI-based personalized medicine? In this section, we present our findings related to this research question.

Our examination reveals that machine learning algorithms and neural networks are commonly utilized in several papers,

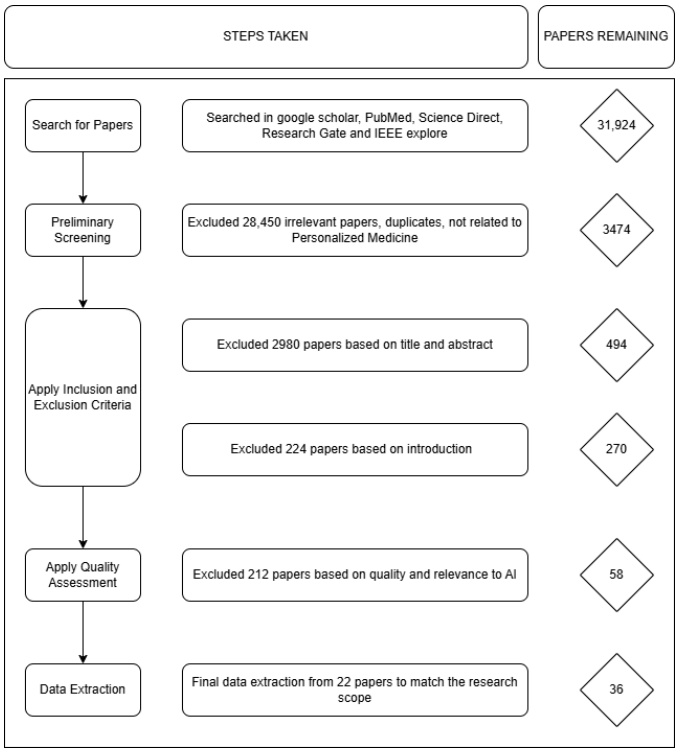


Figure 1. Selection Process of Selected Papers.

showcasing the broad applicability of these techniques. Natural Language Processing (NLP) and computational linguistics are applied in text-based data analysis, while others leverage techniques such as the Quadratic Phenotypic Optimization Platform (QPOP), Bayesian Decision Analysis (BDA), K-Nearest Neighbor (KNN), Support Vector Machine (SVM), and Artificial Neural Network (ANN). Some papers specifically emphasize the use of NLP for biostatistical datasets, while others explore Next-Generation Sequencing (NGS).

In addition, advanced deep learning techniques are widely adopted, including Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Generative Pre-trained Transformers (GPT). Novel approaches such as Evolutionary Enhanced Markov Clustering, AI-assisted Personalized Disease Burden (AI-PDB) assessment tools, and Real-Time Artificial Neural Networks further enrich the methodological landscape. These techniques reflect the interdisciplinary nature of data analysis in AI-driven personalized medicine, showcasing both depth and diversity in technological adoption.

In the discussion below, we divide the techniques in selected papers into two categories: Main Approaches which means the main implementation methods to achieve personalized medicine, and Supporting Techniques which refers to techniques used alongside main approaches in included studies.

1) Main Approaches: In this section, we discuss the main approaches identified from the list of papers in Table I. These approaches represent the key techniques and methodologies employed in the field of AI-based personalized medicine. The reference number in Table II corresponds to the paper numbers

TABLE I. LIST OF INCLUDED STUDIES

Paper #	Paper Title	Publication Year
1	The role of artificial intelligence in shaping future health planning [1]	2025
2	The revolutionary intersection of AI and healthcare: Embracing ChatGPT with caution [15]	2025
3	AI in healthcare: Transforming patient care and outcomes [16]	2025
4	Personalized medicine in urolithiasis: AI chatbot-assisted dietary management of oxalate for kidney stone prevention [17]	2024
5	Revolutionizing healthcare delivery: The role of AI and machine learning in personalized medicine and predictive analytics [18]	2024
6	Synergizing AI and healthcare: Pioneering advances in cancer medicine for personalized treatment [19]	2024
7	AI in healthcare: Revolutionizing diagnosis and treatment through machine learning [20]	2024
8	Artificial Intelligence for Personalized Genetics and New Drug Development: Benefits and Cautions [3]	2023
9	Artificial Intelligence in Brain Tumor Imaging: A Step toward Personalized Medicine [21]	2023
10	Artificial Intelligence based Personalized Predictive Survival Among Colorectal Cancer Patients [22]	2023
11	Artificial Intelligence in Healthcare and Education [23]	2023
12	ChatGPT-4 and the Global Burden of Disease Study: Advancing Personalized Healthcare Through Artificial Intelligence in Clinical and Translational Medicine [24]	2023
13	Deep Learning in Personalized Medicine: Advancements and Applications [7]	2023
14	Generative AI in Healthcare: Advancements in Electronic Health Records, Facilitating Medical Languages, and Personalized Patient Care [8]	2023
15	Personalized Dental Medicine, Artificial Intelligence, and Their Relevance for Dentomaxillofacial Imaging [25]	2023
16	Using ChatGPT to Predict the Future of Personalized Medicine [26]	2023
17	Effectiveness of artificial intelligence for personalized medicine in neoplasms: a systematic review [14]	2022
18	Artificial Intelligence based Algorithms Used for Solving Personalized Medicine Problems in Personalized Medicine Application [9]	2022
19	Artificial Intelligence Toward Personalized Medicine [10]	2021
20	Artificial Intelligence and Hybrid Imaging: The Best Match for Personalized Medicine in Oncology [27]	2020
21	CURATE.AI: Optimizing Personalized Medicine with Artificial Intelligence [28]	2020
22	Enabling Technologies for Personalized and Precision Medicine [11]	2020
23	Use of Machine Learning and Artificial Intelligence to Drive Personalized Medicine Approaches for Spine Care [29]	2020
24	Advancing Personalized Medicine Through the Application of Whole Exome Sequencing and Big Data Analytics [2]	2019
25	Personalized Medicine and the Power of Electronic Health Records [6]	2019
26	Personalized Medicine for Patients with COPD: Where Are We? [5]	2019
27	Personalized Medicine—Concepts, Technologies, and Applications in Inflammatory Skin Diseases [30]	2019
28	Predictive AI Models for Personalized Medicine[31]	2019
29	Using Deep Learning to Model the Hierarchical Structure and Function of a Cell [4]	2019
30	Futuristic Biosensors for Cardiac Health Care: An Artificial Intelligence Approach [32]	2018
31	Pivotal Trial of an Autonomous AI-Based Diagnostic System for Detection of Diabetic Retinopathy in Primary Care Offices [33]	2018
32	Scalable and Accurate Deep Learning with Electronic Health Records [34]	2018
33	The Role of Agent Technologies in Personalized Medicine [35]	2018
34	Integrated Genomic Medicine: A Paradigm for Rare Diseases and Beyond [36]	2017
35	Predicting Effects of Noncoding Variants with Deep Learning–Based Sequence Model [37]	2015
36	E-Health Towards Ecumenical Framework for Personalized Medicine via Decision Support System [38]	2010

 ISBN: 978-1-68558-362-0

TABLE II. MAIN TECHNIQUES AND APPLIED PAPERS

Main Techniques	Applied Papers
Convolutional Neural Network (CNN)	[3] [26] [27] [5] [30] [4] [36] [37]
Support Vector Machines (SVM)	[21] [38]
Deep Learning Neural Networks (DNN)	[25]
Generative Pre-trained Transformer (GPT)	[8]
Bayesian Decision Analysis (BDA)	[23]
Next-Generation Sequencing (NGS)	[25]
Big Data Analytics	[22]

in the table I. The same numbering convention applies to other tables in the paper. Table II summarizes these main techniques.

Several studies [27][11][29][33] referenced the application of Artificial Neural Networks (ANNs) without providing sufficient detail regarding specific subtypes such as Convolutional Neural Networks (CNNs), Deep Neural Networks (DNNs), or Transformer-based models (e.g., GPT). To maintain clarity and avoid redundancy, these studies have been excluded from Table II, but are discussed narratively within the section as part of the broader category of ANN-based approaches.

Although our review includes 36 papers, Table II highlights those that explicitly outline specific AI techniques in the context of personalized medicine. The remaining papers also contribute valuable insights such as conceptual frameworks, application-driven approaches, or interdisciplinary perspectives are discussed narrative within the section to provide a comprehensive understanding of the field.

Artificial Neural Networks (ANNs) and Their Subtypes: Deep Neural Networks (DNNs), including Convolutional Neural Networks (CNNs) and Transformer-based Language Models, fall under the broader category of Artificial Neural Networks (ANNs). CNNs are particularly useful for medical image analysis, extracting features from Magnetic Resonance Imaging (MRI), Computed Tomography (CT), and histopathological scans to assist in diagnostics. Transformer-based models, such as Generative Pre-trained Transformers (GPT), contribute to clinical text mining and genomic sequence interpretation by processing structured and unstructured medical data.

Other Machine Learning Techniques: Beyond neural networks, algorithms such as Support Vector Machines (SVMs) aid in disease classification and treatment response prediction, while Bayesian Decision Analysis (BDA) assists in uncertainty modeling for precision medicine.

Genomic Data Processing: AI plays a significant role in analyzing data generated by sequencing technologies such as Next-Generation Sequencing (NGS) and Whole Exome Sequencing (WES). While NGS and WES are not AI techniques themselves, they produce large-scale genomic datasets that require AI-driven variant classification, pathogenicity prediction, and functional annotation. These AI applications contribute to precision medicine by improving the interpretation of genetic variants.

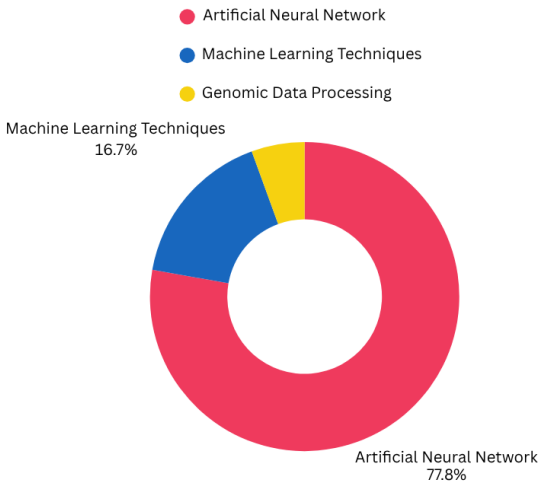


Figure 2. Main Approaches Discussed in Selected Papers.

Big Data Analytics: It was used alongside AI techniques to handle high-volume, heterogeneous datasets. These methods enabled scalable data preprocessing and integration from multiple sources such as electronic health records, genomic repositories, and wearable devices. While not an AI algorithm itself, big data analytics serves as a foundational layer that supports AI-driven personalized healthcare in large scale clinical studies.

Figure 2 summarizes the main techniques adopted in selected papers. The distribution of approaches was determined based on the frequency of each technique's use in the selected papers. As shown in Figure 2, deep learning methods dominate, reflecting a trend toward end-to-end modeling techniques in healthcare.

2) Supporting Techniques: In addition to core AI methodologies, various supporting techniques assist in data pre-processing, feature extraction, and predictive modeling in personalized medicine.

Data-Driven Approaches in Personalized Medicine: Natural Language Processing (NLP) techniques extract meaningful insights from Electronic Health Records (EHRs), clinical notes, and published literature, improving decision-making in personalized treatment plans. Machine learning-based clustering techniques and regression models enhance patient stratification for risk assessment and treatment optimization.

B. Tools

The investigation of AI applications in personalized medicine relies on a diverse array of advanced tools, each designed to enhance different aspects of data processing, analysis, and clinical decision-making. These tools are critical for leveraging large datasets to derive meaningful insights, optimize treatment plans, and improve patient outcomes. By employing various machine learning algorithms, data mining techniques, and specialized tools, researchers can tailor medical care to individual patients needs with higher precision. Table III provides an overview of the key tools used in several studies, illustrating the breadth and depth of technology integration in personalized medicine research.

 ISBN: 978-1-68558-362-0

TABLE III. TOOLS USED IN THE INVESTIGATION

Tools used in the investigation	Applied Papers
ChatGPT	[29] [5]
GATK	[3] [35]
CURATE.AI	[23]
Da Vinci Surgical System	[2]
AssistDent, Diagnocat, CranioCatch (Dentist)	[6]
Sugar.IQ	[32]
Messaging applications (Facebook, Messenger, WeChat	[35]
Golden Helix VarSeq	[3]

The tools listed in the table reflect a broad scope of AI applications in personalized medicine. These range from advanced data analysis platforms to clinical robotics, each serving specialized roles. For instance, genomic analysis software like GATK and Golden Helix VarSeq enable high throughput interpretation of genetic variants, directly supporting precision medicine.

Similarly, AI-driven decision platforms such as CURATE.AI optimize drug dosing tailored to individual patient responses, while conversational AI tools, such as ChatGPT enhances data synthesis and patient interaction. The tools also include domain specific systems like the Da Vinci Surgical System provides robotic precision in surgery, and applications like Sugar.IQ leverage AI for personalized diabetes management. Even everyday platforms such as messaging applications (e.g. Facebook Messenger, WeChat) and diagnostic aids in dentistry (AssistDent, Diagnocat, CranioCatch) have been used, underscoring how AI technologies in disparate domains converge to improve patient-specific outcomes.

C. Specialty

To comprehend the application of AI in personalized medicine, it is essential to understand the medical specialties included in our surveyed papers. Identifying these specialties provides insights into how those technologies are used to help patients in various fields. We can better understand how data-driven approaches improve diagnostic accuracy and enable tailored treatments, thereby supporting clinicians in making precise, patient-specific decisions.

Therefore, the next research question we delve into is what specialties our surveyed papers focus on. Table IV shows the specialties investigated by our included studies.

The reviewed studies span a wide range of medical specialties, reflecting the wide applicability of AI in personalized medicine. Oncology emerged as the field that is the most frequently addressed, indicating a strong focus on AI for cancer diagnosis, treatment planning, and optimization of patient-specific therapy. Genomics was another dominant area, underscoring the central role of AI methods in interpreting genetic data and guiding precision medicine initiatives.

TABLE IV. SPECIALTIES DISCUSSED IN INCLUDED PAPERS

Medical Specialties	Applied Papers
Oncology	[15] [16] [17] [18] [19] [21] [22] [8] [9] [10] [27] [28] [11] [29] [6] [5] [30] [4] [33] [35]
Genomics	[15] [22] [8] [25] [26] [29] [30] [32] [33] [34] [35] [36]
Cardiology	[22] [25] [10] [28] [29] [37]
Radiology	[17] [18] [8] [10] [29] [2] [4]
Neurology	[10] [28] [2]
Dermatology	[18] [32]
Ophthalmology	[18]
Pathology	[18] [4]
Dentistry	[24]
Endocrinology	[28]
Gastroenterology	[28]
Hematology	[18]
Immunology	[32]
Nephrology	[28] [17]
Orthodontics	[24]
Orthopedics	[10]
Pharmacology	[15]
Surgery	[10]
Health Policy & Public Health	[1]
Health Informatics	[16]
Urology	[17]
Radiology	[20]

Beyond oncology and genomics, other specialties were also represented. Fields such as cardiology and radiology had moderate coverage, highlighting the use of AI in personalized cardiac risk assessment and imaging-based diagnostics. Meanwhile, specialties including neurology, dermatology, and dentistry appeared only in a few studies, suggesting an emerging interest. This distribution suggests that while AI-driven personalization is being explored in healthcare care, it is particularly concentrated in data-rich domains such as oncology and genomics, aligning with the high potential for impact in these areas.

VI. THREATS TO VALIDITY

In any research study, there are threats to the validity. These threats can impact the credibility and generalizability of the results. In this section, we discuss three main categories of threats to validity: internal threats, external threats, and construct threats.

A. Internal Threats

In the context of a systematic literature review, these threats consist of issues associated with the selection of studies, data extraction procedures, and synthesizing techniques. For instance, biases in study selection, inconsistencies in data extraction, or subjective interpretations of findings can pose

significant internal threats to the validity of the research. To mitigate these threats, we followed a systematic approach based on established guidelines, ensuring thoroughness and objectivity in our review process.

In our study, another internal threat is related to the included papers. At the time our review was conducted, it was still in the middle of 2025 and, as a result, a complete list of publications for the year was not yet available. Given this limitation, it was not possible to include all relevant studies from 2025, as many papers had not been published or indexed by the time the review was finalized. Therefore, the scope of the review was restricted to studies published up to mid-2025.

B. External Threats

In the context of conducting our systematic review, external threats mainly refer to whether the findings of our review can be applied outside of specific circumstances. Because our review targets existing research, generalizability is not the main concern in our investigation. However, one potential external threat is the potential selection bias. To mitigate this threat, we adopted various resources for paper selection and followed exclusion and inclusion criteria to minimize external threats to validity.

C. Construct Threats

In our study, construct threats mainly refer to the design and execution of the review process. We deal with these issues by using rigorous validation techniques and utilizing standardized protocols.

VII. Challenges

The integration of artificial intelligence into personalized medicine is marked by numerous benefits, but also faces inherent challenges. One such limitation is the lack of a universally accepted definition of personalized medicine. This ambiguity makes it difficult to clearly delineate the scope of our review, which leads to challenges in systematically assessing AI applications across various medical domains.

In addition, when it moves toward customized treatments for individual patients and projects involving whole-genome data, particularly Whole Exome Sequencing (WES), it introduces challenges such as variant complexity and incomplete exome coverage. [2]. Although WES shows significant potential, it creates substantial data challenges that need complete analysis when dealing with complex biological data. Managing and interpreting such data remains a critical challenge.

Another challenge is the adoption of Phenotypic Personalized Medicine (PPM). PPM uses data specific to each individual to determine the best combination and doses of medicines using tools such as CURATE.AI. However, this approach faces challenges because it does not fully account for the differences between individuals and how drugs interact, which cannot be understood by analyzing the results generated from multiple trials [28]. Although PPM aims to optimize results, it faces challenges such as navigating the complex landscape of individual responses and ensuring patient safety.

These challenges underscore the need for advances in data processing, algorithmic refinement, and a deeper understanding of biological systems to exploit the potential of AI in personalized medicine. Despite these challenges, potential benefits, such as precise diagnostics, still highlight the importance of overcoming these challenges and limitations to revolutionize patient care and healthcare outcomes when AI techniques are applied in personalized medicine.

In addition, the cost of personalized medicine, including sequencing, infrastructure, and model maintenance, remains a major barrier to large-scale adoption, particularly in resource-constrained healthcare settings. Another challenge is the readiness of clinical workflows and healthcare systems to adopt AI-driven personalized approaches, which often require integration with electronic health records, clinician training, and organizational change management.

VIII. Future Directions

The future of AI in personalized medicine is closely related to the challenges discussed in our review. Future research efforts must focus on refining AI-based analytics, with a particular focus on improving data handling, storage infrastructure, and algorithmic sophistication to make personalized medicine datasets easier to interpret and more efficient.

Moreover, developing comprehensive ethical frameworks and regulatory paradigms is essential to ensure responsible use of AI in healthcare. These frameworks must address critical issues such as data privacy, fairness, and accountability, aligning with ethical and legal standards. Future research should also explore these barriers while fostering interdisciplinary collaboration among researchers, clinicians, ethicists, and regulators. Such efforts will enhance the equity, effectiveness, and reliability of AI applications in personalized medicine.

IX. Conclusion

Personalized medicine aims to tailor prevention, diagnosis, and treatment to patients' genetic, environmental, and lifestyle characteristics. The integration of AI methods—including machine learning, natural language processing, and deep learning—enables analysis of large, heterogeneous clinical datasets to support risk stratification, treatment-response prediction, and optimization of therapeutic strategies, thereby improving the precision and effectiveness of care.

This systematic literature review synthesizes evidence on the techniques, tools, and clinical specialties that shape current AI-enabled personalized medicine. Key priorities for future research and deployment include strengthening data privacy protections, mitigating algorithmic bias, ensuring regulatory compliance, and promoting equitable access, all of which are essential for responsible and scalable clinical adoption. Collectively, our findings highlight the transformative potential of AI to advance individualized healthcare and inform both methodological development and practice-oriented implementation.

REFERENCES

[1] O. Panahi, "The role of artificial intelligence in shaping future health planning", *Int J Health Policy Plann*, vol. 4, no. 1, pp. 01–05, 2025.

[2] P. Suwinski et al., "Advancing personalized medicine through the application of whole exome sequencing and big data analytics", *Frontiers in genetics*, vol. 10, p. 49, 2019.

[3] C. Gallo, *Artificial intelligence for personalized genetics and new drug development: Benefits and cautions*, 2023.

[4] J. Ma et al., "Using deep learning to model the hierarchical structure and function of a cell", *Nature methods*, vol. 15, no. 4, pp. 290–298, 2018.

[5] F. M. Franssen et al., "Personalized medicine for patients with copd: Where are we?", *International journal of chronic obstructive pulmonary disease*, pp. 1465–1484, 2019.

[6] N. S. Abul-Husn and E. E. Kenny, "Personalized medicine and the power of electronic health records", *Cell*, vol. 177, no. 1, pp. 58–69, 2019.

[7] K. Petrovic, "Deep learning in personalized medicine: Advancements and applications", *Journal of Advanced Analytics in Healthcare Management*, vol. 7, no. 1, pp. 34–50, 2023.

[8] K. Nova, "Generative ai in healthcare: Advancements in electronic health records, facilitating medical languages, and personalized patient care", *Journal of Advanced Analytics in Healthcare Management*, vol. 7, no. 1, pp. 115–131, 2023.

[9] S. Dinesh, B. Raj, and M. Manjunath, "Artificial intelligence based algorithms used for solving personalized medicine problems in personalized medicine application", *Perspectives in Communication, Embedded-systems and Signal-processing-PiCES*, pp. 13–16, 2022.

[10] M. W. Gifari, P. Samodro, and D. W. Kurniawan, "Artificial intelligence toward personalized medicine", *Pharmaceut. Sci. Res*, vol. 8, no. 2, p. 1, 2021.

[11] D. Ho et al., "Enabling technologies for personalized and precision medicine", *Trends in biotechnology*, vol. 38, no. 5, pp. 497–518, 2020.

[12] B. Kitchenham et al., "Procedures for performing systematic reviews", *Keele, UK, Keele University*, vol. 33, no. 2004, pp. 1–26, 2004.

[13] O. Ali et al., "A systematic literature review of artificial intelligence in the healthcare sector: Benefits, challenges, methodologies, and functionalities", *Journal of Innovation & Knowledge*, vol. 8, no. 1, p. 100 333, 2023.

[14] S. Rezayi, S. R. Niakan Kalhori, and S. Saeedi, "Effectiveness of artificial intelligence for personalized medicine in neoplasms: A systematic review", *BioMed Research International*, vol. 2022, no. 1, p. 7 842 566, 2022.

[15] M. T. Khan, "The revolutionary intersection of ai and healthcare: Embracing chatgpt with caution", *Chronicles of Biomedical Sciences*, vol. 2, no. 1, PID35–PID35, 2025.

[16] M. N. Mukabbir, "Ai in healthcare: Transforming patient care and outcomes", *Multidisciplinary Science Journal*, vol. 1, no. 01, pp. 29–39,

[17] N. Aiumtrakul et al., "Personalized medicine in urolithiasis: Ai chatbot-assisted dietary management of oxalate for kidney stone prevention", *Journal of Personalized Medicine*, vol. 14, no. 1, p. 107, 2024.

[18] V. Kolluri, "Revolutionizing healthcare delivery: The role of ai and machine learning in personalized medicine and predictive analytics", *Well Testing Journal*, vol. 33, no. S2, pp. 591–618, 2024.

[19] A. M. K. Sherani, M. Khan, M. U. Qayyum, and H. K. Hussain, "Synergizing ai and healthcare: Pioneering advances in cancer medicine for personalized treatment", *International Journal of Multidisciplinary Sciences and Arts*, vol. 3, no. 2, pp. 270–277, 2024.

[20] X. Chen, "Ai in healthcare: Revolutionizing diagnosis and treatment through machine learning", *MZ Journal of Artificial Intelligence*, vol. 1, no. 2, pp. 1–18, 2024.

[21] M. Cè et al., "Artificial intelligence in brain tumor imaging: A step toward personalized medicine", *Current Oncology*, vol. 30, no. 3, pp. 2673–2701, 2023.

[22] D. Susič, S. Syed-Abdul, E. Dovgan, J. Jonnagaddala, and A. Gradišek, "Artificial intelligence based personalized predictive survival among colorectal cancer patients", *Computer Methods and Programs in Biomedicine*, vol. 231, p. 107 435, 2023.

[23] M. Dave and N. Patel, "Artificial intelligence in healthcare and education", *British dental journal*, vol. 234, no. 10, pp. 761–764, 2023.

[24] M.-H. Temsah, A. Jamal, F. Aljamaan, J. A. Al-Tawfiq, and A. Al-Eyadhy, "Chatgpt-4 and the global burden of disease study: Advancing personalized healthcare through artificial intelligence in clinical and translational medicine", *Cureus*, vol. 15, no. 5, 2023.

[25] K. F. Hung, A. W. K. Yeung, M. M. Bornstein, and F. Schwendicke, "Personalized dental medicine, artificial intelligence, and their relevance for dentomaxillofacial imaging", *Dentomaxillofacial Radiology*, vol. 52, no. 1, p. 20 220 335, 2023.

[26] G. P. Patrinos et al., "Using chatgpt to predict the future of personalized medicine", *The pharmacogenomics journal*, vol. 23, no. 6, pp. 178–184, 2023.

[27] M. Sollini et al., "Artificial intelligence and hybrid imaging: The best match for personalized medicine in oncology", *European journal of hybrid imaging*, vol. 4, no. 1, p. 24, 2020.

[28] A. Blasiak, J. Khong, and T. Kee, "Curate. ai: Optimizing personalized medicine with artificial intelligence", *SLAS TECHNOLOGY: Translating Life Sciences Innovation*, vol. 25, no. 2, pp. 95–105, 2020.

[29] O. Khan, J. H. Badhiwala, G. Grasso, and M. G. Fehlings, "Use of machine learning and artificial intelligence to drive personalized medicine approaches for spine care", *World neurosurgery*, vol. 140, pp. 512–518, 2020.

[30] T. Litman, "Personalized medicine—concepts, technologies, and applications in inflammatory skin diseases", *Apmis*, vol. 127, no. 5, pp. 386–424, 2019.

[31] L. Lella et al., "Predictive ai models for the personalized medicine.", in *HEALTHINF*, 2019, pp. 396–401.

[32] R. Vashistha, A. K. Dangi, A. Kumar, D. Chhabra, and P. Shukla, "Futuristic biosensors for cardiac health care: An artificial intelligence approach", *3 Biotech*, vol. 8, no. 8, p. 358, 2018.

[33] M. D. Abràmoff, P. T. Lavin, M. Birch, N. Shah, and J. C. Folk, "Pivotal trial of an autonomous ai-based diagnostic system for detection of diabetic retinopathy in primary care offices", *NPJ digital medicine*, vol. 1, no. 1, p. 39, 2018.

[34] A. Rajkomar et al., "Scalable and accurate deep learning with electronic health records", *NPJ digital medicine*, vol. 1, no. 1, p. 18, 2018.

[35] M. Ivanovic and M. Semnic, "The role of agent technologies in personalized medicine", in *2018 5th International Conference on Systems and Informatics (ICSAI)*, IEEE, 2018, pp. 299–304.

[36] N. J. Schork and K. Nazor, "Integrated genomic medicine: A paradigm for rare diseases and beyond", *Advances in genetics*, vol. 97, pp. 81–113, 2017.

[37] J. Zhou and O. G. Troyanskaya, "Predicting effects of noncoding variants with deep learning–based sequence model", *Nature methods*, vol. 12, no. 10, pp. 931–934, 2015.

[38] I. Kouris, C. Tsirmpas, S. G. Mougiakakou, D. Iliopoulou, and D. Koutsouris, "E-health towards ecumenical framework for personalized medicine via decision support system", in *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology*, IEEE, 2010, pp. 2881–2885.