algorithms

Special Issue Reprint

AI-Assisted Medical Diagnostics

Edited by
Milan Toma

mdpi.com/journal/algorithms

MDPI

AI-Assisted Medical Diagnostics

AI-Assisted Medical Diagnostics

Guest Editor

Milan Toma

MDPI

Basel • Beijing • Wuhan • Barcelona • Belgrade • Novi Sad • Cluj • Manchester

Guest Editor
Milan Toma
Department of Osteopathic
Manipulative Medicine
New York Institute of
Technology
New York, NY
USA

Editorial Office
MDPI AG
Grosspeteranlage 5
4052 Basel, Switzerland

This is a reprint of the Special Issue, published open access by the journal *Algorithms* (ISSN 1999-4893), freely accessible at: https://www.mdpi.com/journal/algorithms/special_issues/D4501BG2CZ.

For citation purposes, cite each article independently as indicated on the article page online and as indicated below:

Lastname, A.A.; Lastname, B.B. Article Title. *Journal Name* **Year**, *Volume Number*, Page Range.

ISBN 978-3-7258-9057-6 (Hbk)
ISBN 978-3-7258-9058-3 (PDF)
https://doi.org/10.3390/books978-3-7258-9058-3

© 2026 by the authors. Articles in this reprint are Open Access and distributed under the Creative Commons Attribution (CC BY) license. The reprint as a whole is distributed by MDPI under the terms and conditions of the Creative Commons Attribution-NonCommercial-NoDerivs (CC BY-NC-ND) license (https://creativecommons.org/licenses/by-nc-nd/4.0/).

Contents

About the Editor . . . . . **vii**

Preface . . . . . **ix**

Milan Toma
From Interpretable Models to Clinical Implementation: Advances in AI-Assisted Medical Diagnostics
Reprinted from: *Algorithms* **2026**, *19*, 330, https://doi.org/10.3390/a19050330 . . . . . **1**

Aleksandra Łoś, Dorota Bartusik-Aebisher, Wiktoria Mytych and David Aebisher
Applications of Artificial Intelligence in Selected Internal Medicine Specialties: A Critical Narrative Review of the Latest Clinical Evidence
Reprinted from: *Algorithms* **2026**, *19*, 54, https://doi.org/10.3390/a19010054 . . . . . **10**

Hanvedes Daovisan, Charin Suwanwong, Pitchada Prasittichok, Narulmon Prayai and Phaktada Choowan
Reclaiming XAI as an Innovation in Healthcare: Bridging Rule-Based Systems
Reprinted from: *Algorithms* **2025**, *18*, 586, https://doi.org/10.3390/a18090586 . . . . . **51**

Luis F. F. M. Santos, Miguel Ángel Sánchez-Tena, Cristina Alvarez-Peregrina and Clara Martinez-Perez
Artificial Intelligence-Driven Diagnostics in Eye Care: A Random Forest Approach for Data Classification and Predictive Modeling
Reprinted from: *Algorithms* **2025**, *18*, 647, https://doi.org/10.3390/a18100647 . . . . . **75**

Haneen Altartouri, Sahar Qaadan, Yousef Qawqzeh and Mohammed Hamdoun
Probabilities Feature Enrichment for Improved Early Diabetes Prediction
Reprinted from: *Algorithms* **2026**, *19*, 122, https://doi.org/10.3390/a19020122 . . . . . **99**

Lan Wei and Catherine Mooney
DenseNet-Based Classification of EEG Abnormalities Using Spectrograms
Reprinted from: *Algorithms* **2025**, *18*, 486, https://doi.org/10.3390/a18080486 . . . . . **129**

Ilyas Potamitis
Affordable Audio Hardware and Artificial Intelligence Can Transform the Dementia Care Pipeline
Reprinted from: *Algorithms* **2025**, *18*, 787, https://doi.org/10.3390/a18120787 . . . . . **152**

Abdussalam Salama, Reza Saatchi, Maryam Bagheri, Mahpara Saleem and Muhammad Usman Shad
Development and Evaluation of a Real-Time Home Monitoring Application Utilising Long Short-Term Memory Integrated in a Smartphone
Reprinted from: *Algorithms* **2025**, *18*, 780, https://doi.org/10.3390/a18120780 . . . . . **184**

Sungjoon Hong, Mihir Matalia and Milan Toma
Chatting Ain't Diagnosing: Diagnostic Variability and Fundamental Errors in Multimodal LLM Interpretation in Radiology
Reprinted from: *Algorithms* **2026**, *19*, 170, https://doi.org/10.3390/a19030170 . . . . . **205**

About the Editor

Milan Toma

Milan Toma is an Associate Professor at the New York Institute of Technology College of Osteopathic Medicine, specializing in clinical sciences with expertise in computational biomechanics, biomedical engineering, and medical image processing. He earned his Ph.D. in Engineering Science from the Technical University of Lisbon and has held research positions at prestigious institutions including Tokyo University, RIKEN, King Abdullah University of Science and Technology, and Georgia Institute of Technology. His research focuses on cardiovascular fluid–structure interaction, brain injury biomechanics, and maternal–fetal health dynamics. In recent years, Dr. Toma has increasingly concentrated on integrating machine learning and artificial intelligence into medical applications. He is the author of two books on artificial intelligence and medical diagnostics, reflecting his commitment to both advancing research and educating the next generation of healthcare professionals. Dr. Toma is dedicated to advancing medical science through innovative computational methods while bridging the knowledge gap between engineering and clinical practice.

Preface

Artificial intelligence has emerged as one of the most promising frontiers in modern healthcare, offering unprecedented opportunities to enhance diagnostic accuracy, improve patient outcomes, and transform clinical workflows. Yet alongside this remarkable potential comes the critical responsibility to rigorously evaluate these technologies, understand their limitations, and ensure their safe and effective integration into medical practice.

This Reprint was conceived to provide a balanced and comprehensive examination of AI applications in medical diagnostics. Our motivation stems from observing a widening gap between enthusiastic media coverage of AI capabilities and the nuanced reality of clinical implementation. Healthcare professionals, researchers, and technology developers need accessible, peer-reviewed evidence that illuminates both the achievements and challenges facing AI in medicine.

The scope of this collection spans multiple dimensions of AI-driven healthcare innovation. We present research addressing foundational questions about explainability and trustworthiness in clinical AI systems, practical implementations across diverse medical specialties including ophthalmology, endocrinology, and neurology, real-world monitoring solutions for aging populations and chronic disease management, and critical evaluation of diagnostic reliability and potential pitfalls.

This Reprint is addressed to a broad interdisciplinary audience: clinicians seeking to understand AI tools relevant to their practice, researchers developing next-generation diagnostic algorithms, healthcare administrators evaluating AI adoption strategies, computer scientists working on medical AI applications, and students entering the field of health informatics.

Our aim is to foster informed dialog about the responsible advancement of AI in healthcare, grounded in scientific evidence and realistic assessment of both capabilities and limitations.

Milan Toma
Guest Editor

algorithms

MDPI

Editorial

From Interpretable Models to Clinical Implementation: Advances in AI-Assisted Medical Diagnostics

Milan Toma

Algorithmic Medicine Laboratory, Department of Osteopathic Manipulative Medicine, College of Osteopathic Medicine, New York Institute of Technology, Old Westbury, NY 11568, USA; tomamil@tomamil.com

The integration of artificial intelligence into medical diagnostics has evolved from controlled research demonstrations to real-world clinical deployment, creating both unprecedented opportunities and substantial challenges. While AI systems have demonstrated remarkable performance in controlled research settings, their translation into reliable clinical tools demands careful consideration of interpretability, generalizability, safety, and appropriate validation. A central tension persists between model complexity and clinical transparency: sophisticated deep learning architectures may achieve impressive accuracy metrics but often operate as black boxes that resist clinical interpretation, while inherently interpretable models may be perceived as sacrificing predictive power for explainability.

This Special Issue addresses these challenges from theoretical foundations through practical implementation, investigating how AI can augment clinical decision-making while maintaining the transparency, accountability, and reliability that patient care demands. A critical concern is that shortcuts to high accuracy carry unacceptable risks: spurious correlations and dataset biases can produce models that perform well on validation datasets yet fail catastrophically in practice. The papers here emphasize not merely accuracy, but validated reliability, interpretability, and careful alignment of AI approaches to clinical contexts.

This collection can be organized into three thematic categories (Figure 1), with each addressing distinct aspects of AI-assisted medical diagnostics while contributing to an integrated understanding of how artificial intelligence can be responsibly deployed in clinical care.

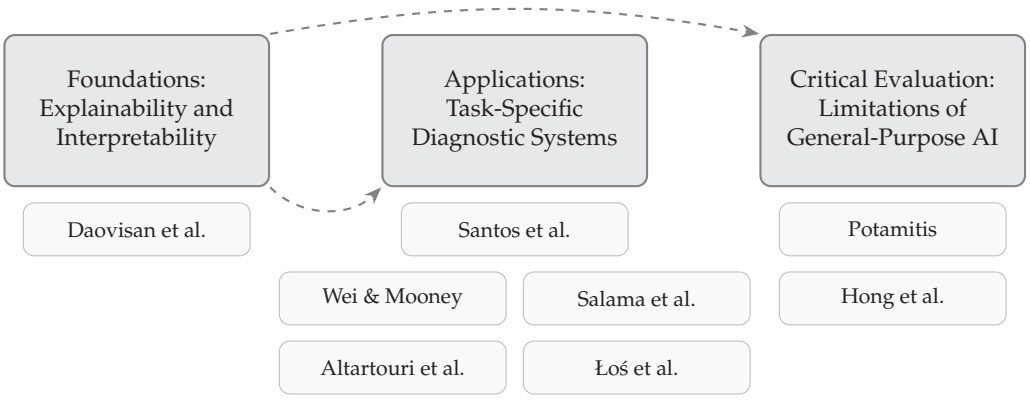


Figure 1. The organization of the three thematic categories for this Special Issue. The foundations pillar establishes theoretical principles that inform the task-specific applications, while critical evaluation provides necessary constraints on technology selection (Contributions 1–8).

The theoretical foundation for responsible clinical AI deployment is established by Daovisan et al. (Contribution 1), who provide a comprehensive scientometric analysis of explainable artificial intelligence (XAI) in healthcare, with particular emphasis on the role of

rule-based systems. Through systematic mapping of 654 Scopus-indexed publications spanning recent years, the authors identify transparency, accountability, and trustworthiness as the central values that must be satisfied for clinical AI integration.

Their analysis demonstrates that rule-based systems (often deployed in hybrid forms that combine statistical learning with explicit logical rules) provide an essential bridge between algorithmic complexity and human interpretability (Figure 2). Daovisan et al. establish a theoretical foundation that resonates throughout this collection: clinicians must not only trust AI recommendations but also explain them to patients, justify decisions in medicolegal contexts, and maintain the ability to override automated suggestions when clinical judgment demands it.

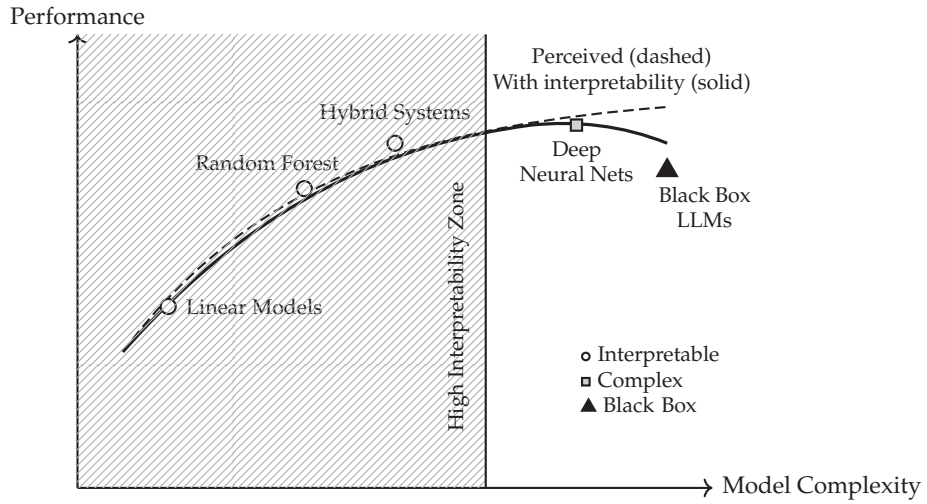


Figure 2. The relationship between model complexity, performance, and interpretability in clinical AI. The dashed curve shows the perceived unlimited performance gain with complexity, while the solid curve represents actual performance when interpretability constraints are considered. Excessive complexity without interpretability mechanisms (shown by the performance decline beyond the optimal point) can reduce clinical utility. The hatched region indicates the high interpretability zone. Symbol key: circles = interpretable models; squares = complex models; triangles = black box systems. The papers in this Special Issue demonstrate that interpretable models can achieve strong performance within clinical constraints.

The authors further identify smart healthcare, digital health, and mobile health (mHealth) as expanding application domains where explainability becomes even more critical, as these contexts often involve patient-facing applications and remote monitoring scenarios where expert oversight may be limited (Contribution 1).

Notably, the analysis reveals that interpretability remains systematically underrepresented in current research despite its recognized importance. This gap has practical consequences for clinical adoption, as physicians are unlikely to rely on systems whose recommendations they cannot justify to patients or colleagues. The subsequent papers in this collection respond to this challenge by demonstrating how interpretability can be operationalized across diverse medical applications.

Four papers present purpose-built AI systems designed for specific clinical challenges, demonstrating how the theoretical principles of interpretability and validation translate into practical diagnostic tools (Figure 3).

Santos et al. (Contribution 2) introduce an AI-based system for automated classification and forecasting of optometric data using Random Forest models, classifying categories including contactology, dry eye, low vision, myopia, pediatrics, and refractive surgery. The epoch-trained model achieves 94.24% accuracy, 94.70% precision, and F1-scores from 92.54%

https://doi.org/10.3390/a19050330

to 98.46%, outperforming single-pass training. Coupled with ARIMA-based forecasting through 2030, this work demonstrates that classical machine learning (here an ensemble method with inherent feature importance quantification) can provide interpretable diagnostic support without deep neural network complexity.

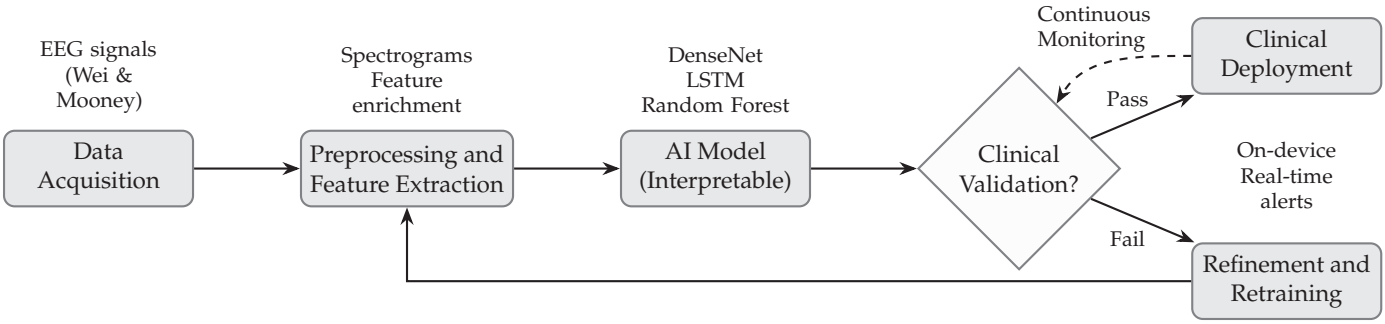


Figure 3. Common pipeline for task-specific diagnostic AI systems illustrated by the applications in this Special Issue. All systems emphasize interpretability and require rigorous clinical validation before deployment, with continuous monitoring post-deployment (Contribution 3).

Wei and Mooney (Contribution 3) address EEG abnormality classification using a DenseNet-based system that transforms signals into time–frequency spectrograms, achieving 79.85% accuracy, 75.0% sensitivity, 83.78% specificity, and 76.92% F1-score, while maintaining interpretability through LIME and Grad-CAM visualization. This integration of deep learning with explainability tools illustrates a practical path forward: leveraging deep pattern recognition while embedding biologically meaningful domain knowledge to promote learning of clinically relevant rather than artifactual features (Contribution 2).

Salama et al. (Contribution 4) present a real-time home monitoring application using LSTM networks on smartphones to detect falls and unexpected inactivity in elderly populations. The personalized model adapts through continuous seven-day training windows, achieving $R^2 = 0.93$ and MAE = 0.05 while maintaining low false positive rates. On-device processing preserves privacy and ensures real-time responsiveness without cloud connectivity (Contributions 1, 4 and 5), while a multi-channel alert system (SMS, email, local notifications) demonstrates attention to practical deployment considerations often overlooked in research prototypes.

Altartouri et al. (Contribution 6) address early diabetes prediction through feature enrichment using Gaussian Mixture Models and Kernel Density Estimation, augmenting the feature space with latent distributional characteristics to achieve substantial gains in sensitivity for minority class (diabetic) cases. Minority class sensitivity is clinically crucial: false negatives carry far greater consequences than false positives in screening, as undetected diabetes leads to progressive complications. This approach offers a middle path between linear interpretability and deep learning flexibility, enhancing performance without sacrificing explainability (Contribution 6).

Łoś et al. (Contribution 7) provide a detailed narrative review synthesizing the most recent evidence on AI applications across multiple internal medicine specialties, including cardiology, pulmonology, neurology, hepatology, gastroenterology, and oncology. The review highlights that as of November 2025, over 850 FDA-cleared AI-enabled medical devices exist, with more than 70% related to imaging applications (Contribution 7). Notable findings include the TAILORED-AF trial demonstrating 88% freedom from atrial fibrillation at 12 months with AI-guided ablation versus 70% with conventional approaches, and the identification of the first AI-generated drug (INS018_055/rentosertib) to demonstrate clinical proof-of-concept in Phase 2 trials for idiopathic pulmonary fibrosis. The authors emphasize critical limitations in current evidence, including the predominance of retrospec-

https://doi.org/10.3390/a19050330

tive studies with limited external validation, concerns about domain shift and algorithmic bias, and the need for large-scale prospective RCTs with hard clinical endpoints before universal implementation.

Two papers in this collection serve essential corrective functions, examining both the practical challenges of AI deployment in complex care scenarios and the fundamental limitations of general-purpose AI systems when applied to specialized medical tasks.

Potamitis (Contribution 5) presents an audio-driven monitoring pipeline for dementia care (Figure 4) that operates on consumer hardware (mobile phones, microcontroller nodes, or smart television sets), combining audio signal processing with AI interpretation. The system integrates voice activity detection, speaker diarization, automatic speech recognition for dialog analysis, and speech emotion recognition. An audio classifier detects care-relevant events including cough, cane taps, thuds, knocks, and speech patterns. Notably, a large language model synthesizes these multimodal inputs alongside a consented household knowledge base to generate daily caregiver reports covering orientation/disorientation (person, place, and time), delusion themes, agitation events, health proxies, and safety flags (e.g., exit seeking, fall detection).

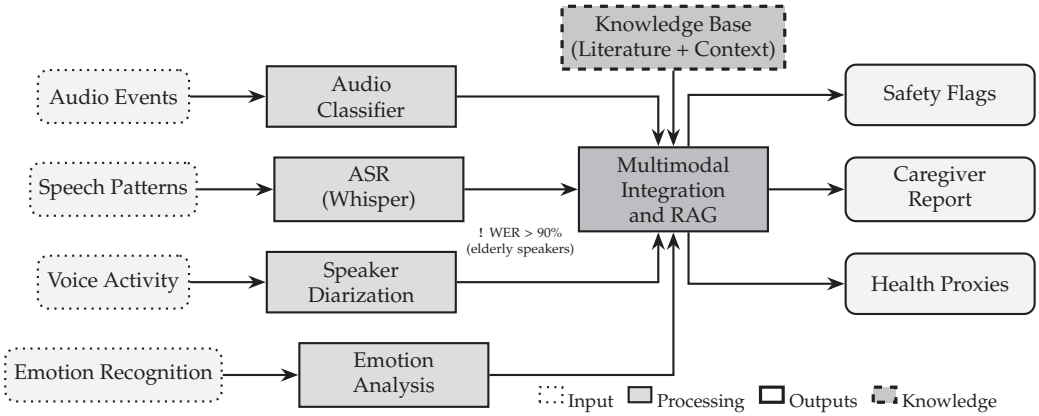


Figure 4. Multimodal integration architecture for dementia care monitoring (Potamitis). The system combines audio event detection, speech recognition, speaker diarization, and emotion analysis with retrieval-augmented generation to produce comprehensive caregiver reports. Node styles: dotted borders indicate input data sources; light gray boxes represent processing modules; thick-bordered boxes denote system outputs; dashed border indicates external knowledge base. The boxed exclamation mark (!WER > 90%) highlights identified performance limitation with elderly speakers.

Dementia care presents particular challenges, including significant inter-individual variability, unpredictable symptom fluctuation, and the need to balance false alarms against missed emergencies. Critically, Whisper large-v3 ASR achieved Word Error Rates exceeding 90% for elderly speakers with dementia due to age-related voice changes and environmental interference (Contribution 5), which is a finding that has major implications for voice-based AI in geriatric care. Retrieval-augmented generation (RAG) over indexed dementia literature partially addresses LLM reliability concerns, producing page-level-cited caregiver reports. Potamitis demonstrates that such systems function best as adjuncts to caregiving rather than autonomous diagnostic devices, requiring human judgment to contextualize AI-generated alerts.

Hong et al. (Contribution 8) provide a critical examination of multimodal large language models (LLMs) for radiological image interpretation, revealing fundamental limitations that distinguish general-purpose AI systems from task-specific diagnostic tools. The study evaluated five leading multimodal LLMs (GPT-5, Gemini 3 Pro, Llama 4 Maverick, Grok 4, and Claude Opus 4.5 Extended) on a standardized non-contrast head CT interpretation task.

https://doi.org/10.3390/a19050330

The results expose a 20% rate of fundamental diagnostic error, with one model misidentifying ischemic stroke as intracerebral hemorrhage with incorrect lateralization, i.e., an error with potentially catastrophic clinical implications, as the management of ischemic stroke differs fundamentally from that of intracerebral hemorrhage. Even among models that reached concordant general diagnoses, clinically meaningful variability persisted in acuity characterization (acute vs. subacute), anatomical localization, and differential diagnosis generation, i.e., differences that would lead to divergent clinical workups and management strategies.

Most significantly, Hong et al. implemented a novel cross-evaluation protocol wherein each LLM graded all five responses. This approach revealed that LLMs cannot reliably agree on radiological ground truth: one model interpreted the image as showing acute infarction with mass effect, while another concluded that it depicted chronic infarction with atrophy, i.e., diametrically opposed interpretations with opposite therapeutic implications. This ground truth disagreement, combined with observed self-evaluation bias and inconsistent grading standards, undermines proposals for using LLMs as automated quality assurance tools or educational assessment systems (Contribution 8).

The diagnostic variability, fundamental interpretive disagreements, and critical errors documented by Hong et al. (Contribution 8) raise a broader question: Are general-purpose multimodal LLMs the appropriate tool for medical image interpretation, or should clinical applications prioritize purpose-built machine learning models trained specifically for defined imaging tasks and pathological conditions? Table 1 summarizes the key distinctions between these two paradigms across seven critical dimensions, providing a framework for understanding when each approach is appropriate.

Table 1 clarifies why general-purpose LLMs are unsuited for autonomous diagnostic image interpretation: despite conversational fluency, they lack validated reliability, regulatory approval, and interpretability mechanisms, reflecting a deeper architectural mismatch in which medical imaging constitutes only a small fraction of heterogeneous training data. Task-specific models, by contrast, are validated against clinical ground truth with interpretability mechanisms such as Class Activation Mapping that reveal which image regions drive predictions. As Table 1 illustrates, LLMs excel at documentation and literature synthesis, while task-specific models suit diagnostic triage and quantitative analysis within their validated scope.

Figure 5 provides a practical decision framework for clinicians and healthcare administrators evaluating which AI technology is appropriate for specific applications.

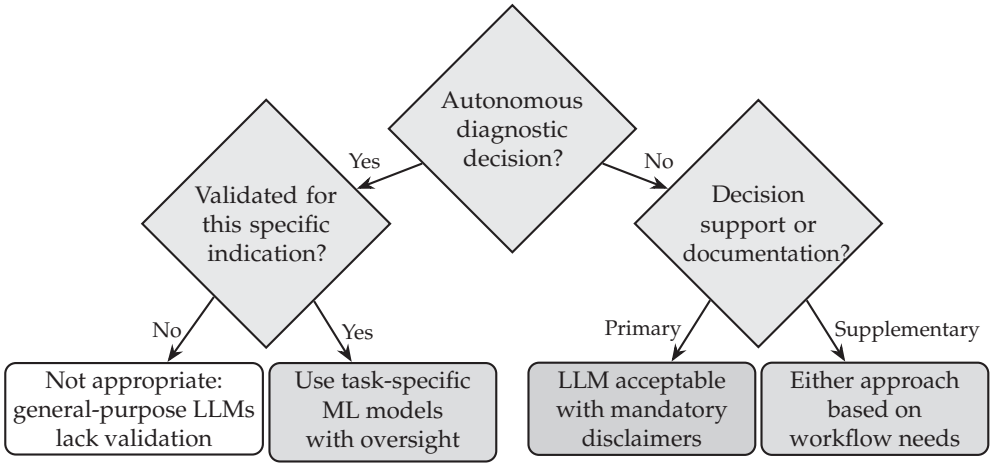


Figure 5. Decision framework for selecting between general-purpose LLMs and task-specific machine learning models in clinical settings. Autonomous diagnostic applications require cleared task-specific models, while LLMs may serve supporting roles with appropriate disclaimers.

https://doi.org/10.3390/a19050330

Table 1. Comparison of general-purpose multimodal LLMs versus task-specific machine learning models for medical image interpretation.

Characteristic	General-Purpose LLMs	Task-Specific ML Models
Training Paradigm	Massive, heterogeneous corpora spanning diverse domains; medical imaging content constitutes a small, potentially unrepresentative fraction	Curated datasets comprising thousands to millions of expert-annotated examples for a specific imaging modality and defined pathological conditions
Feature Optimization	Visual representations not optimized for specific diagnostic features; broad but shallow capabilities across many domains	Learned representations directly optimized for discriminative features relevant to the diagnostic task (e.g., hypodensity vs. hyperdensity, sulcal effacement vs. prominence)
Performance Characteristics	Largely unknown and unvalidated for radiological interpretation; 20% fundamental diagnostic error rate observed in Hong et al. (Contribution 8); failure modes poorly characterized	Quantified sensitivity, specificity, and predictive values; validated against histopathological or clinically confirmed diagnoses; systematic failure mode analysis
Regulatory Status	Not validated or cleared for diagnostic imaging interpretation; no FDA authorization for autonomous clinical use	Multiple systems FDA-cleared (510(k) or De Novo) for specific indications; defined performance benchmarks and intended use populations
Interpretability	Text-based explanations generated post hoc; may not reflect actual visual features driving interpretation (cf. Grok 4 describing "hyperdense" findings on hypodense lesion)	Direct visual attention mapping (CAM, Grad-CAM) highlighting image regions driving predictions; enables verification of reasoning alignment
Clinical Integration	No established workflows; outputs lack standardization for PACS integration; variable disclaimer inclusion	Purpose-built for clinical workflow integration; standardized outputs formatted for radiologist review and PACS compatibility
Appropriate Applications	Report summarization, clinical documentation assistance, natural language interfaces, literature synthesis, workflow orchestration	Diagnostic triage, lesion detection and localization, quantitative analysis, second-reader applications within validated scope

Importantly, only one of the five evaluated LLMs included appropriate safety disclaimers cautioning users about limitations, i.e., a concerning finding given the potential for inappropriate clinical reliance on AI-generated interpretations. This observation reinforces the need for not only technological advancement but also responsible deployment frameworks that include mandatory safety messaging, clear scope limitations, and explicit guidance on appropriate use contexts. These eight contributions advance understanding of AI-assisted medical diagnostics along several interconnected dimensions, while also illuminating critical gaps and challenges that define the future research agenda.

First and foremost, the application papers empirically validate the interpretability principles established by Daovisan et al. (Contribution 1) (Figure 6). Santos (Contribution 2), Wei and Mooney (Contribution 3), Salama et al. (Contribution 4), Altartouri et al. (Contribution 6), and Łoś et al. (Contribution 7) each implement architectures that balance predictive performance with explainability across domains ranging from optometry to neurology to geriatric care—confirming that interpretability is a universal requirement rather than a domain-specific concern.

Second, the Special Issue highlights the importance of distinguishing between appropriate and inappropriate AI technologies for specific clinical tasks. Hong et al.'s evaluation (Contribution 8) reveals that general-purpose language models exhibit unacceptable

https://doi.org/10.3390/a19050330

diagnostic variability for autonomous image interpretation, confirming that technology choice must be matched to deployment context, as outlined in Table 1.

Third, these contributions demonstrate that successful clinical AI requires personalization and continuous adaptation. Salama et al.'s on-device training protocol (Contribution 4) and Altartouri et al.'s feature enrichment approach (Contribution 6) both show how models can be tailored to individual patients or local population characteristics, improving performance beyond what generic pretrained systems can achieve.

Fourth, Potamitis (Contribution 5) demonstrates that AI can support complex care scenarios through multimodal integration, processing audio, speech, and temporal patterns to generate actionable insights for caregivers. This integration of diverse data sources represents a direction for clinical decision support systems.

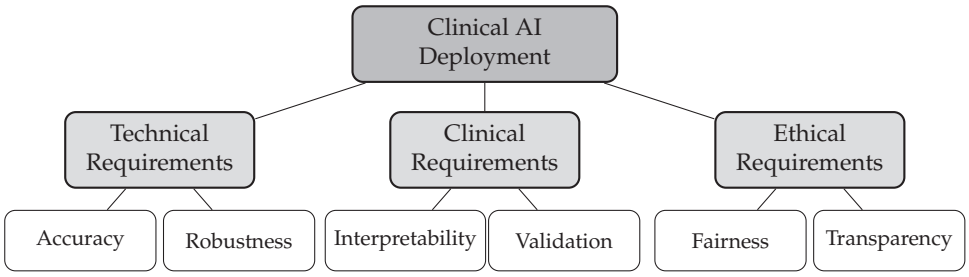


Figure 6. Hierarchical requirements for responsible clinical AI deployment. Technical performance alone is insufficient; clinical and ethical requirements are equally essential. Visual hierarchy: darker shading and thicker borders indicate higher-level concepts. All requirements must be satisfied simultaneously. The papers in this Special Issue address all three categories.

Several important directions for future research emerge from this collection. Large-scale, multi-center prospective trials remain necessary to validate AI systems across diverse populations and practice settings. External validation gaps, particularly for systems trained primarily in high-resource academic centers, must be addressed before universal implementation. Bias mitigation strategies require development to ensure equitable performance across demographic subgroups. The integration of AI systems into existing clinical workflows demands attention to user interface design, cognitive load, and workflow disruption. The regulatory landscape continues to evolve, with implications for approval pathways, post-market surveillance, and liability frameworks.

The field must develop standardized benchmarks and evaluation protocols that go beyond accuracy metrics to assess robustness, fairness, interpretability, and clinical utility. The cross-evaluation methodology introduced by Hong et al. (Contribution 8) provides one example of how assessment can reveal limitations invisible in conventional validation approaches. Education and training programs must prepare clinicians to effectively use AI-assisted diagnostic tools, understanding both their capabilities and limitations.

This Special Issue presents a comprehensive examination of AI-assisted medical diagnostics spanning theoretical foundations, validated applications, and critical evaluation of emerging technologies. Collectively, these works demonstrate that the path toward clinically successful AI is neither linear nor straightforward: it requires simultaneous attention to algorithmic performance, interpretability, validation rigor, and appropriate matching of technology to clinical context.

The consistent thread connecting these diverse contributions is the recognition that medical AI is fundamentally different from AI in other domains. The stakes are higher, the tolerance for error is lower, and the requirement for human understanding is non-negotiable. Achieving high accuracy on benchmark datasets is necessary but insufficient; clinical AI must also be interpretable to practitioners, trustworthy across patient populations, robust to real-world variability, and deployed within systems that acknowledge their limitations.

https://doi.org/10.3390/a19050330

The papers in this collection chart a course forward that embraces both innovation and humility. They demonstrate powerful new capabilities (from real-time home monitoring to automated radiological screening) while simultaneously revealing the limitations of approaches that prioritize fluency over reliability or generality over task-specific validation. This balanced perspective is precisely what the field requires as AI transitions from promising research demonstrations to deployed clinical tools affecting real patient outcomes.

Several actionable implications emerge. For researchers: prioritize interpretability from the outset; validate across diverse populations; and maintain honesty about limitations. For clinicians: demand transparency in AI decision-making; insist on validation evidence from relevant populations; and distinguish between general-purpose LLMs and task-specific validated models. For policymakers: develop regulatory frameworks that distinguish rigorously validated AI tools from those that have not; mandate appropriate safety disclaimers; and support monitoring systems, human oversight protocols, and continuous validation pathways.

Looking forward, we envision a future where AI in medicine is characterized by deliberate pairing of technology to task: task-specific machine learning models for applications requiring validated diagnostic accuracy; general-purpose language models for documentation support and information synthesis; hybrid systems that combine the strengths of different approaches; and, throughout, a commitment to transparency that enables clinicians to understand, trust, and appropriately rely on AI assistance.

We hope this collection serves as both a resource for current practitioners and an inspiration for future work refining and responsibly deploying AI to improve patient care. AI has transitioned from experimental curiosity to a clinically deployed tool, demonstrating measurable improvements in outcomes and efficiency. By maintaining focus on interpretability, validation, and appropriate deployment, the field can deliver on the promise of AI-enhanced healthcare that is more efficient, equitable, and trustworthy.

Acknowledgments: I thank all authors for their contributions and the reviewers for their feedback, which ensured the quality of this Special Issue. I also acknowledge the editorial support provided by the *Algorithms* journal staff throughout the publication process.

Conflicts of Interest: The author declares no conflicts of interest.

List of Contributions:

1. Daovisan, H.; Suwanwong, C.; Prasittichok, P.; Prayai, N.; Choowan, P. Reclaiming XAI as an Innovation in Healthcare: Bridging Rule-Based Systems. *Algorithms* **2025**, *18*, 586. https://doi.org/10.3390/a18090586.
2. Santos, L.; Sánchez-Tena, M.; Alvarez-Peregrina, C.; Martinez-Perez, C. Artificial Intelligence-Driven Diagnostics in Eye Care: A Random Forest Approach for Data Classification and Predictive Modeling. *Algorithms* **2025**, *18*, 647. https://doi.org/10.3390/a18100647.
3. Wei, L.; Mooney, C. DenseNet-Based Classification of EEG Abnormalities Using Spectrograms. *Algorithms* **2025**, *18*, 486. https://doi.org/10.3390/a18080486.
4. Salama, A.; Saatchi, R.; Bagheri, M.; Saleem, M.; Shad, M. Development and Evaluation of a Real-Time Home Monitoring Application Utilising Long Short-Term Memory Integrated in a Smartphone. *Algorithms* **2025**, *18*, 780. https://doi.org/10.3390/a18120780.
5. Potamitis, I. Affordable Audio Hardware and Artificial Intelligence Can Transform the Dementia Care Pipeline. *Algorithms* **2025**, *18*, 787. https://doi.org/10.3390/a18120787.
6. Altartouri, H.; Qaadan, S.; Qawqzeh, Y.; Hamdoun, M. Probabilities Feature Enrichment for Improved Early Diabetes Prediction. *Algorithms* **2026**, *19*, 122. https://doi.org/10.3390/a19020122.

7. Łoś, A.; Bartusik-Aebisher, D.; Mytych, W.; Aebisher, D. Applications of Artificial Intelligence in Selected Internal Medicine Specialties: A Critical Narrative Review of the Latest Clinical Evidence. *Algorithms* **2026**, *19*, 54. https://doi.org/10.3390/a19010054.
8. Hong, S.; Matalia, M.; Toma, M. Chatting Ain't Diagnosing: Diagnostic Variability and Fundamental Errors in Multimodal LLM Interpretation in Radiology. *Algorithms* **2026**, *19*, 170. https://doi.org/10.3390/a19030170.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

algorithms

MDPI

Review

Applications of Artificial Intelligence in Selected Internal Medicine Specialties: A Critical Narrative Review of the Latest Clinical Evidence

Aleksandra Łoś [1], Dorota Bartusik-Aebisher [2], Wiktoria Mytych [1] and David Aebisher [3,*]

[1] English Division Science Club, Collegium Medicum, Faculty of Medicine, University of Rzeszów, 35-310 Rzeszów, Poland
[2] Department of Biochemistry and General Chemistry, Collegium Medicum, Faculty of Medicine, University of Rzeszów, 35-310 Rzeszów, Poland
[3] Department of Photomedicine and Physical Chemisry, Collegium Medicum, Faculty of Medicine, University of Rzeszów, 35-310 Rzeszów, Poland
* Correspondence: daebisher@ur.edu.pl

Abstract

Background: Artificial intelligence (AI) is rapidly transforming clinical medicine by enabling earlier disease detection, personalized risk stratification, precision diagnostics, and optimized therapeutic decision-making across multiple specialties. Methods: This narrative review synthesizes the most recent evidence from prospective randomized controlled trials, large cohort studies, and real-world implementations of AI in cardiology, pulmonology, neurology, hepatology, pancreatic diseases, and other key areas of internal medicine. Studies were selected based on clinical impact, external validation, and regulatory approval status where applicable. Results: AI systems now outperform traditional clinical tools in numerous high-stakes applications: >88% freedom from atrial fibrillation at 1 year with AI-guided ablation, noninferior stent optimization versus OCT guidance, >95% sensitivity for atrial fibrillation and low ejection fraction detection on single-lead ECG, substantial increases in adenoma detection rate and melanoma triage accuracy, automated pancreatic cancer detection on routine CT with 89–90% sensitivity, and significant improvements in palliative care consultation rates and post-PCI outcomes using AI-supported telemedicine. Over 850 FDA-cleared AI devices exist as of November 2025, with cardiology and radiology dominating clinical adoption. Conclusions: AI has transitioned from experimental to clinically indispensable in multiple specialties, delivering measurable reductions in mortality, morbidity, hospitalizations, and healthcare resource utilization. Remaining challenges include external validation gaps, bias mitigation, and the need for large-scale prospective trials before universal implementation.

Keywords: artificial intelligence; health monitoring; disease prediction; prognostics; medical imaging

1. Introduction

Artificial intelligence (AI) is rapidly transforming internal medicine specialties. This review examines the current applications of AI in selected fields, with a focus on high-impact trials and real-world implementations [1]. The number of PubMed-indexed publications containing the term "artificial intelligence" in the title or abstract has increased more than 30-fold since 2010 [2,3]. The widespread adoption of wearable devices and the Internet of

Medical Things has enabled continuous, non-invasive collection of physiological data [4,5]. AI algorithms analyze ECG signals, heart rate, oxygen saturation, physical activity, sleep patterns, and biochemical parameters (Figure 1). Meta-analyses published between 2023 and 2025 report that deep learning models achieve >95% sensitivity and >98% specificity in detecting atrial fibrillation using data from devices such as Apple Watch or Fitbit [6,7]. Recent studies have also explored multimodal health monitoring using federated learning to preserve patient privacy [8,9]. Predictive AI models frequently outperform traditional risk assessment tools. For example, multimodal neural networks combining genetic, laboratory, imaging, and lifestyle data from the UK Biobank achieved an AUC > 0.92 for 10-year prediction of cardiovascular disease, type 2 diabetes, and selected cancers [10,11].

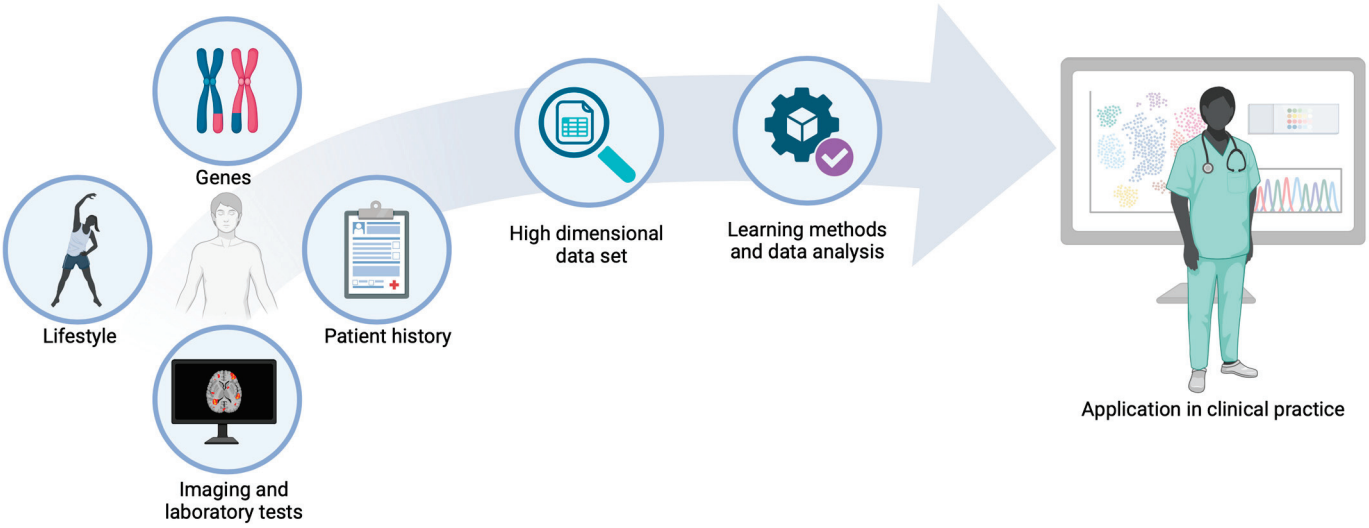


Figure 1. Schematic representation of the integration of artificial intelligence in medicine: from multidimensional patient data collection to clinical application.

In 2025, the first prospective randomized clinical trials demonstrated that AI-driven disease prediction in primary care was associated with a 20–30% reduction in hospitalizations for acute cardiovascular events [12,13]. In oncology, cardiology, and intensive care, AI-based prognostic models often outperform conventional scoring systems. Examples include the DeepMind/Google Health system for predicting acute kidney injury 48 h in advance [14–17] and foundation models such as Med-PaLM-M and GatorTron for in-hospital mortality prediction from unstructured electronic health records [18–20]. Medical imaging remains the most mature area of AI application, with over 850 FDA-cleared AI-enabled medical devices as of November 2025, of which more than 70% are related to imaging [21].

2. Materials and Methods

This is a narrative review synthesizing the most clinically impactful applications of AI in internal medicine, with a focus on high-quality evidence published in recent years. The narrative format was chosen because the field is rapidly evolving, with heterogeneous study designs, outcomes, and regulatory pathways, making a formal meta-analysis or systematic review with quantitative synthesis currently premature and impractical. Studies were selected through a comprehensive review of PubMed, Google Scholar, and relevant conference proceedings (up to November 2025), focusing on high-impact publications (e.g., prospective trials, randomized controlled trials where available, and real-world evidence studies) reporting AI applications in clinical settings. Inclusion criteria prioritized studies demonstrating clinical impact, defined here as improvements in diagnostic accuracy,

prognostic stratification, workflow efficiency, or hard clinical outcomes (e.g., reduced hospitalizations, major adverse events, or mortality). Retrospective studies were included for emerging fields, but emphasis was placed on prospective validations and FDA-cleared devices. Summary tables compile key recent reports per specialty.

Inclusion criteria

Studies were included if they met at least one of the following:

- Prospective randomized controlled trials (RCTs) evaluating AI interventions
- Large prospective cohort studies with real-world implementation of AI tools
- Post hoc analyses of major prospective trials that incorporated
- Studies demonstrating improvement in clinically relevant outcomes or resulting in regulatory approval
- Publication date from 2010

Exclusion criteria

- Purely retrospective studies without prospective external validation
- Proof-of-concept or preclinical studies lacking patient-centered clinical endpoints
- Non-English language publications

No formal quantitative synthesis or meta-analysis was performed. The selection of studies was guided by clinical impact, methodological rigor, and regulatory relevance rather than exhaustive screening of all potentially eligible records.

3. Results

Artificial intelligence shows increasingly mature clinical deployment across internal medicine specialties, with cardiology leading in both volume and outcome evidence. As of November 2025, over 850 AI-enabled medical devices have received FDA clearance, of which more than 70% are imaging-related. Emerging evidence from prospective randomized trials suggests that AI integration into routine care can reduce hard clinical endpoints (hospitalizations, major adverse cardiovascular events, missed cancers) while potentially decreasing radiologist and clinician workload. The following sections present specialty-specific results from the most recent high-impact trials and real-world implementations, organized by organ system and supported by dedicated summary tables.

3.1. Cardiology

Artificial intelligence is increasingly integrated into cardiology, supporting early diagnosis, risk stratification, personalized treatment, and disease management (Figure 2). Table 1 summarizes selected high-impact prospective studies published in recent years [22,23]. Cardiology has been at the forefront of AI adoption in clinical practice. Regarding heart failure (HF), AI leveraging machine learning and deep neural networks enables early diagnosis, phenotyping, and stratification of disease severity by integrating multimodal data, including imaging, ECG, and electronic health records. These approaches achieve significantly better prediction of adverse outcomes and therapeutic response compared to conventional methods [24,25]. Moreover, AI facilitates the identification of asymptomatic structural changes and disease progression, supporting personalized care planning and timely interventions that are increasingly transforming routine clinical practice [26,27]. Diagnostic accuracy in heart failure reaches approximately 85% when using raw imaging data from echocardiography, computed tomography, or magnetic resonance imaging [27].

https://doi.org/10.3390/a19010054

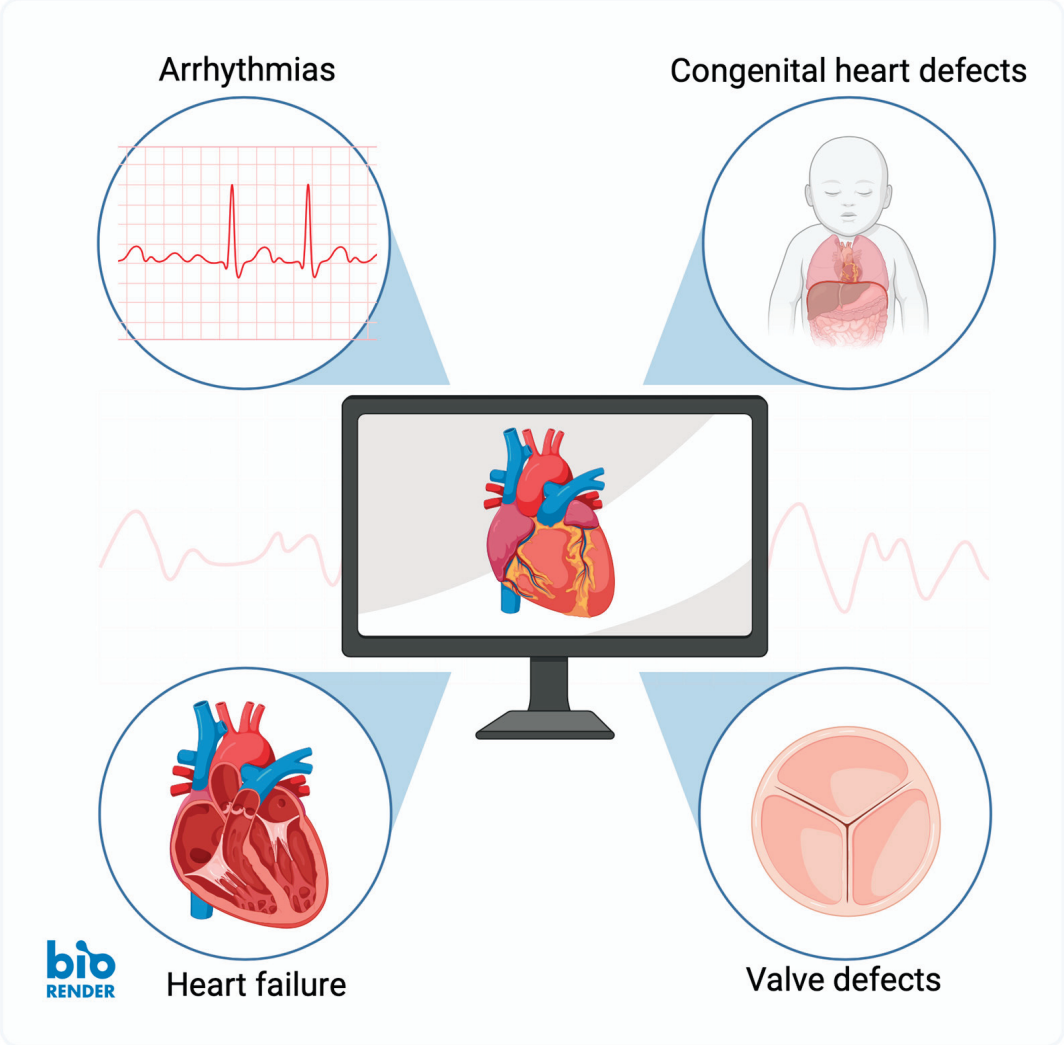


Figure 2. Main applications of artificial intelligence in cardiology.

While AI provides substantial support to clinicians, challenges remain, including discrepancies between diagnostic algorithms (e.g., up to 40% disagreement between HFA-PEFF and HFA2FPEF models in HF with preserved ejection fraction), over-reliance on left ventricular ejection fraction, and interpretive limitations [28,29]. These issues highlight the need for resolution prior to widespread implementation. Beyond heart failure, AI enhances imaging quality and diagnostics in congenital heart defects by reducing inter-observer variability, improving segmentation, and predicting long-term outcomes [30]. In postoperative settings, it detects subtle trends for early complication identification, such as catheter thrombosis. Additionally, AI contributes to drug development and targeted therapies [31], while at the population level, it supports early detection, resource allocation, remote monitoring, and large-scale research through registry and biobank analysis [32].

Table 1. Summary of the latest reports in the field of cardiology.

Authors	Methodology	Results	Significance
Deisenhofer, I., et al. [33]	Multicenter, randomized, controlled, double-blind superiority trial; patients with drug-refractory persistent/long-standing persistent AF were 1:1 randomized to either conventional anatomical PVI-only or PVI plus AI-guided ablation of areas showing spatio-temporal electrogram dispersion.	At 12 months after a single procedure, freedom from documented AF was 88% in the tailored (AI-guided + PVI) arm vs. 70% in the PVI-only arm (log-rank $p < 0.0001$). No significant difference in freedom from any atrial arrhythmia. Procedure and ablation times were twice as long in the tailored arm; safety outcomes were similar between groups.	AI-guided targeting of spatio-temporal dispersion areas significantly improves 1-year AF-free survival compared to PVI alone in persistent/long-standing persistent AF. This establishes a new, more effective ablation strategy beyond standard PVI, although longer procedures and potential need for additional tachycardia ablation should be considered.
Kim, Y., et al. [34]	Multicenter, randomized (1:1), controlled trial involving 400 patients undergoing PCI; comparison of fully automated real-time AI-based quantitative coronary angiography (AI-QCA)-assisted PCI versus intravascular OCT-guided PCI. Primary endpoint: post-PCI minimal stent area (MSA) measured by OCT, tested for noninferiority (margin 0.8 mm^2).	Post-PCI MSA was 6.3 ± 2.2 mm^2 (AI-QCA) vs. 6.2 ± 2.2 mm^2 (OCT) (difference -0.16 mm^2; 95% CI -0.59 to 0.28; P for noninferiority < 0.001). Most OCT-defined endpoints (stent underexpansion, dissection, untreated reference disease) were similar; stent malapposition was higher in the AI-QCA group (13.6% vs. 5.6%, $p = 0.007$).	AI-QCA-assisted PCI is noninferior to OCT-guided PCI in achieving optimal stent expansion (MSA) while being faster and not requiring additional imaging equipment or expertise. It offers a practical, fully automated alternative to intravascular imaging guidance for everyday PCI with comparable stent optimization outcomes.
Liu, W. T., et al. [35]	Open-label, cluster-randomized controlled trial at two hospitals in Taiwan; noncardiologists were randomized by cluster to either receive real-time AI-ECG alerts for undetected AF in at-risk patients (CHA_2DS_2-VASc ≥ 1 M/≥ 2 F) or usual care without alerts (NCT05127460).	In patients with AI-detected AF, NOAC prescription within 90 days was significantly higher in the intervention group (23.3% vs. 12.0%; HR 1.85, 95% CI 1.11–3.07). New AF diagnosis rate was also higher (HR 1.40, 95% CI 1.03–1.90). No differences in echocardiogram ordering, cardiology referrals, ischemic stroke, CV death, or all-cause death.	Simple AI-ECG alerts substantially increased AF detection and guideline-directed NOAC prescribing by noncardiologists, narrowing the care gap with cardiologists. This low-cost, scalable intervention can improve stroke prevention in undiagnosed AF without increasing downstream testing or hard clinical events.
Tsai, D. J., et al. [36]	Pragmatic randomized controlled trial at a single academic center in Taiwan; 13,631 inpatients under non-cardiologist care were 1:1 randomized to AI-ECG interpretation (low-EF probability displayed) versus standard ECG care without AI results.	New low EF ($\leq$50%) diagnoses within 30 days were significantly higher in the intervention group (1.5% vs. 1.1%; HR 1.50, 95% CI 1.11–2.03). Effect was stronger in AI-flagged high-risk patients (13.0% vs. 8.9%; HR 1.55). Positive predictive value of echocardiograms for low EF rose from 20.2% to 34.2% ($p < 0.001$) with no increase in overall echo utilization; cardiology consultations increased in high-risk patients.	A simple AI-ECG tool significantly improved early detection of low ejection fraction in routine inpatient care without raising resource use. It enhanced diagnostic yield of downstream testing and facilitated more appropriate specialist referral, demonstrating an efficient way to close the gap in heart failure diagnosis by non-cardiologists.

https://doi.org/10.3390/a19010054

Table 1. *Cont.*

Authors	Methodology	Results	Significance
Kolossváry, M., et al. [37]	Post hoc analysis of the SCOT-HEART trial; coronary CT angiography from 1750 patients was segmented and analyzed for both conventional attenuation-based plaque burden and advanced radiomic features (eigen radiomic descriptors of plaque morphology). Univariable and multivariable Cox models plus Harrell's C-statistic and time-dependent AUC with cross-validation assessed incremental prognostic value for fatal/nonfatal myocardial infarction over median 8.6 years	82 myocardial infarctions occurred. Eight radiomic features remained independently associated with MI after adjustment for cardiovascular risk score and plaque burden. Adding plaque burden to a clinical model did not improve discrimination (C-statistic 0.70 → 0.70), but further adding radiomic features increased performance to C-statistic 0.74, with significantly higher cumulative/dynamic AUC after year 5.	Radiomics-based detailed plaque morphology characterization substantially improves long-term MI risk prediction beyond clinical factors, calcium score, stenosis, and simple plaque burden. This precision-phenotyping approach from routine coronary CTA identifies higher-risk plaques and could enable better risk stratification and targeted prevention.
Trivedi, R., et al. [38]	Qualitative semistructured interviews with purposive sampling of 30 patients with atrial fibrillation who completed a 6-month fully automated voice-based conversational AI intervention (weekly AI phone calls with speech recognition and natural language processing) as part of the CHAT-AF trial; thematic analysis of transcribed interviews.	Four main themes emerged: (1) AI interactions felt human-like yet limited by scripted responses and trusted because hospital-delivered; (2) engagement depended on personalization, novel content, manageable information volume, and multichannel flexibility; (3) AI improved perceived access to continuous AF care and information; (4) patients felt empowered in self-management through reminders and reassurance from linked rhythm-monitoring devices.	Patients with AF found conversational AI an acceptable and engaging tool for education and self-management support, particularly when personalized and hospital affiliated. Findings highlight the value of voice-based AI in bridging care gaps while identifying key areas (natural dialog flow, tailored content, and information dosing) for future improvement.
Mekonnen, D., et al. [39]	Post hoc imaging sub-analysis of the RESUS-AMI trial; AI-assisted echocardiographic software (CAAS Qardia 2.0) was used to measure GLS (fully automated and semi-automated), LVEF, and volumes in 169 patients after primary PCI for STEMI. Results were correlated with CMR-derived infarct size, LVEF, and volumes ($n = 81$); intra- and inter-observer reproducibility of AI-derived parameters was assessed using ICC, bias, and limits of agreement.	AI-derived GLS showed moderate-to-good correlation with CMR infarct size (r = 0.58 automated, r = 0.64 semi-automated; both $p < 0.001$) and CMR LVEF (r = −0.63 and −0.65). Correlation with echo-derived LVEF was r = −0.51 (automated) and r = −0.67 (semi-automated). Inter- and intra-observer reproducibility of GLS was excellent (ICC 0.93–0.94).	AI-assisted GLS provides a reproducible and reliable marker of infarct size and LV systolic function after STEMI, with good correlation to gold-standard CMR. It enables fast, operator-independent strain assessment in routine post-PCI echocardiography, potentially improving risk stratification and follow-up.

https://doi.org/10.3390/a19010054

Table 1. *Cont.*

Authors	Methodology	Results	Significance
Williams, M. C., et al. [40]	Retrospective analysis of the SCOT-HEART trial (n = 1769); two separate XGBoost machine learning models with 10-fold cross-validation and grid-search hyperparameter tuning were trained on clinical variables (symptoms, demographics, risk factors, ECG, exercise tolerance testing) to predict (1) any coronary artery disease and (2) increased low-attenuation plaque (LAP) burden on CCTA.	ML model predicted any CAD significantly better than the 10-year CV risk score alone (AUC 0.80 vs. 0.75, $p = 0.004$); key features: CV risk score, age, sex, total cholesterol, abnormal ETT. The model predicting high LAP burden showed no significant improvement over the CV risk score (AUC 0.75 vs. 0.72, $p = 0.08$).	Machine learning using readily available clinical data meaningfully improves pre-test prediction of obstructive CAD on CCTA, potentially optimizing selection for imaging. However, clinical variables alone are insufficient to reliably predict high-risk (low-attenuation) plaque burden, suggesting imaging-derived features remain essential for identifying vulnerable plaques.
Saklica, D., et al. [41]	Randomized controlled trial with 52 CAD patients allocated to three groups: telerehabilitation (TRG, n = 18), mobile app-based rehabilitation (MAG, n = 13), or control (physical activity advice only, CG, n = 21). All intervention groups followed a 12-week supervised calisthenic/resistance program (3×/week). Outcomes: exercise capacity (Incremental Shuttle Walk Test), QoL (SF-36), adherence, and patient feedback analyzed with fine-tuned BERT NLP model plus anomaly detection.	Both TRG (+87.2 m) and MAG (+89.4 m) significantly improved ISWT distance vs. CG (+10.9 m; $p = 0.001$). Adherence was markedly higher in TRG (100%) and MAG (80%) than CG (30%; $p < 0.001$). NLP-analyzed patient satisfaction strongly correlated with ISWT gains (r = 0.75, $p < 0.001$); AI anomaly detection identified adherence–outcome mismatches.	Technology-supported cardiac rehabilitation (tele- or app-based) substantially outperforms usual-care advice in improving exercise capacity and adherence. AI tools (NLP for sentiment analysis and anomaly detection) provide objective, scalable enhancement of outcome evaluation and patient engagement monitoring in cardiac rehabilitation.
Trivedi, R., et al. [42]	Single-blinded, 4:1 randomized controlled feasibility trial; 103 post-discharge AF patients allocated to 6 months of fully automated conversational AI phone calls (speech recognition + NLP) with self-management support, symptom monitoring, triggered clinical alerts, supplementary SMS/email surveys, nudges, and an educational website versus usual care.	Trial stopped early (103/385 planned). No significant between-group difference in AFEQT QoL score at 6 months (adjusted mean difference 2.08, 95% CI −7.79 to 11.96; $p = 0.46$). Within the intervention group, AFEQT improved significantly from baseline (69.9 to 79.9; $p = 0.01$). Engagement was moderate (average 4/7 outreaches completed); 88.4% of completed contacts rated useful.	Conversational AI delivered via phone calls is feasible, acceptable, and engaging for post-discharge AF support, with high perceived usefulness and preliminary evidence of within-group QoL improvement. Despite early termination preventing definitive efficacy assessment, it establishes proof-of-concept for scalable, fully automated patient support in chronic AF management.

https://doi.org/10.3390/a19010054

Table 1. *Cont.*

Authors	Methodology	Results	Significance
Fiolet, A. T. L., et al. [43]	Cross-sectional subanalysis of the LoDoCo2 trial; 151 patients with chronic coronary disease on stable therapy underwent coronary CTA after median 28.2 months of blinded low-dose colchicine (0.5 mg/day) or placebo. AI-enabled software quantified pericoronary adipose tissue (PCAT) attenuation and total/detailed plaque volumes (non-calcified, low-attenuation, calcified, dense-calcified) across the entire coronary tree.	No difference in pericoronary inflammation (PCAT attenuation: −79.5 HU colchicine vs. −78.7 HU placebo, $p = 0.236$). Colchicine group showed significantly higher calcified plaque volume (169.6 vs. 113.1 mm^3, $p = 0.041$), calcified plaque burden (9.6% vs. 7.0%, $p = 0.035$), and dense calcified plaque volume (192.8 vs. 144.3 mm^3, $p = 0.048$). Low-attenuation plaque burden was lower with colchicine only in patients on low-intensity statins (p-interaction = 0.037).	Low-dose colchicine does not reduce pericoronary adipose tissue inflammation but promotes coronary plaque calcification and increases dense calcified plaque—features associated with greater plaque stability. This provides a mechanistic explanation for the observed cardiovascular event reduction in LoDoCo2 and supports plaque stabilization as a key anti-atherosclerotic effect of colchicine.
Li, G., et al. [44]	Post hoc blinded analysis of the prospective CAREER trial (NCT04665817); fully automatic AI-based CCTA reconstruction and CT-μFR computation (Murray-law based quantitative flow ratio) performed in 242 patients (657 vessels) who had invasive coronary angiography with FFR or μFR within 30 days. Reference standard: invasive FFR ≤ 0.80 or μFR ≤ 0.80.	Fully automatic CT-μFR was successful in all cases with mean analysis time 1.60 ± 0.34 min/patient. CT-μFR showed good correlation (r = 0.62) and agreement (bias -0.01 ± 0.10) with invasive physiology. Patient-level diagnostic accuracy was 83.0% (sensitivity 84.2%, specificity 81.9%, PPV 82.1%, NPV 84.0%).	Fully automatic, AI-powered CT-μFR provides rapid (~1.6 min), operator-independent functional assessment directly from routine CCTA with high diagnostic performance comparable to invasive FFR. It enables accurate pre-catheterization identification of hemodynamically significant stenoses, potentially reducing unnecessary invasive procedures
Yu, X., et al. [45]	Single-center, open-label, randomized controlled trial (n = 2086 post-PCI CHD patients in China); 1:1 randomization to a comprehensive web-based telemedicine platform (personalized education, medication reminders, vital sign monitoring, AI-assisted consultations) plus usual care versus usual care alone (phone follow-up at 1, 3, 6, 12 months). Primary endpoint: 1-year MACCE (cardiac death, MI, stroke, target vessel revascularization).	Telemedicine significantly reduced 1-year MACCE (3.5% vs. 5.3%, $p = 0.04$), driven by lower cardiac death (1.0% vs. 2.3%, $p = 0.02$) and MI (0.8% vs. 1.8%, $p = 0.03$). Serious bleeding (BARC 3–5) was lower (0.6% vs. 1.6%, $p = 0.03$). Intervention group achieved better BP control, higher adherence to aspirin and ACEI/ARB/ARNI, and reduced alcohol consumption; smoking showed a favorable trend.	A multicomponent AI-supported telemedicine program significantly lowered hard clinical events (MACCE) and improved secondary prevention metrics at 1 year after PCI. This demonstrates that scalable, technology-driven remote management can close implementation gaps and meaningfully improve long-term outcomes in high-risk CHD patients.

https://doi.org/10.3390/a19010054

Table 1. *Cont.*

Authors	Methodology	Results	Significance
Ishiguchi, H., et al. [46]	Post hoc analysis of the WARCEF trial; 2213 HFrEF patients without atrial fibrillation were included. Nine machine learning models (including SVM, XGBoost, LightGBM) were trained on 12 selected clinical/demographic variables to predict incident ischaemic stroke during mean 3.3-year follow-up (74 events). Model performance was compared to CHA_2DS_2-VASc using AUC and decision curve analysis; feature importance assessed via SHAP values.	ML models strongly outperformed CHA_2DS_2-VASc (AUC 0.643). Best-performing models were SVM (AUC 0.874, 95% CI 0.769–0.959) and XGBoost (AUC 0.873, 95% CI 0.783–0.953), with SVM and LightGBM showing consistent net clinical benefit. Top predictive features across models: creatinine clearance, blood urea nitrogen, and warfarin use.	Machine learning substantially improves risk stratification for ischaemic stroke in HFrEF patients in sinus rhythm compared to traditional scores. Renal function markers (CrCl, BUN) and anticoagulation status emerged as dominant predictors, highlighting new targets for stroke prevention beyond AF in this high-risk population.

https://doi.org/10.3390/a19010054

Recent 2025 studies demonstrate progress in the application of artificial intelligence in cardiology, while also highlighting several important limitations and areas requiring further evaluation. In atrial fibrillation ablation, AI improves efficacy [33], but at the cost of twice the procedure length and a higher risk of complications. AI-ECG increases the detection of AF and low ejection fraction [35,36], but this does not translate into a reduction in strokes or deaths—early detection does not always translate into real prevention. Voice AI for patients with AF is promising [38,42], but the studies are small, often uncontrolled, and have been terminated early—the evidence remains weak. In coronary artery disease, AI-QCA produces results like OCT [34], but higher stent malposition raises concerns about long-term safety. Radiomics with CTA improves myocardial infarction prediction [37], but the C-statistic gain is modest ($0.70 \rightarrow 0.74$) and may be insufficient in clinical practice. Automated CT-μFR is fast and accurate [44], but requires validation in large, diverse populations. Studies from 2025 show that artificial intelligence in cardiology is making rapid progress, but many results raise significant concerns. In atrial fibrillation ablation, AI improves efficacy [33], but at the cost of twice the procedure length and a higher risk of complications. AI-ECG increases the detection of AF and low ejection fraction [35,36], but this does not translate into a reduction in stroke or death—early detection does not always translate into real prevention. Voice AI for patients with AF is promising [38,42], but studies are small, often uncontrolled, and terminate early—the evidence remains weak. In coronary artery disease, AI-QCA produces results like OCT [34], but higher stent malapposition raises concerns about long-term safety. Radiomics with CTA improves myocardial infarction prediction [37], but the C-statistic gain is modest ($0.70 \rightarrow 0.74$) and may be insufficient in clinical practice. Automated CT-μFR is fast and accurate [44], but requires validation in large, diverse populations. ML models predict stroke better in HFrEF than CHA_2DS_2-VASc [46], but the lack of randomization and prospective validation limits their readiness for use. Tele-/application-based rehabilitation and telemedicine improve adherence and performance [41,45], but studies are small or conducted in specific settings—scalability in the real world remains uncertain. AI offers valuable tools, but most evidence comes from studies of limited scale, short follow-up, or without hard endpoints (death, stroke, myocardial infarction). We often demonstrate statistical differences, not clinically significant improvements in prognosis. Large, multicenter, randomized trials with long-term follow-up and hard endpoints are needed before routine implementation. For now, AI is a valuable addition to cardiology, but it is far from a true revolution.

3.2. Pulmonology

Building on its success in cardiovascular imaging, AI is now making significant inroads in respiratory medicine (Table 2). It is contributing to advancements in the diagnosis and management of conditions such as lung cancer, interstitial lung disease (ILD), and obstructive lung diseases [47]. Primarily through analysis of chest computed tomography, deep learning algorithms particularly convolutional neural networks (CNNs), enable automatic detection of pulmonary abnormalities, malignancy risk assessment, and longitudinal monitoring [48]. These models often match or approach expert-level accuracy in lung cancer classification, facilitating faster diagnosis, improved prognosis, personalized treatment planning, and even simulation of treatment response using high-resolution reconstructions [49]. Despite these promising results, ongoing challenges include the need for robust clinical validation, seamless system integration, and resolution of ethical concerns.

https://doi.org/10.3390/a19010054

Table 2. Summary of the latest reports in the field of pulmunology.

Authors	Methodology	Results	Significance
Ren, F., et al. [50]	AI-driven target identification selected TNIK as a high-confidence anti-fibrotic target. A generative AI platform designed and optimized the small-molecule TNIK inhibitor INS018_055. The compound was tested in multiple preclinical in vivo models of lung, liver, skin, and kidney fibrosis (oral, inhaled, and topical routes) and in anti-inflammatory models. Safety, tolerability, and pharmacokinetics were assessed in two randomized, double-blind, placebo-controlled phase I trials (NCT05154240 and CTR20221542) involving 78 healthy volunteers.	INS018_055 showed potent pan-organ anti-fibrotic activity and additional anti-inflammatory effects across preclinical models via all tested administration routes. Both phase I trials confirmed favorable safety, tolerability, and pharmacokinetic profiles with no serious adverse events. The entire process from target identification to clinical candidate nomination was completed in approximately 18 months.	This study demonstrates the first successful clinical translation of a generative AI-discovered drug for idiopathic pulmonary fibrosis. By identifying TNIK as a novel anti-fibrotic target and rapidly delivering a safe, orally/inhalably bioavailable inhibitor (INS018_055), it validates an ultra-fast AI-powered drug discovery pipeline capable of dramatically accelerating therapeutic development for high-unmet-need fibrotic diseases.
Xu, Z., et al. [51]	Multicenter, double-blind, randomized, placebo-controlled phase 2a trial (NCT05938920); 71 patients with idiopathic pulmonary fibrosis were randomized 1:1:1:1 to rentosertib (AI-discovered first-in-class TNIK inhibitor) 30 mg QD, 30 mg BID, 60 mg QD, or placebo for 12 weeks. Primary endpoint: safety and tolerability (treatment-emergent adverse events). Key secondary endpoints: pharmacokinetics and change in forced vital capacity (FVC).	Safety profile was favorable and comparable across groups (TEAEs 70–83%); serious treatment-related AEs were rare, with liver toxicity and diarrhea as main reasons for discontinuation. At the highest dose (60 mg QD), FVC increased by +98.4 mL (95% CI 10.9–185.9) versus a decline of −20.3 mL (95% CI −116.1 to 75.6) with placebo, indicating a dose-dependent signal of lung function stabilization/improvement.	Rentosertib is the first fully AI-generated novel small molecule (new target + new chemical entity) to demonstrate clinical proof-of-concept in a phase 2 trial. The positive safety profile and early evidence of FVC benefit in IPF validate generative AI as a viable end-to-end drug discovery engine capable of delivering clinical-stage candidates for previously intractable diseases
Habicher, M., et al. [52]	Single-center, single-blinded randomized controlled trial; 150 patients undergoing lung surgery with single-lung ventilation were 1:1 randomized to either AI-based goal-directed hemodynamic therapy using the Hypotension Prediction Index (HPI) or standard care without a specific protocol.	The HPI-guided intervention significantly reduced the number of hypotensive episodes (0 [0–1] vs. 1 [0–2], $p = 0.01$), duration of hypotension (0 vs. 2.33 min, $p = 0.01$), area under MAP < 65 mmHg (0 vs. 10.67 mmHg·min, $p < 0.01$), and time-weighted average below MAP 65 (0 vs. 0.07 mmHg, $p < 0.01$). Postoperative AKI rates were similar (6.7% vs. 4.2%, $p = 0.72$); MINS showed a strong trend toward reduction (17.1% vs. 31.8%, $p = 0.07$), with a similar trend for postoperative infections.	AI-driven predictive hemodynamic management using the HPI effectively minimized intraoperative hypotension during single-lung ventilation surgery. Although AKI was not significantly reduced, the observed trends toward lower myocardial injury and infections suggest potential clinical benefit, supporting broader adoption of predictive AI tools for perioperative hemodynamic optimization.

https://doi.org/10.3390/a19010054

Table 2. *Cont.*

Authors	Methodology	Results	Significance
Hong, L., et al. [53]	Two prospective randomized controlled trials in COPD patients: 1. A 12-month RCT with 447 patients assessing AI-based medical intervention vs. standard care, with quality-of-life (QoL) and psychological outcomes measured at 4 and 12 months. 2. A separate 9-month RCT with 101 patients randomized to a web-based AI-driven educational and exercise program for prevention of acute exacerbations versus control	At 4 months no significant QoL improvement was observed. At 12 months the AI-intervention group showed significantly better quality of life, emotional status, and psychological well-being compared with controls. In the second trial, the AI-supported group had lower hospitalization rates and shorter length of hospital stay than the control group. Single-factor analysis did not reach statistical significance in all outcomes, but overall results were positive.	AI-supported medical and educational interventions appear feasible and effective in long-term COPD management, improving quality of life, psychological health, and reducing acute exacerbations and hospitalizations after 9–12 months. Although some individual endpoints lacked statistical power, the trials provide preliminary evidence supporting the integration of artificial intelligence into routine COPD care and secondary prevention.
Ladbury, C., Li, R., et al. [54]	Secondary analysis of the RTOG 0617 trial; multiple machine learning models (including XGBoost, random forest, naive Bayes) were trained on clinical and dosimetric variables to predict grade ≥ 3 pulmonary, cardiac, and esophageal toxicity after definitive chemoradiation for locally advanced NSCLC. Best models were interpreted using SHAP values to identify and quantify dosimetric thresholds, validated with logistic regression and bootstrapping.	XGBoost best predicted pulmonary toxicity (AUC 0.739), random forest cardiac (AUC 0.706), and naive Bayes esophageal toxicity (AUC 0.721), all outperforming traditional logistic regression. Key thresholds: lung mean dose > 18 Gy and V20 > 37% for pulmonary toxicity (OR 2.47 and 2.72); esophageal mean dose > 34 Gy and V20 > 37% for esophageal toxicity (OR 4.01 and 3.73). No significant cardiac thresholds identified.	Machine learning combined with explainable AI (SHAP) validated known and identified new clinically actionable dosimetric thresholds for radiation-induced toxicity, outperforming conventional statistical methods. This approach enables data-driven, precise optimization of radiotherapy planning constraints in lung cancer chemoradiation.
Pasipanodya, J. G., et al. [55]	Nested pharmacokinetic substudy within the OFLOTUB trial; 126 patients with drug-susceptible pulmonary TB received a 4-month gatifloxacin-containing regimen. Intensive PK sampling was performed on two days in the first 2 months. Therapy failure was defined as failure to culture-convert, relapse (confirmed by spoligotyping), or death within 24 months. An ensemble machine-learning approach (multiple algorithms) was used to rank 27 clinical/laboratory/PK variables and detect interactions predicting outcome.	19/126 patients (15%) had unfavorable outcomes. Machine learning ranked pyrazinamide and rifampicin exposure (Cmax and AUC) as more important than gatifloxacin exposure. A significant antagonistic 3-way interaction between low concentrations of pyrazinamide, gatifloxacin, and rifampicin was identified; this negative interaction disappeared when rifampicin Cmax exceeded 7 mg/L. Drug concentrations explained 31–75% of outcome variance across sites.	Concentration-dependent antagonism among the three key drugs contributes to treatment failure in shortened TB regimens but can be overcome by higher rifampicin exposure (>7 mg/L). The findings provide a mechanistic explanation for prior trial failures and strongly support dose optimization of both rifampicin and fluoroquinolones to improve efficacy of short-course TB regimens.

https://doi.org/10.3390/a19010054

Table 2. *Cont.*

Authors	Methodology	Results	Significance
Ding, Y., et al. [56]	Retrospective analysis of 424 patients with pulmonary nodules undergoing surgical resection; all had preoperative CT-based AI malignancy probability score, 7-autoantibody (7-AAB) panel, and CEA testing. Patients were randomly split 1:1 into training (n = 212) and validation (n = 212) sets. A logistic regression-based nomogram was built using forward stepwise selection of significant predictors (age, AI score, 7-AAB result, CEA) and internally validated.	The nomogram achieved AUC 0.899, sensitivity 82.3%, specificity 90.5%, and PPV 97.2% in the validation set. It significantly outperformed 7-AAB alone (sensitivity 82.3% vs. 35.9%, $p < 0.001$), CEA alone (sensitivity 82.3% vs. 18.8%, $p < 0.001$), and standalone CT-AI (specificity 90.5% vs. 69.0%, $p = 0.022$). For nodules ≤ 2 cm, nomogram specificity remained high at 90.0% vs. 67.5% for AI alone ($p = 0.022$).	Integration of CT-based AI, 7-AAB panel, age, and CEA into a simple nomogram substantially improves diagnostic accuracy for pulmonary nodules, especially small ($\leq$2 cm) lesions, offering higher sensitivity than biomarkers alone and higher specificity than AI alone. This non-invasive, clinically practical tool can reduce unnecessary surgeries while maintaining excellent malignancy detection.
Bosman, S., et al. [57]	Prospective diagnostic accuracy study at health facilities in Lesotho and South Africa; 1392 symptomatic adults ($\geq$1 TB symptom) underwent digital chest X-ray with CAD4TBv7 analysis, point-of-care CRP, and microbiological confirmation (Xpert MTB/RIF Ultra + liquid culture as composite reference standard). CAD4TBv7 performance was compared to CRP and expert radiologist reading.	CAD4TBv7 AUC was 0.87 (95% CI 0.84–0.91) vs. 0.80 for CRP (95% CI 0.76–0.84). At $\geq$90% sensitivity, CAD4TBv7 specificity was 68.2% (95% CI 65.4–71.0%), nearly meeting WHO TPP (>70%), while CRP specificity was only 38.2%. CAD4TBv7 performed equivalently to expert radiologist interpretation.	CAD4TBv7 (version 7) is a highly accurate, non-sputum triage test that nearly achieves WHO TPP criteria in high TB/HIV-burden settings, enabling rapid rule-out of TB and prioritization of confirmatory testing. It significantly outperforms CRP and matches expert radiology, supporting its deployment as a scalable frontline TB screening tool.

https://doi.org/10.3390/a19010054

In recent years, AI has begun to play an increasingly important role in various stages of the development and implementation of respiratory disease therapies—from the discovery of new molecular targets, through drug design and dose optimization, to diagnostics and treatment monitoring. The first groundbreaking example is the path taken by the TNIK kinase inhibitor INS018_055 (later named rentosertib) from AI-assisted target identification to obtaining clinical evidence in Phase 2. Ren and colleagues [50] were the first to demonstrate that generative AI can complete the entire pathway in just 18 months: from selecting a high-potency target (TNIK) in fibrotic diseases, through designing a new chemical molecule, to nominating a clinical candidate. Importantly, INS018_055 demonstrated potent pan-organ activity (lungs, liver, skin, kidneys) in preclinical models and additionally possessed anti-inflammatory properties. Phase I studies in healthy volunteers confirmed an excellent safety profile and pharmacokinetics after both oral and inhaled/topical administration. The same molecule was subsequently evaluated in a randomized, placebo-controlled phase 2a study in patients with idiopathic pulmonary fibrosis (IPF) [51]. Despite the short treatment period (12 weeks) and small number of patients (n = 71), a dose-dependent clinical signal was observed: at the highest dose (60 mg once daily), there was an increase in FVC of +98.4 mL compared to a decrease of −20.3 mL in the placebo group. Crucially, the safety profile remained very favorable. This is the world's first fully AI-generated novel small molecule (new target + new chemical entity) to achieve clinical proof in phase 2 in a disease with a high unmet need and poor prognosis. It demonstrates that generative AI can dramatically shorten the time from concept to clinical evidence (traditionally 10–15 years → here ~2–3 years to Phase 2). This paves the way for faster treatment of rare and challenging fibrotic diseases. Another area where AI is beginning to provide tangible clinical benefits is hemodynamic management during surgery. Habicher et al. [52] conducted a randomized trial in patients undergoing lung surgery with single-lung ventilation—a very challenging hemodynamic situation. The group using the HPI (Hypotension Prediction Index) algorithm had significantly fewer episodes of hypotension, shorter duration, and less exposure to low mean arterial pressure (MAP < 65 mmHg). Although the incidence of acute kidney injury (AKI) did not differ significantly, a clear trend towards fewer episodes of myocardial infarction (MINS) and postoperative infections was observed. This demonstrates that AI can be proactive, predicting hypotension minutes in advance and enabling staff to intervene quickly. Even if not all complications were significantly reduced, the trends suggest real potential for improved outcomes in high-risk patients. Hong et al. [53] went further, examining whether AI can improve quality of life and clinical outcomes in outpatient COPD patients. In two independent randomized trials (12-month and 9-month), an AI-based intervention (telemedicine, education, exercise, reminders) resulted in significant improvements in quality of life, emotional well-being, and a reduction in the number of exacerbations requiring hospitalization after 9–12 months. These effects only became apparent after a longer period, which is typical of behavioral and educational interventions. This demonstrates that AI can effectively support long-term adherence and secondary prevention in chronic respiratory diseases. Ladbury et al. [54] demonstrated ML + SHAP for determining safe dose thresholds in NSCLC radiotherapy—better AUC than classical logistic regression and specific, clinically useful toxicity thresholds. Pasipanodya et al. [55] also demonstrated ML for detecting complex pharmacokinetic interactions in tuberculosis treatment—explaining why abbreviated regimens with gatifloxacin do not always work and indicating how to improve them. Ding et al. [56] and Bosman et al. [57] used AI combined with biomarkers or AI standalone to improve lung nodule diagnosis and tuberculosis screening with very high accuracy and the potential to reduce unnecessary invasive procedures. This perspective shows that we are witnessing a true paradigm shift

https://doi.org/10.3390/a19010054

in pulmonology from "AI as a curiosity" to "AI as an integral part of modern respiratory care".

3.3. Neurology

The application of AI extends naturally to neurology, where the complexity of brain imaging and signal processing creates opportunities for advanced computational tools. Amid rapid progress in neuroscience and computer science, AI is emerging as a key asset in managing neurological disorders (Figure 3) [58]. This growing collaboration between AI developers and clinicians addresses the demand for precise, objective tools capable of detecting even subtle changes, such as those seen in Alzheimer's disease or epilepsy [59,60].

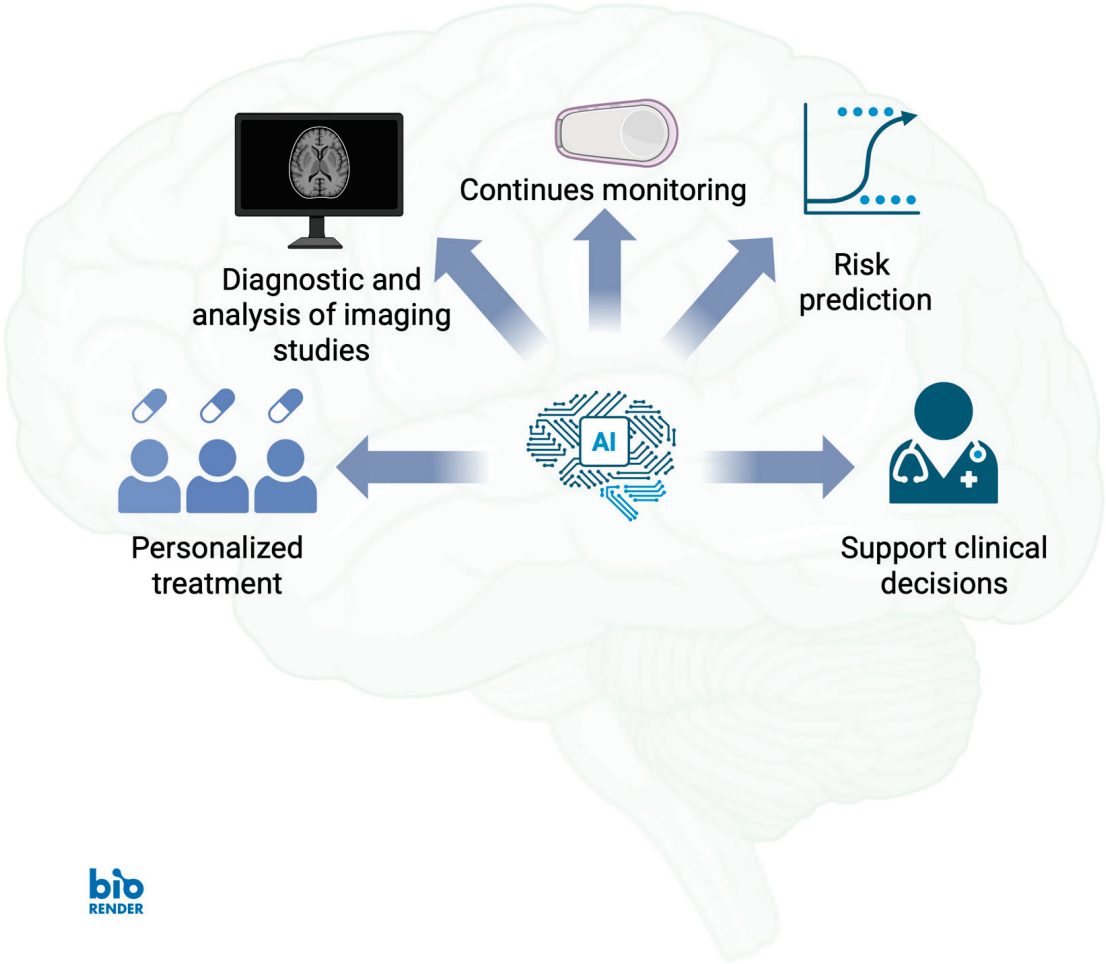


Figure 3. The role of AI in clinical neurology: supporting diagnostic imaging, personalizing treatment, predicting risk, and continuously monitoring patients to optimize clinical decisions.

Primary applications include analysis of MRI scans for anomaly detection, tracking progression of neurodegenerative diseases, decoding EEG signals for brain–computer interfaces (BCI), and supporting cognitive stimulation (Table 3) [61,62].

Table 3. Summary of the latest reports in the field of neurology.

Authors	Methodology	Results	Significance
Gombolay, G. Y., et al. [63]	Randomized, blinded vignette-based study comparing 81 neurologists/child neurologists (members of Child Neurology Society and American Academy of Neurology) with 284 general-population participants. Each received AI-based diagnostic recommendations accompanied by one of eight randomly assigned explainable AI (xAI) methods (decision tree, crowd-sourced agreement, case-based reasoning, probability scores, counterfactuals, feature importance, templated explanations, or no explanation). Primary outcomes: task performance, perceived explainability, trust, and social competence of the DSS.	Decision trees were rated significantly more explainable by neurologists than by the general population ($p < 0.01$) and more explainable than probability scores among neurologists ($p < 0.001$). Higher clinical experience and higher perceived explainability paradoxically correlated with worse task performance ($p = 0.0214$). Performance was driven by perceived explainability rather than the specific xAI technique used.	Different xAI techniques are not equally effective across clinicians and laypeople; neurologists strongly prefer decision-tree explanations, and there is no universal "best" xAI method. Perceived explainability can negatively affect diagnostic accuracy in experts, highlighting the need for clinician-centered, personalized xAI design rather than one-size-fits-all approaches in clinical decision-support systems.
Alagapan, S., et al. [64]	Prospective study in 10 patients with treatment-resistant depression receiving subcallosal cingulate (SCC) DBS with a bidirectional implant allowing chronic local field potential (LFP) recording (NCT01984710). SCC LFPs from 6 patients were analyzed using explainable AI to derive individualized electrophysiological biomarkers of clinical state. Preoperative structural/functional connectivity of the target network was quantified with tractography and resting-state fMRI; objective mood changes were measured via automated video-based facial expression analysis.	At 24 weeks, 90% of patients were responders and 70% achieved remission. Explainable AI identified stable, patient-specific SCC LFP biomarkers that accurately tracked clinical state, discriminated therapeutic from transient stimulation effects, and responded to programming changes. Recovery trajectories strongly correlated with preoperative integrity of the white-matter treatment network and were objectively mirrored by changes in data-driven facial expression metrics.	This is the first demonstration of chronic, individualized electrophysiological biomarkers from the SCC DBS target that objectively track clinical state in TRD. Combined with preoperative connectivity and automated behavioral analysis, these biomarkers enable personalized, biomarker-guided DBS management, reduce reliance on subjective reporting, and explain inter-individual variability in long-term outcome of SCC DBS for severe depression.
Boutet, A., et al. [65]	Prospective observational trial in 67 PD patients with chronic DBS; 3T fMRI was performed under clinically optimized (ON) and deliberately non-optimal (OFF) stimulation settings. Brain response patterns were compared, and a machine learning classifier (trained on 39 patients with a priori known optimal settings) was built to predict optimal vs. non-optimal stimulation from whole-brain fMRI activation maps.	Optimal DBS produced a distinct fMRI signature with strong engagement of the motor network. The ML model classified optimal vs. non-optimal settings with 88% accuracy in the training cohort and successfully generalized to unseen held-out data, including stimulation-naïve patients, correctly predicting the clinically optimal settings.	fMRI-based brain response patterns serve as an objective, patient-specific biomarker of therapeutic DBS efficacy in Parkinson's disease. This proof-of-concept demonstrates the feasibility of imaging-guided DBS programming, potentially reducing the number of clinical visits and accelerating the identification of optimal stimulation parameters.

https://doi.org/10.3390/a19010054

Table 3. *Cont.*

Authors	Methodology	Results	Significance
Cowan, R. P., et al. [66]	Cross-sectional diagnostic accuracy study at three academic headache centers; 212 adults completed both a self-administered web-based Computer-based Diagnostic Engine (CDE) and a semi-structured telephone interview (SSI) by headache specialists, both strictly applying ICHD-3 criteria. Order of administration was randomized; concordance and accuracy metrics (Cohen's kappa, sensitivity, specificity, PPV/NPV) were calculated using SSI as the reference standard.	Excellent concordance between CDE and SSI for migraine/probable migraine diagnosis (κ = 0.83, 95% CI 0.75–0.91). CDE showed sensitivity 90.1% (95% CI 83.6–94.6%), specificity 95.8% (95% CI 88.1–99.1%), PPV 97.0%, and NPV 86.6% at 60% study prevalence. At a population prevalence of 10%, PPV was 70.3% and NPV 98.9%.	A fully automated, self-administered online diagnostic tool using ICHD-3 logic achieves near-specialist-level accuracy in migraine diagnosis. It reliably rules in migraine (high specificity) and rules out migraine (high sensitivity), offering a scalable, valid solution to reduce diagnostic delay and the need for specialist interviews in both clinical and research settings.
Gorenshtein, A., et al. [67]	Prospective single-center randomized controlled trial; 200 patients referred for electrodiagnostic (EDX) studies were 1:1 randomized to physician-only interpretation (control) or physician + AI-assisted interpretation using the multi-agent INSPIRE framework (intervention). Three board-certified physicians rotated across arms. In the intervention arm, physicians reviewed and integrated an AI-generated preliminary report. Primary outcome: report quality measured by AIGERS score (0–1). Secondary outcomes: physician-rated AI integration (PAIR) and usability survey.	I-generated preliminary reports showed only moderate consistency. Final integrated physician-AI reports did not significantly outperform physician-only reports on AIGERS (no statistical difference, $p > 0.05$). Physicians rated trust in AI suggestions moderately (3.7/5) but scored efficiency (2.0/5), ease of use (1.7/5), and workload reduction (1.7/5) poorly, citing interpretability issues and workflow disruption.	In real-world clinical use, the tested AI-assisted multi-agent framework (INSPIRE) did not improve EDX report quality over expert physician interpretation alone and was perceived as cumbersome. Current limitations in usability and workflow integration highlight that AI tools for complex interpretive tasks like EDX require substantial improvement before they can deliver meaningful clinical benefit.
Davidovic, V., et al. [68]	Cross-sectional cohort follow-up of a randomized controlled trial (NCT06273579) at McGill University Neurosurgical Simulation Centre; 87 medical students were block-randomized to three training conditions on the NeuroVR virtual reality simulator: pure AI-tutor feedback, scripted human instruction, or AI-augmented personalized human instruction (human instructor received real-time AI-detected error data). Participants performed repeated simulated tumor resections. Primary measure: feedback frequency (as proxy for errors); secondary: objective performance metrics (healthy tissue removal, instrument separation, aspirator force).	By the third repetition, the AI-augmented personalized instruction group required significantly fewer total feedback instances (IRR 1.50, 95% CI 1.16–1.94, $p < 0.001$) and high-force aspirator corrections (IRR 1.71, 95% CI 1.15–2.55, $p = 0.002$) than in earlier repetitions. Compared to pure AI-tutor instruction, AI-augmented human instruction achieved significantly less healthy tissue removal ($p = 0.01$), smaller instrument tip separation (mean ratio 1.25, $p = 0.008$), and lower aspirator force (mean ratio 1.68, $p <$ 0.001), with sustained improvement across all metrics from baseline.	Real-time AI error detection fed to a human instructor (AI-augmented personalized instruction) markedly reduces feedback frequency (indicating fewer errors) and produces superior technical skill acquisition compared to either AI-only or traditional scripted human teaching. This hybrid approach leverages the strengths of both AI precision and human pedagogical adaptability, establishing a new benchmark for effective VR-based surgical training.

https://doi.org/10.3390/a19010054

Table 3. *Cont.*

Authors	Methodology	Results	Significance
Hassan, A. E., et al. [69]	Observational before-and-after study within the ongoing EMBOLISE trial (NCT04402632) at a single large comprehensive stroke center. Pre-AI period: 153 days (5 May–6 October 2021); post-AI period: 316 days after activation of Viz RECRUIT SDH (6 October 2021–18 August 2022). The AI platform automatically analyzed all non-contrast head CTs, flagged suspected subacute/chronic SDH, and calculated volume, thickness, and midline shift. All AI alerts were manually reviewed to assess positive predictive value (PPV) and enrollment impact.	Pre-AI: 5 patients enrolled (0.99/month), 1 screen failure. Post-AI: 14 patients enrolled (1.35/month), representing a 36% increase in enrollment rate and zero screen failures. Of 6244 processed CTs, 207 SDH detections (3% prevalence); PPV of the algorithm was 81.4% (95% CI 75.3–86.7%). Median response time to alerts: 50% viewed within 1 h, 35% within 10 min.	Real-time AI-based automated screening of routine head CTs (Viz RECRUIT SDH) significantly accelerated patient identification and enrollment in a randomized trial for chronic/subacute SDH by 36%, eliminated screen failures, and demonstrated high real-world performance (PPV > 80%) and rapid clinical response. This validates AI-driven mobile platforms as powerful tools to improve recruitment efficiency in time-sensitive neurosurgical trials.
Macea, J., et al. [70]	Prospective observational study within the SeizeIT2 trial (NCT04284072); 223 in-hospital overnight recordings from 50 epilepsy patients were simultaneously captured with full polysomnography (including standard EEG) and a wearable EEG + accelerometry device. A single deep-learning model performed automated 30 s epoch sleep staging on both modalities. Automated scoring was compared against consensus clinical expert scoring on 20 nights (one per patient) using accuracy, Cohen's kappa, F1-scores, and Bland–Altman analysis. Mixed-effects models compared sleep macrostructure between patients with and without in-hospital seizures.	Automated staging showed moderate agreement with expert scoring on standard EEG (accuracy 0.73, $\kappa = 0.59$) and lower agreement on wearable data (accuracy 0.61, $\kappa = 0.43$). Sensitivity was poor for N1 across both modalities; wearable-based staging systematically underestimated total sleep time and most stages except N2. Patients with seizures slept significantly longer (6.37 h, 95% CI 5.86–7.87) than those without (5.68 h, 95% CI 5.24–6.13; $p = 0.001$) and spent more time in N2.	Wearable-based deep learning can perform automated sleep staging in epilepsy patients with moderate accuracy, sufficient to detect clinically relevant differences (longer sleep and more N2 in patients with seizures). However, current performance (especially poor N1 detection and systematic underestimation by the wearable) requires further model improvement before reliable clinical or research use in epilepsy monitoring.

https://doi.org/10.3390/a19010054

Recent years have shown that artificial intelligence in neurology and neurosurgery is moving from experimental applications to real, clinically significant benefits, although still with clear limitations and the need for further refinement. A groundbreaking example is the first documentation of individual, stable electrophysiological biomarkers following deep-brain stimulation (DBS) in patients with treatment-resistant depression (TRD). Alagapan and colleagues [64] demonstrated that 90% of patients with DBS in the subcallosal cingulate (SCC) achieved a clinical response after 24 weeks, and 70% achieved remission. Explainable AI enabled the extraction of patient-specific LFP biomarkers from chronic recordings, which precisely tracked clinical status, distinguished therapeutic from transient effects, and responded to programming changes. Furthermore, the trajectory of improvement was strongly associated with preoperative white matter network integrity, and objective mood changes were reflected in automated analysis of facial expressions. This is the first evidence that DBS in TRD can be performed in a biomarker-guided manner, with minimal reliance on subjective patient assessment—paving the way for personalized, precise neuromodulation in severe psychiatric disorders. Concurrently, AI is beginning to support DBS programming in Parkinson's disease. Boutet et al. [65] demonstrated that optimal stimulation induces a characteristic pattern of motor network activation visible in fMRI, and a machine learning model trained on these patterns achieved 88% accuracy in distinguishing optimal from suboptimal settings—and, most importantly, generalized to treatment-naive patients. This is proof of concept that functional imaging can become an objective tool for rapid and precise selection of stimulation parameters, potentially reducing the time and number of visits needed to optimize DBS.In neurological diagnostics, AI also achieves results close to those of experts. Cowan et al. [66] demonstrated that a fully automated, web-based diagnostic engine (CDE) based on the ICHD-3 criteria achieved excellent agreement ($\kappa = 0.83$) with a telephone interview conducted by specialists in diagnosing migraine, with very high sensitivity (90%) and specificity (96%). This tool can significantly shorten diagnostic time and reduce specialist workload—especially in countries with limited access to neurologists. At the same time, research shows that AI's success is not automatic. Gorenshtein et al. [67], in a randomized trial comparing physician-assisted electromyography (EDX) interpretation with AI-assisted interpretation, demonstrated that the addition of AI did not improve the quality of reports compared to experts, and physicians rated the tool as inconvenient and of limited use (low ratings for effectiveness and workflow integration). This is an important signal that even advanced AI systems must be designed with real clinical needs and physician ergonomics in mind. In neurosurgical education, Davidovic et al. [68] demonstrated the clear advantage of a hybrid approach: in simulated brain tumor resection training, students who received personalized instructions from a human tutor supported by real-time AI data (errors detected by the algorithm) achieved significantly better results than with an AI tutor alone or traditional training. This indicates that AI is best suited as a precise support for human learning, not as a complete replacement. In clinical trial recruitment, Hassan et al. [69] demonstrated that automated, real-time AI-assisted head CT scan analysis (Viz RECRUIT SDH) increased the rate of patient enrollment in a randomized trial of subdural hematomas by 36%, eliminated enrollment errors, and achieved a PPV of >80%. This demonstrates how AI can accelerate clinical trials in neurosurgery and neurology. AI in neurology and neurosurgery already provides personalized biomarkers, supports the precise programming of implantable devices, accelerates diagnostics and research recruitment, and in education, it works best in hybrid form with humans. At the same time, it is becoming clear that not every AI application automatically brings benefits—success depends on explainability, adaptability to workflows, and collaboration with clinicians. We are on the threshold of an era in which AI is becoming an integral element of precise, personalized neurology and neurosurgery.

https://doi.org/10.3390/a19010054

3.4. Hepatology

Similarly, in hepatology, AI is proving valuable for handling large datasets and complex imaging interpretations (Table 4). Using machine learning, convolutional neural networks, and deep learning, AI effectively analyzes clinical data, imaging, and histopathological studies, frequently matching or surpassing traditional techniques [71]. These capabilities translate into potential enhancements in diagnosis, prognosis, treatment optimization (Figure 4), and automated detection/classification of lesions on ultrasound, CT, or MRI [72].



Figure 4. Advantages and disadvantages of using AI in the diagnosis and treatment of liver diseases.

In liver transplantation, AI aids donor–recipient matching and risk assessment for rejection, potentially improving outcomes by enabling early complication detection from electronic records [73].

Table 4. Summary of the latest reports in the field of hepatology.

Authors	Methodology	Results	Significance
Ratziu, V., et al. [74]	Post hoc analysis of digitized liver biopsies from 251 patients with biopsy-confirmed NASH (F1–F3) enrolled in a 72-week phase 2 RCT of once-daily semaglutide (NCT02970942). Paired baseline and week-72 biopsies were scored by two expert pathologists using conventional NASH CRN criteria and independently analyzed by PathAI's machine-learning (ML) NASH model. Both categorical (ordinal) scores and continuous quantitative feature scores for fibrosis, steatosis, inflammation, and ballooning were generated. Treatment effects on the two co-primary endpoints (NASH resolution without worsening of fibrosis; fibrosis improvement ≥ 1 stage without worsening of NASH) were compared between pathologist and ML assessments.	Both pathologist and ML categorical scoring detected significantly higher rates of NASH resolution with semaglutide 0.4 mg versus placebo (pathologist: 58.5% vs. 22.0%, $p < 0.0001$; ML: 36.9% vs. 11.9%, $p = 0.0015$). Fibrosis improvement trended higher but was not significant with either method. ML-derived continuous scores revealed significant semaglutide-induced reductions in fibrosis ($p = 0.0099$) and other features that were not detectable by conventional categorical pathologist or ML ordinal assessments.	ML-based digital pathology reproduces expert pathologist categorical assessments of treatment response in NASH but provides superior sensitivity via continuous quantitative scoring, uncovering an antifibrotic effect of semaglutide missed by traditional histopathology. This demonstrates the value of AI-powered continuous metrics as more responsive endpoints in NASH clinical trials.
Tiyarattanachai, T., et al. [75]	Single-center prospective randomized controlled trial (TCTR20201230003); 504 patients (260 with FLLs for non-experts, 244 for experts) underwent real-time ultrasound twice: with and without AI assistance from a CNN-based system. Detection rates and false positives were compared using McNemar's test; non-experts were trainees; experts were board-certified radiologists.	For non-experts, AI assistance increased FLL detection rate (36.9% vs. 21.4%, $p < 0.001$) without significantly raising false positives (14.2% vs. 9.2%, $p = 0.08$). For experts, AI did not significantly improve detection (66.7% vs. 63.3%, $p = 0.32$) or alter false positives (8.6% vs. 9.0%, $p = 0.85$).	AI assistance significantly boosts focal liver lesion detection during ultrasound for non-expert operators, potentially enabling high-quality HCC surveillance in resource-limited settings with limited expert availability. No added benefit for experts suggests AI's primary value in democratizing access to reliable imaging.
Zhang, Y., et al. [76]	Prospective randomized diagnostic trial: 100 patients with suspected small hepatocellular carcinoma ($\leq$3 cm) were 1:1:1:1 randomized to four ultrasound modalities: color Doppler alone, contrast-enhanced ultrasound (CEUS) alone, elastography alone, or multimodal (all three combined). All images were processed and segmented by a Mask R-CNN deep-learning algorithm. Diagnostic performance was compared against pathological biopsy as gold standard. Additionally, EZH2 and p57 expression were measured by immunohistochemistry in tumor tissue, peritumoral tissue, and normal liver.	Mask R-CNN achieved the highest segmentation accuracy (97.23%) and average precision (71.90%). The multimodal ultrasound group outperformed single-modality groups with sensitivity 88.87%, specificity 90.91%, accuracy 89.47%, and Cohen's κ 0.68 (all $p < 0.05$). Multimodal imaging features of malignancy: irregular shape, unclear borders, uneven internal echo, grade 1–2 blood flow, elasticity score 4–5, and fast-in fast-out contrast pattern. EZH2 was overexpressed (75.95% positive) and p57 underexpressed (80.79% negative) in tumor tissue; p57 negativity was significantly higher in poorly differentiated HCC.	Mask R-CNN-enhanced multimodal ultrasound (Doppler + CEUS + elastography) provides excellent diagnostic performance for small HCC, significantly superior to any single modality. High EZH2 and low p57 expression are strongly associated with oncogenesis and poor differentiation in small HCC, supporting their potential as diagnostic and prognostic biomarkers. This AI-augmented multimodal approach offers a highly accurate, non-invasive tool for early HCC detection.
Briceño, J., et al. [77]	Multicenter retrospective analysis of 1003 liver transplants from 11 Spanish centers. Sixty-four donor and recipient variables were used to develop two complementary artificial neural network (ANN) models via Neural Net Evolutionary Programming (NNEP): a positive-survival model (NN-CCR) and a negative-loss model (NN-MS) to predict 3-month graft survival/loss for each donor–recipient pair. Performance was compared against six established prognostic scores (MELD, D-MELD, DRI, P-SOFT, SOFT, BAR) using ROC curves and AUROC.	ANN models significantly outperformed all conventional scores. NN-CCR predicted 3-month graft survival with 90.79% accuracy (AUROC 0.80) and NN-MS predicted graft loss with 71.42% accuracy (AUROC 0.82). AUROCs of traditional scores ranged from 0.41 to 0.67 (all $p < 0.001$ vs. ANN). ANN also outperformed multiple regression models.	Artificial neural networks provide substantially superior, individualized prediction of 3-month graft survival compared to all currently validated prognostic scores. This ANN-based approach offers a more objective, accurate, and equitable tool for donor–recipient matching and organ allocation, with potential to optimize justice, utility, and transplant outcomes in clinical practice.

https://doi.org/10.3390/a19010054

In recent years, artificial intelligence has been revolutionizing hepatology and liver transplantation, both in diagnosis and in assessing treatment response and optimizing organ allocation. The most groundbreaking are two independent post hoc analyses of the same phase 2 clinical trial with semaglutide in NASH [75]. In 251 patients with liver biopsies before and after 72 weeks of treatment, PathAI's machine-learning NASH model analyzed scanned slides in parallel with two expert pathologists. Traditional categorical ratings (CRN) confirmed a higher rate of NASH resolution without worsening fibrosis in the semaglutide 0.4 mg group (58.5% vs. 22.0% by pathologists; 36.9% vs. 11.9% by ML). However, only continuous, quantitative metrics generated by AI revealed a significant reduction in fibrosis ($p = 0.0099$) and improvement in all histological features (steatosis, inflammation, ballooning)—an antifibrotic effect of semaglutide that had completely eluded traditional categorical methods. This demonstrates that AI-assisted digital pathology not only replicates expert assessment but is significantly more sensitive and can detect subtle changes that determine drug success in NASH clinical trials. As a result, AI may accelerate drug development and change the way histological endpoints are defined in hepatology. Simultaneously, AI is beginning to significantly improve liver imaging diagnostics. In a randomized trial involving 504 patients, Tiyarattanachai et al. [75] demonstrated that assisting non-experts (residents) with a real-time CNN system during ultrasound examination increased the detection rate of focal liver lesions (FLL) from 21.4% to 36.9% without a significant increase in false positives. In expert (certified radiologists), AI did not provide additional benefit, meaning that its greatest value lies in the hands of less experienced operators. This paves the way for improved HCC screening in countries with limited access to specialists. Zhang et al. [76] went further and demonstrated that multimodal ultrasound (Doppler + CEUS + elastography) aided by Mask R-CNN segmentation achieved excellent diagnostic accuracy for small HCCs ($\leq$3 cm): sensitivity 88.9%, specificity 90.9%, and accuracy 89.5%. This result was significantly better than with either modality alone. Furthermore, EZH2 overexpression and p57 deficiency in tumor tissue were confirmed to be strongly associated with carcinogenicity and poorer prognosis—a finding that may support both diagnostics and risk stratification in the future. In transplantology, Briceño et al. [77] developed two complementary neural network models (NN-CCR and NN-MS) that, based on 64 donor and recipient variables, predict graft survival or loss within 3 months with an accuracy of 90.8% (AUROC 0.80) and 71.4% (AUROC 0.82). Both models significantly outperformed all previous classical scales (MELD, DRI, SOFT, BAR—AUROC 0.41–0.67). This is the first evidence that AI can offer individual, precise prediction of transplant prognosis, potentially enabling more equitable and efficient organ allocation. AI in hepatology is moving from supporting diagnostics to providing new, more sensitive endpoints in clinical trials, improving the detection of hepatocellular carcinoma in less experienced physicians, and finally to precise, individual donor–recipient matching in transplantology. We are witnessing a paradigm shift in which AI not only supports but is beginning to redefine standards for diagnosis, treatment, and organ allocation in liver diseases.

3.5. Pancreatic Disease

AI applications continue in the challenging domain of pancreatic diseases (Table 5), where early detection is critical. Ranging from pancreatic cancer to polyps and acute pancreatitis, AI primarily leverages imaging modalities such as endoscopic ultrasound (EUS), CT, and MRI to enable rapid identification of pathological changes [78,79]. Notably, 3D-CNN models detect pancreatic ductal adenocarcinoma (PDAC) and preinvasive lesions with sensitivity exceeding 90%, even on pre-diagnostic scans [80–82]. Deep learning achieves Dice coefficients above 0.92 for automatic segmentation of solid lesions and cysts. In assessing malignancy risk for intraductal papillary mucinous neoplasms (IPMN),

radiomics-based models yield AUC values of 0.88–0.95 [83,84]. Furthermore, tools like the EASY-APP model predict multiple organ failure risk within the first 24 h of hospitalization using clinical and biochemical data, with an accuracy of 85–92% [85].

AI is already achieving results that not only support physicians but, in many cases, equal or exceed their performance, while simultaneously paving the way for new generations of medical devices. The most spectacular results are seen in endoscopic diagnostics of the pancreas. Cui et al. [86], in a randomized, crossover multicenter study involving 12 endoscopists and 130 patients with solid pancreatic lesions, demonstrated that a multimodal AI model (combining EUS images with clinical data) achieved an external AUC of 0.955–0.976. In real-world conditions, AI assistance significantly improved the diagnostic accuracy of junior endoscopists ($p < 0.001$) without compromising the performance of experienced specialists. Importantly, explainability elements reduced skepticism among seniors and enabled seamless human-AI collaboration. This demonstrates that a well-designed, multimodal, and explainable AI system can become the standard in pancreatic endoscopy, especially in centers with varying levels of experience. Equally impressive results are achieved in the diagnosis of pancreatic cancer using computed tomography. Chen et al. [87] developed a fully automated deep-learning system that, in a nationwide validation using 1473 CT examinations in Taiwan, achieved a sensitivity of 89.7% and a specificity of 92.8% (AUC 0.95)—comparable to original radiologist reports. Crucially, the tool detected 74.7% of tumors smaller than 2 cm lesions that are often missed in daily practice. This result confirms that AI can be a reliable second reader or triage tool that significantly reduces the number of missed pancreatic cancers. In diabetes, AI is also revolutionizing automated glucose control systems. Kovatchev and colleagues [88], in their first human study, demonstrated that a neural network approximating a clinically validated model prediction algorithm (UMPC) named NAP achieved nearly identical glycemic control results (TIR 86% vs. 87%) with six times less computational demand. This breakthrough paves the way for replacing traditional control algorithms with artificial intelligence, which could significantly accelerate the development and implementation of future generations of artificial pancreas. The foundations for this path were previously demonstrated by Atlas et al. [89] in a feasibility study of the MD-Logic Artificial Pancreas system based on fuzzy logic. Under controlled, 24 h closed-loop conditions, a TIR of 73% was achieved without any episodes of symptomatic hypoglycemia, despite mealtime challenges. Although the results were even lower than those of current systems, this study demonstrated that a model-free, expert-based, closed-loop approach is safe and effective, opening the door to later, more advanced solutions. AI is evolving from a supportive role to a key diagnostic and therapeutic tool. Multimodal AI in EUS improves accuracy even for less experienced physicians, automated DL systems on CT save lives by detecting small pancreatic tumors, and neural approximations of traditional control algorithms enable the faster development of modern automated glucose control systems. We are witnessing a paradigm shift, with AI becoming an integral part of the precise diagnosis and treatment of pancreatic diseases and type 1 diabetes.

3.6. Other Applications of AI (Table 6)

Finally, AI demonstrates broad utility across additional medical domains. A growing body of evidence highlights its effectiveness in diagnostic imaging for cancers of the breast, colon, stomach, and melanoma, where it improves detection rates while reducing clinician workload [90,91]. Emerging roles in behavioral therapy and palliative care suggest potential outcome improvements, whereas results in medical training and patient self-care remain mixed. AI offers scalability but does not consistently outperform established standards [92–95]. Overall, AI appears safe and effective in addressing healthcare workforce shortages, though further real-world studies are essential to confirm long-term benefits.

https://doi.org/10.3390/a19010054

Table 5. Summary of the latest reports in the field of pancreatic diseases.

Authors	Methodology	Results	Significance
Cui, H., et al. [86]	Multicenter randomized crossover trial (4 centers in China, Jan–Jun 2023); 12 endoscopists (junior and senior) diagnosed 130 prospective patients with solid pancreatic lesions twice: conventionally and with assistance from a multimodal joint-AI model. The AI was trained/validated on EUS images + clinical data from 439 internal patients (2014–2022) and externally tested on 189 patients from 3 other institutions. Primary outcome: diagnostic accuracy with vs. without AI; secondary: AUC of the joint-AI model and human-AI interaction effects.	Joint-AI model achieved outstanding performance: internal AUC 0.996, external AUCs 0.955–0.976. In prospective crossover testing, AI assistance significantly improved the diagnostic accuracy of novice endoscopists ($p < 0.001$) with no decline in senior endoscopists. Explainability features reduced skepticism among experienced users, demonstrating positive human-AI collaboration.	The multimodal joint-AI model (EUS images + clinical data) dramatically outperforms single-modality AI and significantly boosts real-world diagnostic accuracy, especially for less-experienced endoscopists. It establishes a new benchmark for AI-assisted EUS diagnosis of solid pancreatic lesions and shows that well-designed, explainable multimodal AI can be readily adopted across expertise levels in clinical practice.
Chen, P. T., et al. [87]	Retrospective multicenter development and validation of an end-to-end deep learning (DL) system combining a segmentation CNN and a classifier ensemble of five CNNs. Training/validation used 546 pancreatic cancer CTs (2006–2018) and 733 normal pancreas CTs (2004–2019). Internal testing and comparison with original radiologist reports were performed, followed by nationwide external real-world validation on 1473 CT studies from institutions across Taiwan.	Internal test set: DL tool achieved 89.9% sensitivity (98/109) and 95.9% specificity (141/147), AUC 0.96, with no significant difference from original radiologist sensitivity (96.1%, $p = 0.11$). Nationwide real-world test set (n = 1473): sensitivity 89.7% (600/669), specificity 92.8% (746/804), AUC 0.95. For tumors < 2 cm, sensitivity was 74.7% (68/91).	This fully automated DL tool reliably detects pancreatic cancer on contrast-enhanced CT, including a large proportion of sub-2 cm tumors that are frequently missed by radiologists in routine practice. It offers performance comparable to or exceeding human interpretation and has proven generalizability across Taiwan, establishing it as a robust second-reader or triage tool to reduce missed pancreatic cancers in clinical workflow.
Kovatchev, B., et al. [88]	Pilot randomized crossover feasibility trial: 15 adults with T1D on commercial AID systems underwent two identical 20 h supervised hotel stays. The University of Virginia Model-Predictive Control (UMPC) algorithm was encoded into a neural network to create its Neural-Net Artificial Pancreas (NAP) approximation. Participants were randomly assigned to receive either NAP or the original UMPC algorithm during each session (crossover design with washout).	NAP and UMPC achieved nearly identical glycemic outcomes: TIR 86% vs. 87% (adjusted difference 1 percentage point), time < 70 mg/dL 2.0% vs. 1.8%, and CV 29.3% vs. 29.1%. Mean absolute difference in insulin delivery was only 0.031 U/h under identical inputs. NAP required sixfold lower computational resources than UMPC. No serious adverse events occurred with either controller.	A neural-network-encoded version of a clinically validated model-predictive control AID algorithm (NAP) replicated the performance of the original UMPC algorithm in real-world conditions while dramatically reducing computational burden. This first-in-human demonstration opens regulatory and clinical pathways for modern machine-learning techniques to replace traditional control algorithms in artificial pancreas systems, enabling faster innovation and deployment of next-generation AID.
Atlas, E., et al. [89]	Early feasibility clinical study of the MD-Logic Artificial Pancreas (MDLAP), a closed-loop system based on fuzzy logic theory that mimics diabetes caregiver reasoning using control-to-range and control-to-target strategies. Seven young adults with well-controlled T1D (age 19–30, A1C 6.6 $\pm$ 0.7%) underwent a total of 14 supervised sessions: 8 h sessions (overnight fasting + meal challenges) and 24 h full closed-loop sessions in a clinical research center.	Mean postprandial peak glucose was 224 $\pm$ 22 mg/dL, returning to <180 mg/dL within 2.6 $\pm$ 0.6 h and remaining stable in range for $\geq$1 h. During 24 h closed-loop control, 73% of sensor values were in 70–180 mg/dL, 27% > 180 mg/dL, and 0% < 70 mg/dL. No symptomatic hypoglycemia occurred in any trial.	The MD-Logic Artificial Pancreas, one of the earliest fuzzy-logic-based fully closed-loop systems, demonstrated safe and effective automated glucose control in a controlled setting, with good postprandial handling and zero hypoglycemia despite meal challenges. This proof-of-concept study established the clinical feasibility of non-model-based, expert-knowledge-driven artificial pancreas technology and paved the way for subsequent real-world and long-term trials.

https://doi.org/10.3390/a19010054

Table 6. Other examples of AI applications in medicine.

Authors	Methodology	Results	Significance
Lång, K., et al. [96]	Prospective, population-based, two-arm randomized controlled trial (MASAI, NCT04838756) at four Swedish screening sites. 80,033 women aged 40–80 were randomized 1:1 to AI-supported screening (Transpara v1.7.0 risk score 1–10 used to triage: scores 1–9 single reading, score 10 double reading; CAD marks shown for scores 8–10) versus standard double reading without AI. Participants and radiographers were masked; radiologists were not. Prespecified clinical safety analysis performed after 80,000 enrolments, focusing on cancer detection rate, recall rate, false-positive rate, PPV, and screen-reading workload.	AI-supported arm (n = 39,996): cancer detection 6.1/1000 (244 cancers), recall 2.2%, false-positive rate 1.5%, PPV 28.3%, with 44.3% reduction in screen readings (46,345 vs. 83,231). Standard double-reading arm (n = 40,024): cancer detection 5.1/1000 (203 cancers), recall 2.0%, false-positive rate 1.5%, PPV 24.8%. Cancer detection ratio 1.2 (95% CI 1.0–1.5, p = 0.052), exceeding the prespecified safety threshold (>3/1000 in AI arm). Proportion of invasive vs. in situ cancers was similar.	In the first large-scale randomized trial of AI in mammography screening, AI-supported reading proved clinically safe with a comparable (slightly higher) cancer detection rate, similar recall/false-positive rates, and a 44% reduction in radiologist workload. This establishes AI as a viable and efficient alternative to standard double reading, with potential to address global mammography workforce shortages while maintaining or improving screening performance.
Dembrower, K., et al. [97]	Prospective, population-based, paired-screening non-inferiority trial (ScreenTrustCAD, NCT04778670) at a single mammography unit in Stockholm. 55,581 consecutive screening women aged 40–74 years (Apr 2021–Jun 2022) had their mammograms independently read by two radiologists (standard double reading) and by AI (Transpara) + one radiologist, AI alone, and two radiologists + AI. Non-inferiority margin was a $\leq$15% relative reduction in cancer detection rate compared with standard double reading by two radiologists.	269 screen-detected cancers (0.49%). One radiologist + AI: 261 cancers (relative proportion 1.04, 95% CI 1.00–1.09)—non-inferior and numerically 4% higher. AI alone: 246 cancers (relative proportion 0.98, 95% CI 0.93–1.04)—non-inferior. Two radiologists + AI: 269 cancers (relative proportion 1.08, 95% CI 1.04–1.11)—superior. False-positive rates and recall rates were comparable across strategies.	In the first prospective trial of AI as an independent reader in population-based screening, replacing one of the two radiologists with AI maintained (and slightly improved) cancer detection while halving radiologist workload. AI alone was also non-inferior to standard double reading. These findings support safe, controlled implementation of AI to address radiologist shortages and increase screening capacity without compromising detection performance.
Sadeh-Sharvit, S., et al. [98]	Single-site randomized controlled trial at a U.S. community mental health clinic; 47 adults with primary depression or anxiety disorders starting outpatient individual CBT were 1:1 randomized to therapy augmented by the Eleos Health AI platform or treatment-as-usual (TAU) for the first 2 months. The AI platform automatically transcribed sessions, measured fidelity to evidence-based practices, integrated patient-reported outcome measures, and auto-drafted progress notes.	Patients in the AI-augmented group attended 67% more sessions (mean 5.24 vs. 3.14). Depression (PHQ-9) scores decreased 34% in the AI group vs. 20% in TAU, and anxiety (GAD-7) scores decreased 29% vs. 8%, both with large effect sizes favoring AI-augmented therapy. Therapists using the AI platform submitted progress notes 55 h earlier on average ($p < 0.001$). Treatment satisfaction and perceived helpfulness were equivalent between groups.	In the first randomized trial of an AI platform designed specifically for behavioral health, AI-augmented CBT significantly improved patient engagement, symptom reduction, and clinician documentation efficiency compared with standard care. These results establish clinical proof-of-concept that AI tools can meaningfully enhance the delivery and outcomes of routine outpatient psychotherapy in real-world community settings.
Repici, A., et al. [99]	Prospective, randomized, controlled non-inferiority trial (AID-2); 10 non-expert endoscopists (<2000 lifetime colonoscopies) performed 660 screening/surveillance/diagnostic colonoscopies in patients aged 40–80 years, randomized 1:1 to high-definition colonoscopy with or without real-time AI CADe (GI Genius, Medtronic). Primary endpoint: adenoma detection rate (ADR). Post hoc pooled analysis combined these data with the previously published AID-1 trial (6 expert endoscopists, similar design).	In non-experts (AID-2), CADe significantly increased ADR (53.3% vs. 44.5%; RR 1.22, 95% CI 1.04–1.40; $p < 0.01$ for non-inferiority, $p = 0.02$ for superiority), adenomas per colonoscopy, and detection of small/distal lesions, without increasing non-neoplastic resections. Pooled analysis of 1020 patients (AID-1 + AID-2) confirmed CADe as a strong independent predictor of higher ADR (RR 1.29, 95% CI 1.16–1.42), while endoscopist experience level was not significant (RR 1.02, 95% CI 0.89–1.16).	Real-time AI CADe substantially improves ADR in less experienced colonoscopists to levels exceeding those of unaided experts, and pooled data show that the benefit of CADe is largely independent of physician experience. This supports universal implementation of CADe to standardize and elevate colonoscopy quality across all skill levels, especially during training periods.

https://doi.org/10.3390/a19010054

Table 6. *Cont.*

Authors	Methodology	Results	Significance
Marcuzzi, A., et al. [100]	Single-center, three-arm randomized clinical trial at a Danish multidisciplinary specialist outpatient clinic. 294 adults (≥18 years) with persistent neck and/or low back pain on the waiting list for specialist care were randomized 1:1:1 to: SELFBACK app + usual care (AI-based, individually tailored weekly exercise, physical-activity, and education plans) e-Help + usual care (non-tailored generic web-based self-management information) Usual care alone Primary outcome: change in Musculoskeletal Health Questionnaire (MSK-HQ) score at 3 months. Follow-up: 6 weeks, 3 months, 6 months	At 3 months (82.7% complete data): Adjusted mean difference SELFBACK vs. usual care: +0.62 MSK-HQ points (95% CI −1.66 to 2.90; $p = 0.60$) SELFBACK vs. e-Help: +1.08 points (95% CI −1.24 to 3.41; $p = 0.36$) No significant differences were found in the primary outcome or any secondary outcomes (pain-related disability, pain intensity, catastrophizing, fear-avoidance, quality of life) at any time point.	In patients already referred to specialist care, adding an AI-based, individually tailored self-management app (SELFBACK) to usual care did not improve musculoskeletal health, pain, or function more than usual care alone or simple nontailored web-based information. This negative trial suggests that highly personalized digital self-management support may offer little additional benefit in a secondary/tertiary care setting where patients already expect specialist intervention, highlighting the importance of context when deploying digital health tools.
Wallace, M. B., et al. [101]	Multicenter (Italy, UK, US), randomized (1:1), tandem colonoscopy trial in 230 screening/surveillance patients. Same-day back-to-back procedures were performed with or without real-time AI assistance (deep-learning CADe), with order randomized (AI-first vs. colonoscopy-first). Primary endpoint: adenoma miss rate (AMR) = adenomas detected only on second colonoscopy/total adenomas from both colonoscopies. Secondary endpoints: mean adenomas at second colonoscopy, false-negative patient rate, and adverse events.	Overall AMR was significantly lower when AI was used first (15.5% vs. 32.4%; adjusted OR 0.38, 95% CI 0.23–0.62). AI benefit was greatest for ≤5 mm (15.9% vs. 35.8%; OR 0.34) and nonpolypoid lesions (16.8% vs. 45.8%; OR 0.24), and consistent in proximal (18.3% vs. 32.5%) and distal colon. Mean adenomas missed at second colonoscopy were halved with AI-first (0.33 vs. 0.70, $p < 0.001$). False-negative patient rate dropped from 29.6% to 6.8% (OR 0.17). No difference in adverse events.	In this rigorous tandem-colonoscopy design, real-time AI approximately halved the miss rate of colorectal adenomas, especially small/subtle lesions that commonly escape human perception. This provides direct evidence that AI reduces perceptual errors during standard colonoscopy, supporting its role in substantially improving colorectal cancer prevention through higher neoplasia detection.
Liaw, S. Y., et al. [102]	Single-center randomized controlled trial with 64 nursing students 1:1 allocated to sepsis team training in virtual reality simulation with either an AI-powered virtual doctor or a human-controlled virtual doctor (played by medical students). Both groups underwent identical sepsis scenarios. Outcomes assessed: sepsis care knowledge, interprofessional communication knowledge, self-efficacy in interprofessional communication (pre- and post-intervention), and objective performance in sepsis care and communication during a post-intervention simulation-based test.	Both groups significantly improved communication knowledge and self-efficacy from baseline. Only the AI-powered group significantly improved sepsis care knowledge ($p < 0.001$ vs. $p = 0.16$ in human-controlled group). Posttest sepsis knowledge: AI-powered group significantly higher (mean 9.06 vs. 7.75, $p = 0.009$) Sepsis care performance: no difference (13.63 vs. 12.75, $p = 0.39$) Interprofessional communication performance: no difference (29.34 vs. 27.06, $p = 0.21$) Self-efficacy in interprofessional communication: human-controlled group higher (69.6 vs. 60.1, $p = 0.008$).	An AI-powered virtual doctor was not inferior to a human-controlled (medical student) virtual doctor in training nursing students for sepsis care and interprofessional communication, demonstrating feasibility and scalability of AI-driven team training when medical student availability is limited. The AI group achieved superior sepsis knowledge retention, but human partners provided a greater boost in perceived communication self-efficacy. This supports hybrid models that combine AI scalability with human sociability for optimal interprofessional sepsis training.

https://doi.org/10.3390/a19010054

Table 6. *Cont.*

Authors	Methodology	Results	Significance
Luo, H., et al. [103]	Multicenter diagnostic study across six Chinese hospitals of varying tiers. A total of 1,036,496 standard white-light endoscopy images from 84,424 patients were used to develop and validate GRAIDS, a deep-learning convolutional neural network system for real-time detection of upper gastrointestinal cancers (oesophageal and gastric). Training and tuning: images from Sun Yat-sen University Cancer Center Validation: internal, prospective (same centre), and five external datasets from primary-care hospitals Performance of GRAIDS was directly compared with endoscopists of three expertise levels (expert, competent, trainee).	GRAIDS achieved: Diagnostic accuracy: 91.5–97.8% across all validation sets (internal 95.5%, prospective 92.7%) Sensitivity: 94.2% (not significantly different from expert endoscopists 94.5%, $p = 0.692$) Sensitivity markedly superior to competent (85.8%, $p < 0.0001$) and trainee (72.2%, $p < 0.0001$) endoscopists Specificity: 91–98% Positive predictive value: 81.4% Negative predictive value: 97.8%	GRAIDS is the first deep-learning system prospectively validated across multiple tiers of hospitals for real-time detection of upper gastrointestinal cancers using standard endoscopy. It performs at expert-endoscopist level for sensitivity and significantly outperforms non-expert endoscopists, offering a scalable, high-performance tool to improve early cancer detection in community and primary-care settings where endoscopic expertise is often limited. This represents a major step toward AI-assisted universal upper GI cancer screening
Wilson, P. M., et al. [104]	Pragmatic, cluster-randomized, stepped-wedge trial across 12 inpatient units at two U.S. hospitals (Aug 2019–Nov 2020). An AI/ML tool continuously screened the electronic health record and predicted need for palliative care consultation. Units were sequentially randomized to activate the intervention: when triggered, the tool sent a real-time Best Practice Advisory to the primary team recommending palliative care consultation. Control periods used usual care (no AI alert). Primary outcome: documented palliative care consultation. Secondary outcomes: hospital length of stay, ICU transfers, and 30-/60-/90-day readmissions.	3183 hospitalizations enrolled: 1717 retained for analysis (1212 intervention, 1332 usual-care). Palliative care consultation rate was significantly higher with AI support (IRR 1.44, 95% CI 1.11–1.92). Exploratory analyses showed reduced 60-day (OR 0.75, 95% CI 0.57–0.97) and 90-day readmissions (OR 0.72, 95% CI 0.55–0.93). No significant differences in length of stay or ICU transfers.	This is the first large-scale randomized trial demonstrating that a real-time AI/ML decision-support tool safely and effectively increases palliative care consultation rates in general hospital inpatients. The associated reduction in 60- and 90-day readmissions suggests that proactive, AI-triggered palliative care involvement improves downstream utilization and potentially quality of care. The pragmatic design and integration into routine workflow support scalability across diverse hospital settings.
Papachristou, P., et al. [105]	Prospective multicentre diagnostic trial at 36 primary care centres in Sweden. Primary care physicians used a smartphone-based AI clinical decision support tool (dermoscopic photo → dichotomous text output: "evidence of melanoma" or not) on 253 skin lesions of concern in 228 patients. All lesions were managed according to standard care (excision or specialist referral) regardless of app result. Final histopathological or specialist diagnoses were retrieved from medical records and compared with the AI output.	253 lesions: 21 melanomas (11 invasive, 10 in situ) Overall melanoma detection: AUROC 0.960 (95% CI 0.928–0.980) → max sensitivity 95.2%, specificity 84.5% Invasive melanomas only: AUROC 0.988 (95% CI 0.965–0.997) → max sensitivity 100%, specificity 92.6%	This is one of the first prospective trials of an AI-based melanoma detection tool used by primary care physicians in real-world clinical practice. The smartphone app achieved extremely high diagnostic accuracy (near perfect for invasive melanomas) when assessing dermoscopic images of lesions already flagged as suspicious. These results suggest that AI decision support can safely and effectively augment primary care physicians' ability to triage and detect melanoma, potentially reducing diagnostic delay and unnecessary excisions in primary care settings.

https://doi.org/10.3390/a19010054

In diagnostic imaging and clinical care, AI has been achieving breakthrough results in recent years, not only supporting physicians but, in many cases, maintaining or improving the quality of care while significantly reducing the workload of medical staff. The most important evidence comes from large, randomized trials of mammography screening. In the MASAI study [96]—the first multicenter RCT involving over 80,000 women—AI-assisted mammogram triage (Transpara) detected more cancers (6.1 vs. 5.1 per 1000) with similar false positive and recall rates, while simultaneously reducing the number of readings required by radiologists by 44%. Similar conclusions were drawn from the ScreenTrustCAD study [97], in which replacing one radiologist with an AI in a double reading did not reduce cancer detection rates, but slightly increased them, and the AI as a stand-alone reader was no worse than two human readers. These results clearly demonstrate that AI can safely and effectively address the shortage of radiologists in breast cancer screening programs, increasing the availability and quality of examinations. AI has demonstrated equally impressive results in colonoscopy. Repici et al. [99] in the AID-2 study demonstrated that the CADe system (GI Genius) significantly improved adenoma detection (ADR 53.3% vs. 44.5%) in inexperienced endoscopists, and a combined analysis with the AID-1 study showed that the benefit of AI was independent of operator experience. Wallace et al. [101] in a tandem colonoscopy study confirmed that AI halved the rate of missed adenomas (AMR 15.5% vs. 32.4%), particularly for small and flat lesions—critical for colorectal cancer prevention. Together, these studies demonstrate that real-time CADe is becoming a tool for standardizing and improving the quality of colonoscopy, regardless of the physician's experience. AI is also bringing tangible benefits in other areas. Sadeh-Sharvit et al. [98] in a randomized trial demonstrated that the Eleos Health platform (automatic transcription of CBT sessions, assessment of EBP fidelity, and note generation) increased patient attendance by 67%, improved symptoms of depression and anxiety, and sped up documentation by dozens of hours. Luo et al. [103] created the GRAIDS system, which, in prospective validation in hospitals of various levels, achieved expert-level sensitivity for detecting esophageal and gastric cancer (94.2%), significantly outperforming trainee and competent endoscopists. Papachristou et al. [105] demonstrated that a smartphone app in the hands of primary care physicians achieved an AUROC of 0.988 for invasive melanoma, opening the door to rapid triage of skin lesions in primary care. Not all trials yielded positive results. Marcuzzi et al. [100] demonstrated that the individually tailored SELFBACK app did not improve outcomes for patients with back and neck pain already referred to a specialist, illustrating that AI's success depends on the clinical context. Wilson et al. [104] demonstrated that proactive AI alerts increased palliative consultation rates by 44% and reduced readmissions at 60 and 90 days, confirming AI's potential to improve the quality of hospital care. Well-designed AI systems for mammography, colonoscopy, and upper gastrointestinal endoscopy already meet or exceed expert results, significantly reduce staff workload, and improve cancer detection (Table 7). In psychiatry and palliative care, AI increases patient engagement and treatment efficiency, and in medical education and primary care, it opens new opportunities for scalable support.

https://doi.org/10.3390/a19010054

Table 7. A summary of applications of artificial intelligence in various fields of medicine—an overview of current methods, goals and implementation status.

Medical Area	AI Methods	Purposes	Status
Radiology	Deep Learning (DL), Convolutional Neural Networks (CNN), Diffusion Models, Machine Learning (ML), Computer Vision, Large Multimodal Models (LMMs)	Image analysis, diagnosis (e.g., cancer detection in MRI/CT), report generation, anomaly detection, image enhancement, segmentation, triaging acute diseases (e.g., stroke, pneumothorax).	Established with emerging generative AI for synthetic imaging and bias mitigation [106–108].
Pathology	DL, CNN, ML, Neural Networks, Virtual Staining	Histopathology classification, mutation prediction, diagnosis (e.g., tumor subclassification), prognosis prediction, biomarker analysis.	Established, with emerging applications in precision pathology and immunotherapy outcome prediction [109–111].
Ophthalmology	DL, ML, CNN	Diabetic retinopathy screening, macular edema detection, glaucoma staging, refractive error detection.	Emerging, fastest-growing field with high publication growth [112–114].
Cardiology	ML, DL, LMMs, CNN	ECG interpretation, arrhythmia detection, cardiovascular risk prediction, echocardiography analysis.	Established, with emerging multimodal analysis for improved accuracy [115–117].
Oncology	DL, ML, CNN, Generative AI	Cancer detection/subtyping, treatment planning, prognosis, drug response prediction, precision therapy.	Established, emerging in genomic subtyping and AI for chemotherapy dose optimization [118–120].
Neurology	DL, LMMs, Transformer-based Models, ML	Lesion detection, brain condition distinction, stroke triage, neurosurgery outcome prediction.	Emerging, with multimodal integration for diagnostics [121–123].
Gastroenterology	CNN, Support Vector Machines (SVM), ML, DL	Neoplasia detection, polyp identification, inflammatory bowel disease prediction, endoscopy assistance.	Established, emerging integration with genomic data [124–126].
Dermatology	CNN, DL, ML	Skin lesion classification, melanoma detection, triage of lesions.	Established, addressing health disparities through data mining [127–129].
Mental Health/Psychiatry	NLP, LLMs, ML, Chatbots	Therapy delivery, psychosis prediction, mental health monitoring, patient engagement.	Emerging, with AI chatbots for accessible care and relapse prediction [130–132].
Drug Discovery & Development	ML, DL, NLP, Generative AI	Drug repurposing, target identification, toxicity prediction, vaccine design, clinical trial management.	Emerging, accelerating personalized medicine and synthetic data generation [133–135].
Orthopedics	ML, CNN, Diffusion Models	Fracture detection, osteoarthritis prediction, surgical outcome prediction, gait analysis.	Emerging, patient-specific biomechanical testing [136–138].
Emergency Medicine	ML, CNN, DL	Patient triage, risk stratification, hyperglycemic crises prediction, decision support.	Established, faster triage for acute conditions [139–141].
Preventive Medicine	ML, DL, Predictive Analytics	Risk assessment, disease progression prediction, population health management.	Emerging, second-fastest-growing field with wearable integration [142–144].
Anesthesiology	ML, DL	Depth of anesthesia monitoring, pain identification, intraoperative support.	Established, with emerging decision support techniques [145–147].

4. Discussion

Artificial intelligence offers significant potential to support diagnosis, prognosis, treatment planning, and workflow efficiency in medicine. Successful applications are characterized by high-quality training data, external validation, integration into clinical workflows,

https://doi.org/10.3390/a19010054

and collaboration with clinicians rather than replacement of human judgment. A credible validation of the artificial intelligence models is one of the core principles of the safe clinical use of the models. One such threat is that of overfitting, i.e., overfitting the model, hence high performance in the training group but poor performance in new patients. Data leakage may make this issue worse, as it involves accidental data leakage between the training and test sets that results in falsely high results. The learning curve analysis, in turn, is a complement to the effectiveness evaluation that shows whether the model in question becomes better in its performance as the volume of data increases. Calibration is also essential, and it identifies whether the projected probabilities are sufficient for the actual clinical risk. The single high values of discrimination measures like AUC cannot guarantee the reliability of the model in clinical practice. To ensure reliable evaluation, it is necessary to check whether the model identifies meaningful patterns in the data, rather than simply remembering cases used in the training set. This includes checking learning curve analysis, checking on separate cohorts and checking the calibration of the modeled risk. The representativeness of the training data also defines the effectiveness of the model because, in the case of the absence of diversity of the population, diagnostic errors may appear and, therefore, limited practical value can be obtained. All these factors need to be taken into consideration to evaluate the plausibility and viability of AI implementation in clinical care. In cardiology, pulmonology, neurology, hepatology, and gastroenterology, AI frequently achieves diagnostic accuracy comparable to or exceeding that of human experts in selected tasks. However, evidence for a reduction in hard clinical endpoints (mortality, major morbidity) remains limited. At the height of the artificial intelligence growth in the medical field, validation of the presented models is important. The study of the mentioned studies demonstrates a distinct separation as certain papers provide solid proof based on prospective, randomized clinical trials with external validation (e.g., TAILORED-AF [33], FLASH [34], MASAI [96], or Phase II studies of a drug identified with the help of generative AI [50,51]) which reduces the risk of the overfitting and states the real clinical effect. Conversely, many post hoc-based and retrospective cohort studies (e.g., radiomics studies in SCOT-HEART [37], predictive ML models in WARCEF [46], or subanalyses of RTOG 0617 [54]) report primarily final performance scores (e.g., C-statistic increase by 0.04 or AUC > 0.90) derived on in-house data or with limited cross-validation. When this happens, the likelihood of overfitting increases, especially when one applies high-dimensional features (e.g., radiomics) and thus erroneous hopes about their generalizability and actual clinical use. The difference highlights the necessity to focus on potential, multicenter, RCTs with solid endpoints prior to the regular application of AI in clinical practice. The absence of such validation of numerous potential good publications in the present weakens the evidence and demands careful interpretation of the reported advantages. Generalizability across diverse populations, bias mitigation, and legal responsibility require further attention. The problem of generalizability of AI models trained within a setting to local settings (so-called domain shift) due to differences in equipment, imaging protocols, patient demographics, and clinical practices is one issue that remains a challenge. Algorithms tested on data of highly developed centers tend to become slow in other environments, which restricts their safe use. The other important aspect is the detection of patient subcategories (e.g., age, gender, race/ethnicity or socioeconomic status) where the model does not perform optimally, resulting in care disparities. Subgroup analysis should also be performed regularly (e.g., with explainable AI, e.g., SHAP) or by identifying subgroups with worse performance using data, to identify such disparities. The absence of such evaluation in most studies (mostly retrospective or post hoc) is a severe drawback and explains the necessity to conduct prospective, multi-centric validation and extended observation of the results, post implementation. Although there is a rising trend in the

use of artificial intelligence in the field of medicine, the practical clinical use of artificial intelligence tools still has its pitfalls. The existing regulatory framework and recommendations are too broad and lack any reliable guidelines on the circumstances in which AI tool recommendations are not compatible with the opinion with an experienced clinician, which poses a severe problem associated with their application and operation in everyday clinical practice. In case the recommendations of AI tools challenge the judgment of a clinician, a physician may follow their judgment and experience, record discrepancies, and examine the potential error sources [148]. The implementation of these systems does not shift the burden of responsibility onto physicians in making their decisions and any mistakes may weigh down on the employees and the organizations employing these systems [149]. There are also challenges encountered in the course of testing algorithms in new, local clinical sites where there are differences between training data and the real population of patients. This will require the local validation and constant post-implementation observation. Validation of an algorithm can be used to determine the effectiveness of these algorithms using other centers or population data. Clinical implementation should be validated by a multicenter or prospective validation prior to clinical implementation [150,151]. The problem of the dissimilar efficiency of the models among particular patients is discussed in the research community, yet, in clinical practice, the number of studies to offer particular solutions to eliminate the inequalities is still low [152]. Tools such as AI-guided atrial fibrillation ablation [33], automated PCI optimization [34], and radiomics analysis of atherosclerotic plaques in CCTA [37] highlight the potential of AI to exceed the limitations of standard anatomical parameters. Low-threshold solutions, such as AI-ECG, which support the clinical practice of non-cardiologists in detecting AF and reduced ejection fraction [35,36], demonstrate that even the simplest algorithms can effectively reduce diagnostic gaps. Increasing patient experience with conversational AI [38] and automated echocardiographic measurements [39] emphasizes the importance of interpretability, personalization, and integration with workflows for ongoing clinical adaptation. The effectiveness of these tools depends primarily on appropriate data selection and clinical context. Models based solely on clinical variables can support patient selection for diagnosis [40], but imaging studies remain essential for identifying high-risk features. The greatest clinical benefits are observed in the use of tools that combine analytics with education, monitoring, and patient support in rehabilitation and outpatient care [41–43,45]. Fully automated tools such as CT-μFR [44] or stroke prediction models in patients with HFrEF [46] outperform traditional scales and reveal new risk determinants. In pulmonology, AI is useful at every stage of care, from drug discovery, through treatment optimization, to disease management. Generative AI in the antifibrotic drug discovery process [50,51] requires interpretation for long-term durability and safety. The use of predictive models in surgery and radiotherapy [52,54] also requires interpretability and clinical accountability. In chronic pulmonary diseases such as COPD, the benefits of AI only become apparent with long-term follow-up [53]. Hybrid tools combining imaging results and biomarkers, such as in the diagnosis of pulmonary nodules [56] or tuberculosis screening [57], achieve the greatest clinical value. In neurology, the value of AI depends not only on accuracy but also on the presentation of results and personalization to the patient. Explainable AI demonstrates that results can be interpreted differently by physicians and laypeople [63]. AI-guided neuromodulation enables objectification of DBS programming in treatment-resistant depression and Parkinson's disease [64,65], thus reducing subjective assessment. The main limitations to its effectiveness include negative experiences with electrodiagnostics [67] and the limited effectiveness of signal analysis from external devices [70], as well as the lack of integration into workflows. Tools supporting decisions and care organization, e.g., in the diagnosis of migraines [66], and hybrid neurosurgical training [68], reveal the greatest potential of AI in

https://doi.org/10.3390/a19010054

neurology. In the field of hepatology and transplantology, AI allows for overcoming the traditional limitations of expert assessments through continuous, quantitative measures and complex decision models. Digital analyses in NASH have shown that continuous scoring reveals subtle therapeutic effects that are undetectable by classical tools [74,75]. In diagnostic imaging, AI equalizes the competence of operators [76], and the integration of deep learning algorithms with a multimodal approach allows for the detection of small hepatocellular tumors with high accuracy [77]. Neural networks in transplantology outperform traditional prognostic scales [78], thus emphasizing the need for transparency and decision-making accountability. The primary value of AI comes from endpoint control, risk stratification, and clinical decision support, rather than from automated single assessments. In gastroenterology and diabetology, the greatest clinical benefits are provided by multimodal, validated, and supportive systems. Rather than replacing physician decision-making. Combining clinical data with EUS images significantly improves the diagnostic accuracy of less experienced endoscopists [86], and DL systems for pancreas imaging maintain their performance in nationwide validation [87]. Modern algorithms for type 1 diabetes significantly replicate the effectiveness of traditional control algorithms, reducing computational complexity and maintaining safety [88,89]. In diagnostic imaging, endoscopy, mental health, and palliative care, AI works best when applied in clinical processes and designed as part of the care team, rather than an automatic decision-maker. In mammography, AI can safely replace one of the operators, maintaining cancer detection rates and thus reducing staffing burden [96,97]. In endoscopy, advanced CADe and GRAIDS systems reduce perception errors and allow for compensating for differences in operator experience [99,101,103]. Algorithms do not always deliver benefits in populations already receiving care, as evidenced by the negative results of SELFBACK [100]. In mental health and palliative care, AI improves patient engagement, treatment outcomes, and supports clinical decision-making [98,102,104]. In everyday primary care practice, it allows for shortening the diagnostic process for melanoma and making the correct diagnosis [105]. The use of artificial intelligence in medicine shows significant potential to improve the diagnostic process, treatment, and efficiency of healthcare in many areas. Successful AI implementation requires not only precise algorithms but also sustainable generalizability, integration into daily workflows, and the definition of clinical responsibility.

5. Limitations

This narrative review has several important limitations. The rapid evolution of artificial intelligence in medicine means that some of the most recent studies, particularly those published after the final literature search, may not be included. The field is dominated by imaging-based applications, which are overrepresented in the current literature due to their relative maturity and regulatory approvals, while evidence in other domains (e.g., long-term outcome prediction, non-imaging biomarkers, or real-world implementation in low-resource settings) remains more limited. Many of the cited studies are single-center, have relatively short follow-up periods, or lack large-scale, multicenter prospective validation. Long-term clinical outcome data, particularly hard endpoints such as mortality or major morbidity, are still scarce for most AI tools. Additionally, potential publication bias toward positive results and the predominance of high-income country settings may limit generalizability. Finally, while this review focuses on clinically impactful, high-quality evidence, the heterogeneity of study designs and outcomes precludes formal meta-analysis, and important ongoing trials or emerging foundation models may alter the landscape soon. These limitations should be considered when interpreting the current state of AI in clinical medicine.

https://doi.org/10.3390/a19010054

6. Conclusions

AI, particularly ML and DL, shows growing integration into clinical practice, offering tools with high diagnostic and prognostic accuracy across multiple internal medicine specialties. This review summarizes current applications in cardiology, pulmonology, neurology, hepatology, pancreatic diseases, gastroenterology, and related fields. Promising evidence from recent studies demonstrates improvements in diagnostic performance as well as workflow efficiency. As of late 2025, over 1300 AI-enabled medical devices have received FDA clearance, predominantly in imaging-related applications (more than 75%), with cardiology and radiology leading in regulatory approvals. Despite strong performance metrics in controlled settings (AUC often 0.90–0.98) and increasing real-world implementations, most of the evidence derives from retrospective studies or external validations, with a limited number of randomized controlled trials. Direct improvements in hard clinical outcomes remain preliminary and require further confirmation. Widespread clinical adoption is currently constrained by challenges, including the need for robust prospective validation, generalizability across diverse populations and settings, bias mitigation, interpretability, ethical considerations, legal liability, and clinician acceptance. Large-scale, multicenter, prospective randomized controlled trials are essential to establish sustained benefits on patient outcomes, optimize integration into clinical workflows, and realize the full potential of AI as a supportive tool that enhances rather than replaces human clinical judgment. Successful future implementation will depend on high-quality data, transparent algorithms, multidisciplinary collaboration, and ongoing post-market surveillance.

Author Contributions: Conceptualization, A.Ł., D.B.-A., W.M. and D.A.; methodology, A.Ł., D.B.-A., W.M. and D.A.; software, A.Ł., D.B.-A., W.M. and D.A.; validation, A.Ł., D.B.-A., W.M. and D.A.; formal analysis, A.Ł., D.B.-A., W.M. and D.A.; resources, A.Ł., D.B.-A., W.M. and D.A.; data curation, A.Ł., D.B.-A., W.M. and D.A.; writing—original draft preparation, A.Ł., D.B.-A., W.M. and D.A.; writing—review and editing, A.Ł., D.B.-A., W.M. and D.A.; visualization, A.Ł., D.B.-A., W.M. and D.A.; supervision, D.A.; project administration, A.Ł., D.B.-A., W.M. and D.A.; funding acquisition, D.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All data has been included.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AF	Atrial fibrillation
AI	Artificial intelligence
AKI	Acute kidney injury
AUC	Area under the (receiver operating characteristic) curve
CADe	Computer-aided detection
CAD	Coronary artery disease
CNN	Convolutional neural network
CT	Computed tomography

DL	Deep learning
ECG	Electrocardiogram
EHR	Electronic health record
EUS	Endoscopic ultrasound
FDA	U.S. Food and Drug Administration
HF	Heart failure
HFpEF	Heart failure with preserved ejection fraction
HPI	Hypotension Prediction Index
IPMN	Intraductal papillary mucinous neoplasm
LVEF	Left ventricular ejection fraction
MACCE	Major adverse cardiac and cerebrovascular events
ML	Machine learning
MRI	Magnetic resonance imaging
NOAC	Non-vitamin K antagonist oral anticoagulant
NSCLC	Non-small cell lung cancer
PCI	Percutaneous coronary intervention
PDAC	Pancreatic ductal adenocarcinoma
PPV	Positive predictive value
QoL	Quality of life
RCT	Randomized controlled trial
T1D	Type 1 diabetes

References

1. Bajwa, J.; Munir, U.; Nori, A.; Williams, B. Artificial intelligence in healthcare: Transforming the practice of medicine. *Future Healthc. J.* **2021**, *8*, e188–e194. [CrossRef]
2. Shi, J.; Bendig, D.; Vollmar, H.C.; Rasche, P. Mapping the Bibliometrics Landscape of AI in Medicine: Methodological Study. *J. Med. Internet Res.* **2023**, *25*, e45815. [CrossRef] [PubMed]
3. Howard, F.M.; Li, A.; Riffon, M.F.; Garrett-Mayer, E.; Pearson, A.T. Characterizing the Increase in Artificial Intelligence Content Detection in Oncology Scientific Abstracts From 2021 to 2023. *JCO Clin. Cancer Inform.* **2024**, *8*, e2400077. [CrossRef] [PubMed]
4. Shajari, S.; Kuruvinashetti, K.; Komeili, A.; Sundararaj, U. The Emergence of AI-Based Wearable Sensors for Digital Health Technology: A Review. *Sensors* **2023**, *23*, 9498. [CrossRef] [PubMed]
5. Huang, G.; Chen, X.; Liao, C. AI-Driven Wearable Bioelectronics in Digital Healthcare. *Biosensors* **2025**, *15*, 410. [CrossRef] [PubMed]
6. Hannun, A.Y.; Rajpurkar, P.; Haghpanahi, M.; Tison, G.H.; Bourn, C.; Turakhia, M.P.; Ng, A.Y. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nat. Med.* **2019**, *25*, 65–69. [CrossRef] [PubMed]
7. Virella Pérez, Y.I.; Medlow, S.; Ho, J.; Steinbeck, K. Mobile and Web-Based Apps That Support Self-Management and Transition in Young People with Chronic Illness: Systematic Review. *J. Med. Internet Res.* **2019**, *21*, e13579. [CrossRef] [PubMed]
8. Rieke, N.; Hancox, J.; Li, W.; Milletarì, F.; Roth, H.R.; Albarqouni, S.; Bakas, S.; Galtier, M.N.; Landman, B.A.; Maier-Hein, K.; et al. The future of digital health with federated learning. *npj Digit. Med.* **2020**, *3*, 119. [CrossRef] [PubMed]
9. Ziller, A.; Usynin, D.; Braren, R.; Makowski, M.; Rueckert, D.; Kaissis, G. Medical imaging deep learning with differential privacy. *Sci. Rep.* **2021**, *11*, 13524. [CrossRef] [PubMed]
10. Yadalam, A.K.; Liu, C.; Hui, Q.; Razavi, A.C.; Sperling, L.S.; Quyyumi, A.A.; Sun, Y.V. Large-Scale Proteomics-Based Risk Score for the Prediction of Incident Cardio-Kidney-Metabolic Disease Risk. *Circ. Genom. Precis. Med.* **2025**, *18*, e005125. [CrossRef]
11. Singh, M.; Kumar, A.; Khanna, N.N.; Laird, J.R.; Nicolaides, A.; Faa, G.; Johri, A.M.; Mantella, L.E.; Fernandes, J.F.E.; Teji, J.S.; et al. Artificial intelligence for cardiovascular disease risk assessment in personalised framework: A scoping review. *EClinicalMedicine* **2024**, *73*, 102660. [CrossRef]
12. Hadida Barzilai, D.; Sudri, K.; Goshen, G.; Klang, E.; Zimlichman, E.; Barbash, I.; Cohen Shelly, M. Randomized Controlled Trials Evaluating Artificial Intelligence in Cardiovascular Care: A Systematic Review. *JACC Adv.* **2025**, *4*, 102152. [CrossRef]
13. Komorowski, M.; Cecconi, M. Deploying AI in the ICU: Learning from successes and failures. *Intensive Care Med.* **2025**, *51*, 2410–2413. [CrossRef] [PubMed]
14. Tomašev, N.; Glorot, X.; Rae, J.W.; Zielinski, M.; Askham, H.; Saraiva, A.; Mottram, A.; Meyer, C.; Ravuri, S.; Protsyuk, I.; et al. A clinically applicable approach to continuous prediction of future acute kidney injury. *Nature* **2019**, *572*, 116–119. [CrossRef] [PubMed]

15. Kellum, J.A.; Bihorac, A. Artificial intelligence to predict AKI: Is it a breakthrough? *Nat. Rev. Nephrol.* **2019**, *15*, 663–664. [CrossRef]
16. Tiwari, A.; Mishra, S.; Kuo, T.R. Current AI technologies in cancer diagnostics and treatment. *Mol. Cancer* **2025**, *24*, 159. [CrossRef] [PubMed]
17. Zhang, J.; Che, Y.; Liu, R.; Wang, Z.; Liu, W. Deep learning-driven multi-omics analysis: Enhancing cancer diagnostics and therapeutics. *Brief. Bioinform.* **2025**, *26*, bbaf440. [CrossRef]
18. Olang, O.; Mohseni, S.; Shahabinezhad, A.; Hamidianshirazi, Y.; Goli, A.; Abolghasemian, M.; Shafiee, M.A.; Aarabi, M.; Alavinia, M.; Shaker, P. Artificial Intelligence-Based Models for Prediction of Mortality in ICU Patients: A Scoping Review. *J. Intensive Care Med.* **2025**, *40*, 1240–1246. [CrossRef]
19. Singhal, K.; Tu, T.; Gottweis, J.; Sayres, R.; Wulczyn, E.; Amin, M.; Hou, L.; Clark, K.; Pfohl, S.R.; Cole-Lewis, H.; et al. Toward expert-level medical question answering with large language models. *Nat. Med.* **2025**, *31*, 943–950. [CrossRef]
20. Yang, X.; Chen, A.; PourNejatian, N.; Shin, H.C.; Smith, K.E.; Parisien, C.; Compas, C.; Martin, C.; Costa, A.B.; Flores, M.G.; et al. A large language model for electronic health records. *npj Digit. Med.* **2022**, *5*, 194. [CrossRef]
21. Sivakumar, R.; Lue, B.; Kundu, S. FDA Approval of Artificial Intelligence and Machine Learning Devices in Radiology: A Systematic Review. *JAMA Netw. Open* **2025**, *8*, e2542338. [CrossRef] [PubMed]
22. Shuja, M.H.; Shakil, F.; Shuja, S.H.; Hasan, M.; Edhi, M.; Abbasi, A.F.; Jawaid, A.; Shakil, S. Harnessing Artificial Intelligence in Cardiology: Advancements in Diagnosis, Treatment, and Patient Care. *Heart Views* **2024**, *25*, 241–248. [CrossRef]
23. Ahmad, A.; Ahmad, S.; Ahmad, R.; Bodi, J.; Mohamed, A.; Wasim, A. Artificial Intelligence in Cardiovascular Diagnosis: Innovations and Impact on Disease Screenings. *J. Pharm. Bioallied Sci.* **2025**, *17*, S1900–S1903. [CrossRef]
24. Medhi, D.; Kamidi, S.R.; Mamatha Sree, K.P.; Shaikh, S.; Rasheed, S.; Thengu Murichathil, A.H.; Nazir, Z. Artificial Intelligence and Its Role in Diagnosing Heart Failure: A Narrative Review. *Cureus* **2024**, *16*, e59661. [CrossRef] [PubMed]
25. Udoy, I.A.; Hassan, O. AI-Driven Technology in Heart Failure Detection and Diagnosis: A Review of the Advancement in Personalized Healthcare. *Symmetry* **2025**, *17*, 469. [CrossRef]
26. Almansouri, N.E.; Awe, M.; Rajavelu, S.; Jahnavi, K.; Shastry, R.; Hasan, A.; Hasan, H.; Lakkimsetti, M.; AlAbbasi, R.K.; Gutiérrez, B.C.; et al. Early Diagnosis of Cardiovascular Diseases in the Era of Artificial Intelligence: An In-Depth Review. *Cureus* **2024**, *16*, e55869. [CrossRef] [PubMed]
27. Yasmin, F.; Shah, S.M.I.; Naeem, A.; Shujauddin, S.M.; Jabeen, A.; Kazmi, S.; Siddiqui, S.A.; Kumar, P.; Salman, S.; Hassan, S.A.; et al. Artificial intelligence in the diagnosis and detection of heart failure: The past, present, and future. *Rev. Cardiovasc. Med.* **2021**, *22*, 1095–1113. [CrossRef]
28. Sanders-van Wijk, S.; Barandiarán Aizpurua, A.; Brunner-La Rocca, H.P.; Henkens, M.T.H.M.; Weerts, J.; Knackstedt, C.; Uszko-Lencer, N.; Heymans, S.; van Empel, V. The HFA-PEFF and H_2 FPEF scores largely disagree in classifying patients with suspected heart failure with preserved ejection fraction. *Eur. J. Heart Fail.* **2021**, *23*, 838–840. [CrossRef]
29. Bourazana, A.; Xanthopoulos, A.; Briasoulis, A.; Magouliotis, D.; Spiliopoulos, K.; Athanasiou, T.; Vassilopoulos, G.; Skoularigis, J.; Triposkiadis, F. Artificial Intelligence in Heart Failure: Friend or Foe? *Life* **2024**, *14*, 145. [CrossRef] [PubMed]
30. Holt, D.B.; El-Bokl, A.; Stromberg, D.; Taylor, M.D. Role of Artificial Intelligence in Congenital Heart Disease and Interventions. *J. Soc. Cardiovasc. Angiogr. Interv.* **2025**, *4*, 102567. [CrossRef]
31. Paul, D.; Sanap, G.; Shenoy, S.; Kalyane, D.; Kalia, K.; Tekade, R.K. Artificial intelligence in drug discovery and development. *Drug Discov. Today* **2021**, *26*, 80–93. [CrossRef]
32. Meder, B.; Asselbergs, F.W.; Ashley, E. Artificial intelligence to improve cardiovascular population health. *Eur. Heart J.* **2025**, *46*, 1907–1916. [CrossRef]
33. Deisenhofer, I.; Albenque, J.P.; Busch, S.; Gitenay, E.; Mountantonakis, S.E.; Roux, A.; Horvilleur, J.; Bakouboula, B.; Oza, S.; Abbey, S.; et al. Artificial intelligence for individualized treatment of persistent atrial fibrillation: A randomized controlled trial. *Nat. Med.* **2025**, *31*, 1286–1293. [CrossRef]
34. Kim, Y.; Yoon, H.J.; Suh, J.; Kang, S.H.; Lim, Y.H.; Jang, D.H.; Park, J.H.; Shin, E.S.; Bae, J.W.; Lee, J.H.; et al. Artificial Intelligence-Based Fully Automated Quantitative Coronary Angiography vs Optical Coherence Tomography-Guided PCI: The FLASH Trial. *JACC. Cardiovasc. Interv.* **2025**, *18*, 187–197. [CrossRef]
35. Liu, W.T.; Lin, C.; Lee, C.C.; Chang, C.H.; Fang, W.H.; Tsai, D.J.; Lin, W.Y.; Hung, Y.; Chen, K.C.; Lee, C.H.; et al. Artificial Intelligence-Enabled ECGs for Atrial Fibrillation Identification and Enhanced Oral Anticoagulant Adoption: A Pragmatic Randomized Clinical Trial. *J. Am. Heart Assoc.* **2025**, *14*, e042106. [CrossRef] [PubMed]
36. Tsai, D.J.; Lin, C.; Liu, W.T.; Lee, C.C.; Chang, C.H.; Lin, W.Y.; Liu, Y.L.; Chang, D.W.; Hsieh, P.H.; Tsai, C.S.; et al. Artificial intelligence-assisted diagnosis and prognostication in low ejection fraction using electrocardiograms in inpatient department: A pragmatic randomized controlled trial. *BMC Med.* **2025**, *23*, 342. [CrossRef] [PubMed]
37. Kolossváry, M.; Lin, A.; Kwiecinski, J.; Cadet, S.; Slomka, P.J.; Newby, D.E.; Dweck, M.R.; Williams, M.C.; Dey, D. Coronary Plaque Radiomic Phenotypes Predict Fatal or Nonfatal Myocardial Infarction: Analysis of the SCOT-HEART Trial. *JACC Cardiovasc. Imaging* **2025**, *18*, 308–319. [CrossRef]

38. Trivedi, R.; Shaw, T.; Sheahen, B.; Chow, C.K.; Laranjo, L. Patient Perspectives on Conversational Artificial Intelligence for Atrial Fibrillation Self-Management: Qualitative Analysis. *J. Med. Internet Res.* **2025**, *27*, e64325. [CrossRef] [PubMed]
39. Mekonnen, D.; Spitzer, E.; McFadden, E.P.; Caplice, N.M.; Ren, C.B. Artificial intelligence-assisted left ventricular global longitudinal strain assessment in patients with acute myocardial infarction: A RESUS-AMI trial sub-analysis. *Int. J. Cardiovasc. Imaging* **2025**, *41*, 1225–1236. [CrossRef]
40. Williams, M.C.; Guimaraes, A.R.M.; Jiang, M.; Kwieciński, J.; Weir-McCall, J.R.; Adamson, P.D.; Mills, N.L.; Roditi, G.H.; van Beek, E.J.R.; Nicol, E.; et al. Machine learning to predict high-risk coronary artery disease on CT in the SCOT-HEART trial. *Open Heart* **2025**, *12*, e003162. [CrossRef]
41. Saklica, D.; Vardar-Yagli, N.; Saglam, M.; Yuce, D.; Ates, A.H.; Yorgun, H. The Impact of Technology-Based Cardiac Rehabilitation on Exercise Capacity and Adherence in Patients with Coronary Artery Disease: An Artificial Intelligence Analysis. *Arq. Bras. Cardiol.* **2025**, *122*, e20240765. [CrossRef]
42. Trivedi, R.; Laranjo, L.; Marschner, S.; Thiagalingam, A.; Thomas, S.; Kumar, S.; Shaw, T.; Chow, C.K. Conversational AI Phone Calls to Support Patients with Atrial Fibrillation: Randomized Controlled Trial. *JMIR Cardio* **2025**, *9*, e64326. [CrossRef]
43. Fiolet, A.T.L.; Lin, A.; Kwiecinski, J.; Tutein Nolthenius, J.; McElhinney, P.; Grodecki, K.; Kietselaer, B.; Opstal, T.S.; Cornel, J.H.; Knol, R.J.; et al. Effect of low-dose colchicine on pericoronary inflammation and coronary plaque composition in chronic coronary disease: A subanalysis of the LoDoCo2 trial. *Heart* **2025**, *111*, 1156–1163. [CrossRef] [PubMed]
44. Li, G.; Weng, T.; Sun, P.; Li, Z.; Ding, D.; Guan, S.; Han, W.; Gan, Q.; Li, M.; Qi, L.; et al. Diagnostic performance of fully automatic coronary CT angiography-based quantitative flow ratio. *J. Cardiovasc. Comput. Tomogr.* **2025**, *19*, 40–47. [CrossRef]
45. Yu, X.; Cao, J.; Xu, J.; Xu, Q.; Chen, H.; Yu, D.; Ou, A.; Hu, Y.; Ma, L. Efficacy of Telemedical Interventional Management in Patients with Coronary Heart Disease Undergoing Percutaneous Coronary Intervention: Randomized Controlled Trial. *J. Med. Internet Res.* **2025**, *27*, e63350. [CrossRef] [PubMed]
46. Ishiguchi, H.; Chen, Y.; Huang, B.; Gue, Y.; Correa, E.; Homma, S.; Thompson, J.L.P.; Qian, M.; Lip, G.Y.H.; Abdul-Rahim, A.H. Machine learning for stroke in heart failure with reduced ejection fraction but without atrial fibrillation: A post-hoc analysis of the WARCEF trial. *Eur. J. Clin. Investig.* **2025**, *55*, e14360. [CrossRef] [PubMed]
47. Sindhu, A.; Jadhav, U.; Ghewade, B.; Bhanushali, J.; Yadav, P. Revolutionizing Pulmonary Diagnostics: A Narrative Review of Artificial Intelligence Applications in Lung Imaging. *Cureus* **2024**, *16*, e57657. [CrossRef]
48. Cellina, M.; Cacioppa, L.M.; Cè, M.; Chiarpenello, V.; Costa, M.; Vincenzo, Z.; Pais, D.; Bausano, M.V.; Rossini, N.; Bruno, A.; et al. Artificial Intelligence in Lung Cancer Screening: The Future Is Now. *Cancers* **2023**, *15*, 4344. [CrossRef]
49. Ma, K.; Zheng, M.; Chen, W.; Qi, Y.; Rong, H. Research progress in computer-aided diagnosis systems for lung cancer. *npj Digit. Med.* **2025**, *8*, 722. [CrossRef]
50. Ren, F.; Aliper, A.; Chen, J.; Zhao, H.; Rao, S.; Kuppe, C.; Ozerov, I.V.; Zhang, M.; Witte, K.; Kruse, C.; et al. A small-molecule TNIK inhibitor targets fibrosis in preclinical and clinical models. *Nat. Biotechnol.* **2025**, *43*, 63–75. [CrossRef]
51. Xu, Z.; Ren, F.; Wang, P.; Cao, J.; Tan, C.; Ma, D.; Zhao, L.; Dai, J.; Ding, Y.; Fang, H.; et al. A generative AI-discovered TNIK inhibitor for idiopathic pulmonary fibrosis: A randomized phase 2a trial. *Nat. Med.* **2025**, *31*, 2602–2610. [CrossRef]
52. Habicher, M.; Denn, S.M.; Schneck, E.; Akbari, A.A.; Schmidt, G.; Markmann, M.; Alkoudmani, I.; Koch, C.; Sander, M. Perioperative goal-directed therapy with artificial intelligence to reduce the incidence of intraoperative hypotension and renal failure in patients undergoing lung surgery: A pilot study. *J. Clin. Anesth.* **2025**, *102*, 111777. [CrossRef]
53. Hong, L.; Cheng, X.; Zheng, D. Application of Artificial Intelligence in Emergency Nursing of Patients with Chronic Obstructive Pulmonary Disease. *Contrast Media Mol. Imaging* **2021**, *2021*, 6423398. [CrossRef] [PubMed]
54. Ladbury, C.; Li, R.; Danesharasteh, A.; Ertem, Z.; Tam, A.; Liu, J.; Hao, C.; Li, R.; McGee, H.; Sampath, S.; et al. Explainable Artificial Intelligence to Identify Dosimetric Predictors of Toxicity in Patients with Locally Advanced Non-Small Cell Lung Cancer: A Secondary Analysis of RTOG 0617. *Int. J. Radiat. Oncol. Biol. Phys.* **2023**, *117*, 1287–1296. [CrossRef] [PubMed]
55. Pasipanodya, J.G.; Smythe, W.; Merle, C.S.; Olliaro, P.L.; Deshpande, D.; Magombedze, G.; McIlleron, H.; Gumbo, T. Artificial intelligence-derived 3-Way Concentration-dependent Antagonism of Gatifloxacin, Pyrazinamide, and Rifampicin During Treatment of Pulmonary Tuberculosis. *Clin. Infect. Dis.* **2018**, *67*, S284–S292. [CrossRef]
56. Ding, Y.; Zhang, J.; Zhuang, W.; Gao, Z.; Kuang, K.; Tian, D.; Deng, C.; Wu, H.; Chen, R.; Lu, G.; et al. Improving the efficiency of identifying malignant pulmonary nodules before surgery via a combination of artificial intelligence CT image recognition and serum autoantibodies. *Eur. Radiol.* **2023**, *33*, 3092–3102. [CrossRef]
57. Bosman, S.; Ayakaka, I.; Muhairwe, J.; Kamele, M.; van Heerden, A.; Madonsela, T.; Labhardt, N.D.; Sommer, G.; Bremerich, J.; Zoller, T.; et al. Evaluation of C-Reactive Protein and Computer-Aided Analysis of Chest X-rays as Tuberculosis Triage Tests at Health Facilities in Lesotho and South Africa. *Clin. Infect. Dis.* **2024**, *79*, 1293–1302. [CrossRef] [PubMed]
58. Kalani, M.; Anjankar, A. Revolutionizing Neurology: The Role of Artificial Intelligence in Advancing Diagnosis and Treatment. *Cureus* **2024**, *16*, e61706. [CrossRef]

59. Kale, M.; Wankhede, N.; Pawar, R.; Ballal, S.; Kumawat, R.; Goswami, M.; Khalid, M.; Taksande, B.; Upaganlawar, A.; Umekar, M.; et al. AI-driven innovations in Alzheimer's disease: Integrating early diagnosis, personalized treatment, and prognostic modelling. *Ageing Res. Rev.* **2024**, *101*, 102497. [CrossRef]
60. Onciul, R.; Tataru, C.-I.; Dumitru, A.V.; Crivoi, C.; Serban, M.; Covache-Busuioc, R.-A.; Radoi, M.P.; Toader, C. Artificial Intelligence and Neuroscience: Transformative Synergies in Brain Research and Clinical Applications. *J. Clin. Med.* **2025**, *14*, 550. [CrossRef]
61. Zhang, H.; Jiao, L.; Yang, S.; Li, H.; Jiang, X.; Feng, J.; Zou, S.; Xu, Q.; Gu, J.; Wang, X.; et al. Brain-computer interfaces: The innovative key to unlocking neurological conditions. *Int. J. Surg.* **2024**, *110*, 5745–5762. [CrossRef]
62. Buttar, A.M.; Shaheen, Z.; Gumaei, A.H.; Mosleh, M.A.A.; Gupta, I.; Alzanin, S.M.; Akbar, M.A. Enhanced neurological anomaly detection in MRI images using deep convolutional neural networks. *Front. Med.* **2024**, *11*, 1504545. [CrossRef] [PubMed]
63. Gombolay, G.Y.; Silva, A.; Schrum, M.; Gopalan, N.; Hallman-Cooper, J.; Dutt, M.; Gombolay, M. Effects of explainable artificial intelligence in neurology decision support. *Ann. Clin. Transl. Neurol.* **2024**, *11*, 1224–1235. [CrossRef] [PubMed]
64. Alagapan, S.; Choi, K.S.; Heisig, S.; Riva-Posse, P.; Crowell, A.; Tiruvadi, V.; Obatusin, M.; Veerakumar, A.; Waters, A.C.; Gross, R.E.; et al. Cingulate dynamics track depression recovery with deep brain stimulation. *Nature* **2023**, *622*, 130–138. [CrossRef]
65. Boutet, A.; Madhavan, R.; Elias, G.J.B.; Joel, S.E.; Gramer, R.; Ranjan, M.; Paramanandam, V.; Xu, D.; Germann, J.; Loh, A.; et al. Predicting optimal deep brain stimulation parameters for Parkinson's disease using functional MRI and machine learning. *Nat. Commun.* **2021**, *12*, 3043. [CrossRef]
66. Cowan, R.P.; Rapoport, A.M.; Blythe, J.; Rothrock, J.; Knievel, K.; Peretz, A.M.; Ekpo, E.; Sanjanwala, B.M.; Woldeamanuel, Y.W. Diagnostic accuracy of an artificial intelligence online engine in migraine: A multi-center study. *Headache* **2022**, *62*, 870–882. [CrossRef] [PubMed]
67. Gorenshtein, A.; Weisblat, Y.; Khateb, M.; Kenan, G.; Tsirkin, I.; Fayn, G.; Geller, S.; Shelly, S. AI-Based EMG Reporting: A Randomized Controlled Trial. *J. Neurol.* **2025**, *272*, 586. [CrossRef]
68. Davidovic, V.; Giglio, B.; Albeloushi, A.; Alhaj, A.K.; Alhantoobi, M.; Saeedi, R.; Deraiche, S.; Yilmaz, R.; Tee, T.; Fazlollahi, A.M.; et al. Effect of Artificial Intelligence-Augmented Human Instruction on Feedback Frequency and Surgical Performance During Simulation Training. *J. Surg. Educ.* **2025**, *82*, 103743. [CrossRef]
69. Hassan, A.E.; Ravi, S.; Desai, S.; Saei, H.M.; Mckennon, E.; Tekle, W.G. An artificial intelligence (AI)-based approach to clinical trial recruitment: The impact of Viz RECRUIT on enrollment in the EMBOLISE trial. *Interv. Neuroradiol. J. Perither. Neuroradiol. Surg. Proced. Relat. Neurosci.* **2025**, *31*, 739–744. [CrossRef]
70. Macea, J.; Heremans, E.R.M.; Proost, R.; De Vos, M.; Van Paesschen, W. Automated Sleep Staging in Epilepsy Using Deep Learning on Standard Electroencephalogram and Wearable Data. *J. Sleep Res.* **2025**, *34*, e70061. [CrossRef]
71. Kröner, P.T.; Engels, M.M.; Glicksberg, B.S.; Johnson, K.W.; Mzaik, O.; van Hooft, J.E.; Wallace, M.B.; El-Serag, H.B.; Krittanawong, C. Artificial intelligence in gastroenterology: A state-of-the-art review. *World J. Gastroenterol.* **2021**, *27*, 6794–6824. [CrossRef]
72. Bian, Y.; Li, J.; Ye, C.; Jia, X.; Yang, Q. Artificial intelligence in medical imaging: From task-specific models to large-scale foundation models. *Chin. Med. J.* **2025**, *138*, 651–663. [CrossRef]
73. Vivek, K.; Papalois, V. AI and Machine Learning in Transplantation. *Transplantology* **2025**, *6*, 23. [CrossRef]
74. Ratziu, V.; Francque, S.; Behling, C.A.; Cejvanovic, V.; Cortez-Pinto, H.; Iyer, J.S.; Krarup, N.; Le, Q.; Sejling, A.S.; Tiniakos, D.; et al. Artificial intelligence scoring of liver biopsies in a phase II trial of semaglutide in nonalcoholic steatohepatitis. *Hepatology* **2024**, *80*, 173–185. [CrossRef]
75. Tiyarattanachai, T.; Apiparakoon, T.; Chaichuen, O.; Sukcharoen, S.; Yimsawad, S.; Jangsirikul, S.; Chaikajornwat, J.; Siriwong, N.; Burana, C.; Siritaweechai, N.; et al. Artificial intelligence assists operators in real-time detection of focal liver lesions during ultrasound: A randomized controlled study. *Eur. J. Radiol.* **2023**, *165*, 110932. [CrossRef]
76. Zhang, Y.; Cui, J.; Wan, W.; Liu, J. Multimodal Imaging under Artificial Intelligence Algorithm for the Diagnosis of Liver Cancer and Its Relationship with Expressions of EZH2 and p57. *Comput. Intell. Neurosci.* **2022**, *2022*, 4081654. [CrossRef]
77. Briceño, J.; Cruz-Ramírez, M.; Prieto, M.; Navasa, M.; Ortiz de Urbina, J.; Orti, R.; Gómez-Bravo, M.Á.; Otero, A.; Varo, E.; Tomé, S.; et al. Use of artificial intelligence as an innovative donor-recipient matching model for liver transplantation: Results from a multicenter Spanish study. *J. Hepatol.* **2014**, *61*, 1020–1028. [CrossRef] [PubMed]
78. Yang, M.; Zhao, Y.; Li, C.; Weng, X.; Li, Z.; Guo, W.; Jia, W.; Feng, F.; Hu, J.; Sun, H.; et al. Multimodal integration of liquid biopsy and radiology for the noninvasive diagnosis of gallbladder cancer and benign disorders. *Cancer Cell* **2025**, *43*, 398–412.e4. [CrossRef]
79. Goyal, H.; Mann, R.; Gandhi, Z.; Perisetti, A.; Zhang, Z.; Sharma, N.; Saligram, S.; Inamdar, S.; Tharian, B. Application of artificial intelligence in pancreaticobiliary diseases. *Ther. Adv. Gastrointest. Endosc.* **2021**, *14*, 2631774521993059. [CrossRef] [PubMed]
80. Podină, N.; Gheorghe, E.C.; Constantin, A.; Cazacu, I.; Croitoru, V.; Gheorghe, C.; Balaban, D.V.; Jinga, M.; Țieranu, C.G.; Săftoiu, A. Artificial Intelligence in Pancreatic Imaging: A Systematic Review. *United Eur. Gastroenterol. J.* **2025**, *13*, 55–77. [CrossRef]
81. Lopez-Ramirez, F.; Syailendra, E.A.; Tixier, F.; Kawamoto, S.; Fishman, E.K.; Chu, L.C. Early detection of pancreatic cancer on computed tomography: Advancements with deep learning. *Radiol. Adv.* **2025**, *2*, umaf028. [CrossRef]

82. Korfiatis, P.; Suman, G.; Patnam, N.G.; Trivedi, K.H.; Karbhari, A.; Mukherjee, S.; Cook, C.; Klug, J.R.; Patra, A.; Khasawneh, H.; et al. Automated Artificial Intelligence Model Trained on a Large Data Set Can Detect Pancreas Cancer on Diagnostic Computed Tomography Scans As Well As Visually Occult Preinvasive Cancer on Prediagnostic Computed Tomography Scans. *Gastroenterology* **2023**, *165*, 1533–1546.e4. [CrossRef]
83. Udriștoiu, A.L.; Podină, N.; Ungureanu, B.S.; Constantin, A.; Georgescu, C.V.; Bejinariu, N.; Pirici, D.; Burtea, D.E.; Gruionu, L.; Udriștoiu, S.; et al. Deep learning segmentation architectures for automatic detection of pancreatic ductal adenocarcinoma in EUS-guided fine-needle biopsy samples based on whole-slide imaging. *Endosc. Ultrasound* **2024**, *13*, 335–344. [CrossRef]
84. Shi, Y.J.; Zhang, H.; Wang, L.L.; Liu, Y.L.; Zhu, H.T.; Li, X.T.; Wei, Y.Y.; Sun, Y.S. Deep learning automatic segmentation and radiomics model for diagnosing pancreatic solid neoplasms in MRI. *BMC Cancer* **2025**, *25*, 1563. [CrossRef]
85. Kui, B.; Pintér, J.; Molontay, R.; Nagy, M.; Farkas, N.; Gede, N.; Vincze, Á.; Bajor, J.; Gódi, S.; Czimmer, J.; et al. EASY-APP: An artificial intelligence model and application for early and easy prediction of severity in acute pancreatitis. *Clin. Transl. Med.* **2022**, *12*, e842. [CrossRef]
86. Cui, H.; Zhao, Y.; Xiong, S.; Feng, Y.; Li, P.; Lv, Y.; Chen, Q.; Wang, R.; Xie, P.; Luo, Z.; et al. Diagnosing Solid Lesions in the Pancreas With Multimodal Artificial Intelligence: A Randomized Crossover Trial. *JAMA Netw. Open* **2024**, *7*, e2422454. [CrossRef] [PubMed]
87. Chen, P.T.; Wu, T.; Wang, P.; Chang, D.; Liu, K.L.; Wu, M.S.; Roth, H.R.; Lee, P.C.; Liao, W.C.; Wang, W. Pancreatic Cancer Detection on CT Scans with Deep Learning: A Nationwide Population-based Study. *Radiology* **2023**, *306*, 172–182. [CrossRef] [PubMed]
88. Kovatchev, B.; Castillo, A.; Pryor, E.; Kollar, L.L.; Barnett, C.L.; DeBoer, M.D.; Brown, S.A. Neural-Net Artificial Pancreas: A Randomized Crossover Trial of a First-in-Class Automated Insulin Delivery Algorithm. *Diabetes Technol. Ther.* **2024**, *26*, 375–382. [CrossRef]
89. Atlas, E.; Nimri, R.; Miller, S.; Grunberg, E.A.; Phillip, M. MD-logic artificial pancreas system: A pilot study in adults with type 1 diabetes. *Diabetes Care* **2010**, *33*, 1072–1076. [CrossRef] [PubMed]
90. Ali, A.; Alghamdi, M.; Marzuki, S.S.; Tengku Din, T.A.D.A.A.; Yamin, M.S.; Alrashidi, M.; Alkhazi, I.S.; Ahmed, N. Exploring AI Approaches for Breast Cancer Detection and Diagnosis: A Review Article. *Breast Cancer* **2025**, *17*, 927–947. [CrossRef]
91. Debellotte, O.; Dookie, R.L.; Rinkoo, F.; Kar, A.; Salazar González, J.F.; Saraf, P.; Aflahe Iqbal, M.; Ghazaryan, L.; Mukunde, A.C.; Khalid, A.; et al. Artificial Intelligence and Early Detection of Breast, Lung, and Colon Cancer: A Narrative Review. *Cureus* **2025**, *17*, e79199. [CrossRef]
92. Barrett, M.; Boyne, J.; Brandts, J.; Brunner-La Rocca, H.P.; De Maesschalck, L.; De Wit, K.; Dixon, L.; Eurlings, C.; Fitzsimons, D.; Golubnitschaja, O.; et al. Artificial intelligence supported patient self-care in chronic heart failure: A paradigm shift from reactive to predictive, preventive and personalised care. *EPMA J.* **2019**, *10*, 445–464. [CrossRef]
93. Garzonis, K.; Mann, E.; Wyrzykowska, A.; Kanellakis, P. Improving Patient Outcomes: Effectively Training Healthcare Staff in Psychological Practice Skills: A Mixed Systematic Literature Review. *Eur. J. Psychol.* **2015**, *11*, 535–556. [CrossRef]
94. Arioz, U.; Allsop, M.J.; Goodman, W.D.; Timmons, S.; Simbirtseva, K.; Mlakar, I.; Mocnik, G. Artificial intelligence-based approaches for advance care planning: A scoping review. *BMC Palliat. Care* **2025**, *24*, 268. [CrossRef]
95. Bienefeld, N.; Keller, E.; Grote, G. AI Interventions to Alleviate Healthcare Shortages and Enhance Work Conditions in Critical Care: Qualitative Analysis. *J. Med. Internet Res.* **2025**, *27*, e50852. [CrossRef]
96. Lång, K.; Josefsson, V.; Larsson, A.M.; Larsson, S.; Högberg, C.; Sartor, H.; Hofvind, S.; Andersson, I.; Rosso, A. Artificial intelligence-supported screen reading versus standard double reading in the Mammography Screening with Artificial Intelligence trial (MASAI): A clinical safety analysis of a randomised, controlled, non-inferiority, single-blinded, screening accuracy study. *Lancet Oncol.* **2023**, *24*, 936–944. [CrossRef] [PubMed]
97. Dembrower, K.; Crippa, A.; Colón, E.; Eklund, M.; Strand, F.; ScreenTrustCAD Trial Consortium. Artificial intelligence for breast cancer detection in screening mammography in Sweden: A prospective, population-based, paired-reader, non-inferiority study. *Lancet Digit. Health* **2023**, *5*, e703–e711. [CrossRef] [PubMed]
98. Sadeh-Sharvit, S.; Camp, T.D.; Horton, S.E.; Hefner, J.D.; Berry, J.M.; Grossman, E.; Hollon, S.D. Effects of an Artificial Intelligence Platform for Behavioral Interventions on Depression and Anxiety Symptoms: Randomized Clinical Trial. *J. Med. Internet Res.* **2023**, *25*, e46781. [CrossRef]
99. Repici, A.; Spadaccini, M.; Antonelli, G.; Correale, L.; Maselli, R.; Galtieri, P.A.; Pellegatta, G.; Capogreco, A.; Milluzzo, S.M.; Lollo, G.; et al. Artificial intelligence and colonoscopy experience: Lessons from two randomised trials. *Gut* **2022**, *71*, 757–765. [CrossRef]
100. Marcuzzi, A.; Nordstoga, A.L.; Bach, K.; Aasdahl, L.; Nilsen, T.I.L.; Bardal, E.M.; Boldermo, N.Ø.; Falkener Bertheussen, G.; Marchand, G.H.; Gismervik, S.; et al. Effect of an Artificial Intelligence-Based Self-Management App on Musculoskeletal Health in Patients with Neck and/or Low Back Pain Referred to Specialist Care: A Randomized Clinical Trial. *JAMA Netw. Open* **2023**, *6*, e2320400. [CrossRef] [PubMed]

101. Wallace, M.B.; Sharma, P.; Bhandari, P.; East, J.; Antonelli, G.; Lorenzetti, R.; Vieth, M.; Speranza, I.; Spadaccini, M.; Desai, M.; et al. Impact of Artificial Intelligence on Miss Rate of Colorectal Neoplasia. *Gastroenterology* **2022**, *163*, 295–304.e5. [CrossRef]
102. Liaw, S.Y.; Tan, J.Z.; Bin Rusli, K.D.; Ratan, R.; Zhou, W.; Lim, S.; Lau, T.C.; Seah, B.; Chua, W.L. Artificial Intelligence Versus Human-Controlled Doctor in Virtual Reality Simulation for Sepsis Team Training: Randomized Controlled Study. *J. Med. Internet Res.* **2023**, *25*, e47748. [CrossRef] [PubMed]
103. Luo, H.; Xu, G.; Li, C.; He, L.; Luo, L.; Wang, Z.; Jing, B.; Deng, Y.; Jin, Y.; Li, Y.; et al. Real-time artificial intelligence for detection of upper gastrointestinal cancer by endoscopy: A multicentre, case-control, diagnostic study. *Lancet Oncol.* **2019**, *20*, 1645–1654. [CrossRef] [PubMed]
104. Wilson, P.M.; Ramar, P.; Philpot, L.M.; Soleimani, J.; Ebbert, J.O.; Storlie, C.B.; Morgan, A.A.; Schaeferle, G.M.; Asai, S.W.; Herasevich, V.; et al. Effect of an Artificial Intelligence Decision Support Tool on Palliative Care Referral in Hospitalized Patients: A Randomized Clinical Trial. *J. Pain Symptom Manag.* **2023**, *66*, 24–32. [CrossRef]
105. Papachristou, P.; Söderholm, M.; Pallon, J.; Taloyan, M.; Polesie, S.; Paoli, J.; Anderson, C.D.; Falk, M. Evaluation of an artificial intelligence-based decision support for the detection of cutaneous melanoma in primary care: A prospective real-life clinical trial. *Br. J. Dermatol.* **2024**, *191*, 125–133. [CrossRef] [PubMed]
106. Kazemzadeh, K. Artificial intelligence in ophthalmology: Opportunities, challenges, and ethical considerations. *Med. Hypothesis Discov. Innov. Ophthalmol. J.* **2025**, *14*, 255–272. [CrossRef]
107. Hosny, A.; Parmar, C.; Quackenbush, J.; Schwartz, L.H.; Aerts, H.J.W.L. Artificial intelligence in radiology. *Nat. Rev. Cancer* **2018**, *18*, 500–510. [CrossRef]
108. Bhandari, A. Revolutionizing Radiology with Artificial Intelligence. *Cureus* **2024**, *16*, e72646. [CrossRef]
109. Försch, S.; Klauschen, F.; Hufnagl, P.; Roth, W. Artificial Intelligence in Pathology. *Dtsch. Arztebl. Int.* **2021**, *118*, 194–204. [CrossRef]
110. Shafi, S.; Parwani, A.V. Artificial intelligence in diagnostic pathology. *Diagn. Pathol.* **2023**, *18*, 109. [CrossRef]
111. Sajithkumar, A.; Thomas, J.; Saji, A.M.; Ali, F.; E.K, H.H.; Adampulan, H.A.G.; Sarathchand, S. Artificial Intelligence in pathology: Current applications, limitations, and future directions. *Ir. J. Med. Sci.* **2024**, *193*, 1117–1121. [CrossRef]
112. Hashemian, H.; Peto, T.; Ambrósio, R., Jr.; Lengyel, I.; Kafieh, R.; Muhammed Noori, A.; Khorrami-Nejad, M. Application of Artificial Intelligence in Ophthalmology: An Updated Comprehensive Review. *J. Ophthalmic Vis. Res.* **2024**, *19*, 354–367. [CrossRef]
113. Lim, J.I.; Regillo, C.D.; Sadda, S.R.; Ipp, E.; Bhaskaranand, M.; Ramachandra, C.; Solanki, K. Artificial Intelligence Detection of Diabetic Retinopathy: Subgroup Comparison of the EyeArt System with Ophthalmologists' Dilated Examinations. *Ophthalmol. Sci.* **2022**, *3*, 100228. [CrossRef]
114. Li, Z.; Wang, L.; Wu, X.; Jiang, J.; Qiang, W.; Xie, H.; Zhou, H.; Wu, S.; Shao, Y.; Chen, W. Artificial intelligence in ophthalmology: The path to the real-world clinic. *Cell Rep. Med.* **2023**, *4*, 101095. [CrossRef]
115. Johnson, K.W.; Torres Soto, J.; Glicksberg, B.S.; Shameer, K.; Miotto, R.; Ali, M.; Ashley, E.; Dudley, J.T. Artificial Intelligence in Cardiology. *J. Am. Coll. Cardiol.* **2018**, *71*, 2668–2679. [CrossRef] [PubMed]
116. Karatzia, L.; Aung, N.; Aksentijevic, D. Artificial intelligence in cardiology: Hope for the future and power for the present. *Front. Cardiovasc. Med.* **2022**, *9*, 945726. [CrossRef]
117. Patrascanu, O.S.; Tutunaru, D.; Musat, C.L.; Dragostin, O.M.; Fulga, A.; Nechita, L.; Ciubara, A.B.; Piraianu, A.I.; Stamate, E.; Poalelungi, D.G.; et al. Future Horizons: The Potential Role of Artificial Intelligence in Cardiology. *J. Pers. Med.* **2024**, *14*, 656. [CrossRef] [PubMed]
118. Lotter, W.; Hassett, M.J.; Schultz, N.; Kehl, K.L.; Van Allen, E.M.; Cerami, E. Artificial Intelligence in Oncology: Current Landscape, Challenges, and Future Directions. *Cancer Discov.* **2024**, *14*, 711–726. [CrossRef]
119. Kann, B.H.; Hosny, A.; Aerts, H.J.W.L. Artificial intelligence for clinical oncology. *Cancer Cell* **2021**, *39*, 916–927. [CrossRef] [PubMed]
120. Shimizu, H.; Nakayama, K.I. Artificial intelligence in oncology. *Cancer Sci.* **2020**, *111*, 1452–1460. [CrossRef]
121. Rizzo, M.; Dawson, J.D. AI in Neurology: Everything, Everywhere, All at Once Part 1: Principles and Practice. *Ann. Neurol.* **2025**, *98*, 211–230. [CrossRef]
122. Khalilian, M.; Godefroy, O.; Roussel, M.; Mousavi, A.; Aarabi, A. Post-stroke outcome prediction based on lesion-derived features. *NeuroImage Clin.* **2025**, *45*, 103747. [CrossRef] [PubMed]
123. Voigtlaender, S.; Pawelczyk, J.; Geiger, M.; Vaios, E.J.; Karschnia, P.; Cudkowicz, M.; Dietrich, J.; Haraldsen, I.R.J.H.; Feigin, V.; Owolabi, M.; et al. Artificial intelligence in neurology: Opportunities, challenges, and policy implications. *J. Neurol.* **2024**, *271*, 2258–2273. [CrossRef]
124. Shrestha, U.K. Emerging role of artificial intelligence in gastroenterology and hepatology. *World J. Gastroenterol.* **2025**, *31*, 111495. [CrossRef]
125. Urquhart, S.A.; Christof, M.; Coelho-Prabhu, N. The impact of artificial intelligence on the endoscopic assessment of inflammatory bowel disease-related neoplasia. *Ther. Adv. Gastroenterol.* **2025**, *18*, 17562848251348574. [CrossRef] [PubMed]

126. El-Sayed, A.; Lovat, L.B.; Ahmad, O.F. Clinical Implementation of Artificial Intelligence in Gastroenterology: Current Landscape, Regulatory Challenges, and Ethical Issues. *Gastroenterology* **2025**, *169*, 518–530. [CrossRef]
127. Nahm, W.J.; Sohail, N.; Burshtein, J.; Goldust, M.; Tsoukas, M. Artificial Intelligence in Dermatology: A Comprehensive Review of Approved Applications, Clinical Implementation, and Future Directions. *Int. J. Dermatol.* **2025**, *64*, 1568–1583. [CrossRef] [PubMed]
128. Young, A.T.; Xiong, M.; Pfau, J.; Keiser, M.J.; Wei, M.L. Artificial Intelligence in Dermatology: A Primer. *J. Investig. Dermatol.* **2020**, *140*, 1504–1512. [CrossRef]
129. Fliorent, R.; Fardman, B.; Podwojniak, A.; Javaid, K.; Tan, I.J.; Ghani, H.; Truong, T.M.; Rao, B.; Heath, C. Artificial intelligence in dermatology: Advancements and challenges in skin of color. *Int. J. Dermatol.* **2024**, *63*, 455–461. [CrossRef]
130. Cruz-Gonzalez, P.; He, A.W.; Lam, E.P.; Ng, I.M.C.; Li, M.W.; Hou, R.; Chan, J.N.; Sahni, Y.; Vinas Guasch, N.; Miller, T.; et al. Artificial intelligence in mental health care: A systematic review of diagnosis, monitoring, and intervention applications. *Psychol. Med.* **2025**, *55*, e18. [CrossRef]
131. Prégent, J.; Chung, V.H.; El Adib, I.; Désilets, M.; Hudon, A. Applications of Artificial Intelligence in Psychiatry and Psychology Education: Scoping Review. *JMIR Med. Educ.* **2025**, *11*, e75238. [CrossRef] [PubMed]
132. Garcia, G. The role of AI in transforming psychiatric-mental health care: Enhancing the role of psychiatric-mental health nurse practitioners. *Nurs. Outlook* **2025**, *73*, 102461. [CrossRef]
133. Niazi, S.K.; Mariam, Z. Artificial intelligence in drug development: Reshaping the therapeutic landscape. *Ther. Adv. Drug Saf.* **2025**, *16*, 20420986251321704. [CrossRef]
134. Zhang, K.; Yang, X.; Wang, Y.; Yu, Y.; Huang, N.; Li, G.; Li, X.; Wu, J.C.; Yang, S. Artificial intelligence in drug development. *Nat. Med.* **2025**, *31*, 45–59. [CrossRef]
135. Fu, C.; Chen, Q. The future of pharmaceuticals: Artificial intelligence in drug discovery and development. *J. Pharm. Anal.* **2025**, *15*, 101248. [CrossRef]
136. Baghbani, S.; Mehrabi, Y.; Movahedinia, M.; Babaeinejad, E.; Joshaghanian, M.; Amiri, S.; Shahrezaee, M. The revolutionary impact of artificial intelligence in orthopedics: Comprehensive review of current benefits and challenges. *J. Robot. Surg.* **2025**, *19*, 511. [CrossRef]
137. Wu, S.; Miao, Y.; Mei, J.; Xiong, S. The Rise of Artificial Intelligence in Orthopedics: A Bibliometric and Visualization Analysis. *J. Multidiscip. Healthc.* **2025**, *18*, 6037–6050. [CrossRef]
138. Song, J.; Wang, G.C.; Wang, S.C.; He, C.R.; Zhang, Y.Z.; Chen, X.; Su, J.C. Artificial intelligence in orthopedics: Fundamentals, current applications, and future perspectives. *Mil. Med. Res.* **2025**, *12*, 42. [CrossRef]
139. Smith, M.E.; Zalesky, C.C.; Lee, S.; Gottlieb, M.; Adhikari, S.; Goebel, M.; Wegman, M.; Garg, N.; Lam, S.H.F. Artificial Intelligence in Emergency Medicine: A Primer for the Nonexpert. *J. Am. Coll. Emerg. Physicians Open* **2025**, *6*, 100051. [CrossRef] [PubMed]
140. Stewart, J.; Sprivulis, P.; Dwivedi, G. Artificial intelligence and machine learning in emergency medicine. *Emerg. Med. Australas. EMA* **2018**, *30*, 870–874. [CrossRef] [PubMed]
141. Amiot, F.; Potier, B. Artificial Intelligence (AI) and Emergency Medicine: Balancing Opportunities and Challenges. *JMIR Med. Inform.* **2025**, *13*, e70903. [CrossRef] [PubMed]
142. Kerth, J.L.; Bischops, A.C.; Hagemeister, M.; Reinhart, L.; Konrad, K.; Heinrichs, B.; Meissner, T. Künstliche Intelligenz in der Gesundheitsvorsorge von Kindern und Jugendlichen—Anwendungsmöglichkeiten und Akzeptanz [Artificial intelligence in preventive medicine for children and adolescents-applications and acceptance]. *Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz* **2025**, *68*, 907–914. [CrossRef] [PubMed]
143. Dinc, R.; Ardic, N. The Next Frontiers in Preventive and Personalized Healthcare: Artificial Intelligent-powered Solutions. *J. Prev. Med. Public Health = Yebang Uihakhoe Chi* **2025**, *58*, 441–452. [CrossRef] [PubMed]
144. Whiteson, H.Z.; Frishman, W.H. Artificial Intelligence in the Prevention and Detection of Cardiovascular Disease. *Cardiol. Rev.* **2025**, *33*, 239–242. [CrossRef]
145. Lopes, S.; Rocha, G.; Guimarães-Pereira, L. Artificial intelligence and its clinical application in Anesthesiology: A systematic review. *J. Clin. Monit. Comput.* **2024**, *38*, 247–259. [CrossRef]
146. Dost, A.; Alaraj, R.; Mayet, R.; Agrawal, D.K. Reshaping Anesthesia with Artificial Intelligence: From Concept to Reality. *Anesth. Crit. Care* **2025**, *7*, 77–90. [CrossRef]
147. Hashimoto, D.A.; Witkowski, E.; Gao, L.; Meireles, O.; Rosman, G. Artificial Intelligence in Anesthesiology: Current Techniques, Clinical Applications, and Limitations. *Anesthesiology* **2020**, *132*, 379–394. [CrossRef]
148. Scott, I.A.; van der Vegt, A.; Lane, P.; McPhail, S.; Magrabi, F. Achieving large-scale clinician adoption of AI-enabled decision support. *BMJ Health Care Inform.* **2024**, *31*, e100971. [CrossRef]
149. Fihn, S.; Saria, S.; Mendonça, E.; Hain, S.; Matheny, M.; Shah, N.; Liu, H.; Auerbach, A. Deploying artificial intelligence in clinical settings. In *Artificial Intelligence in Health Care: The Hope, the Hype, the Promise, the Peril*; Whicher, D., Ahmed, M., Israni, S.T., Eds.; National Academies Press: Washington, DC, USA, 2023. Available online: https://www.ncbi.nlm.nih.gov/books/NBK605954/ (accessed on 20 December 2025).

150. Bahl, M. Artificial intelligence in clinical practice: Implementation considerations and barriers. *J. Breast Imaging* **2022**, *4*, 632–639. [CrossRef]
151. de Hond, A.A.H.; Leeuwenberg, A.M.; Hooft, L.; Kant, I.M.J.; Nijman, S.W.J.; van Os, H.J.A.; Aardoom, J.J.; Debray, T.P.A.; Schuit, E.; van Smeden, M.; et al. Guidelines and quality criteria for artificial intelligence-based prediction models in healthcare: A scoping review. *npj Digit. Med.* **2022**, *5*, 2. [CrossRef]
152. Liu, M.; Ning, Y.; Teixayavong, S.; Liu, X.; Mertens, M.; Shang, Y.; Li, X.; Miao, D.; Liao, J.; Xu, J.; et al. A scoping review and evidence gap analysis of clinical AI fairness. *npj Digit. Med.* **2025**, *8*, 360. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

algorithms

MDPI

Systematic Review

Reclaiming XAI as an Innovation in Healthcare: Bridging Rule-Based Systems

Hanvedes Daovisan *, Charin Suwanwong, Pitchada Prasittichok, Narulmon Prayai and Phaktada Choowan

Behavioral Science Research Institute, Srinakharinwirot University, Bangkok 10110, Thailand
* Correspondence: hanvedes@g.swu.ac.th

Abstract

The adoption of explainable artificial intelligence (XAI) in healthcare has been increasingly framed as dependent on transparency, trustworthiness, and accountability. The objective of this study was the reclamation of rule-based systems within XAI as innovations aligned with healthcare accountability. A scientometric mapping analysis was conducted using publications indexed in Scopus between 1 January 2018 and 20 May 2025. The search strategy was applied to 1034 records. From these, 892 were screened, 238 duplicates were removed, and 654 studies were retained in accordance with the PRISMA 2020 framework. Thematic cluster analysis, co-authorship structures, and keyword co-occurrence patterns were visualised through VOSviewer 1.6.20. Transparency, accountability, and trustworthiness were established as central values for clinical integration. Expanding domains were identified in smart healthcare, digital health, healthcare technology, and mHealth, while interpretability was observed to remain underrepresented. Rule-based systems, frequently in hybrid forms, were demonstrated to bridge algorithmic complexity with interpretability. This bridging was interpreted as reinforcing physician confidence, regulatory compliance, and patient safety. It was concluded that the advancement of XAI in healthcare has been shaped by the interplay of ethical principles, methodological innovation, and digital health applications. Practical implications, theoretical contributions, and potential limitations were systematically addressed.

Keywords: healthcare; innovation; explainable artificial intelligence; rule-based systems; scientometric mapping analysis

1. Introduction

Over the past decade, diagnostic precision, treatment personalisation, and predictive analytics in healthcare have been reshaped through advances in machine learning and data-driven modelling. Within this transformation, XAI has been established as a central mechanism for transparency, accountability, and clinician trust [1,2]. Decision pathways have been rendered opaque by deep learning architectures, thereby intensifying the significance of XAI. The demand for interpretable models has been reinforced by regulatory mandates, documented cases of algorithmic bias, and the diffusion of clinical artificial intelligence into safety-critical domains [3]. Historical reliance on rule-based systems has preserved trustworthiness through the embedding of expert-defined heuristics and explicit reasoning processes. Auditability and compliance with clinical guidelines have similarly been enabled by such systems [4]. In this study, XAI has been positioned not merely as a compliance instrument but as a conceptual bridge between the transparency of rule-based frameworks and the predictive strength of contemporary artificial intelligence. Through

this bridging role, epistemic reliability and operational safety in healthcare decision-making have been advanced.

Advances in computational performance have not been accompanied by the removal of opacity that constrains interpretability, reproducibility, and regulatory compliance in healthcare artificial intelligence systems [5]. Black-box architectures, particularly deep neural networks, have not been associated with clinically actionable explanations and have therefore been restricted in integration into diagnostic and therapeutic workflows [2]. Deficits in transparency have been linked to the erosion of clinician confidence, the limitation of patient comprehension, and the intensification of institutional exposure to legal and ethical risks when algorithmic outputs diverge from clinical judgement. Historically, these challenges were mitigated by rule-based systems through the provision of explicit decision pathways and guideline-conformant reasoning. Predictive accuracy, however, was reduced in complex and data-rich environments owing to restricted adaptability [6]. The present study has addressed this gap by reconciling the epistemic rigour and auditability of rule-based frameworks with the adaptive capacity of modern artificial intelligence. Through this reconciliation, healthcare innovation has been advanced as both technologically progressive and socially accountable.

XAI has been defined as a suite of computational methods through which decision-making by artificial intelligence is rendered transparent, interpretable, and aligned with human cognitive frameworks [7]. Rule-based systems have been characterised by explicitly codified if–then logic grounded in domain expertise, thereby ensuring traceability and regulatory auditability in clinical practice [8]. Techniques of deep learning—such as saliency mapping, surrogate modelling, and counterfactual explanation—have been developed to approximate interpretability; however, the causal clarity intrinsic to rule-based reasoning has not been preserved [5]. The model advanced in this study has been positioned not as a post hoc interpretative layer but as an integrative design paradigm in which adaptive statistical learning is embedded within transparent, rule-governed structures. Within this configuration, interpretability has been maintained without loss of predictive accuracy. This synthesis has been proposed as a mechanism through which clinical trust is sustained, informed decision-making is supported, and compliance with emerging frameworks for healthcare artificial intelligence governance is ensured [9].

The integration of XAI with rule-based systems has been positioned as a transformative innovation through which the epistemic foundations of clinical artificial intelligence are reconceptualised. Adaptive statistical inference has been embedded within inherently interpretable logical structures so that the dual demands for transparency, accountability, auditability, and ethically aligned decision-making in healthcare are met [7]. Broader implications have been carried by this integration, including regulatory harmonisation, cross-institutional interoperability, and the consolidation of sustained stakeholder trust. These outcomes have been expected to facilitate the adoption of artificial intelligence in domains that have previously exhibited resistance to opaque algorithms [10]. The synthesis has been defined by a dual commitment to predictive performance and interpretability. Through the establishment of this balance, a methodological precedent has been set for artificial intelligence integration with the potential to reshape clinical workflows, enhance shared decision-making, and reinforce the social licence under which healthcare artificial intelligence operates.

The urgency of advancing XAI—understood as understandable, usable, interpretable, responsible, and accountable artificial intelligence—has been generated by the convergence of rapid algorithmic adoption with intensifying regulatory, ethical, and trust imperatives [11]. Conventional approaches have treated interpretability as a secondary, post hoc extension of predictive modelling, and transparency has therefore not been embedded

as a structural property of healthcare artificial intelligence systems [12,13]. A departure from this convention has been proposed in the present study through a hybrid paradigm in which the explicit logical reasoning of rule-based systems is structurally integrated with the adaptive capacity of statistical learning models. Within this configuration, causal interpretability, regulatory auditability, and clinician usability have been preserved without compromise in predictive performance [14,15]. The aim has been defined not only as the establishment of a model architecture but also as the articulation of a design philosophy that addresses existing transparency deficits while creating a replicable precedent for responsible artificial intelligence deployment. Through this advance, healthcare innovation policy, clinical workflow integration, and cross-domain governance of intelligent systems have been positioned to be reshaped.

Research Questions

The resurgence and evolving role of rule-based systems within XAI in healthcare are examined. A combined approach of systematic review and scientometric mapping of Scopus-indexed literature is employed to investigate the integration of rule-based logic into contemporary applications. Implications for transparency, trust, and clinical adoption are assessed. The inquiry is directed by the following research questions:

RQ1: What publication trends emerge in Scopus-indexed healthcare literature on XAI and rule-based systems, and which authors, institutions, and countries contribute most prominently to the identified thematic domains?

RQ2: What research gaps remain in connecting rule-based models to interpretability in healthcare, and which areas are underexplored in the current literature?

RQ3: Which thematic clusters emerge from systematically mapping the integration of rule-based systems into XAI for healthcare?

Interpretability in XAI is reframed in this study as a theoretical contribution extending beyond narrow technical accounts. A methodological advance is developed through hybrid modelling, in which rule-based reasoning is integrated with data-driven approaches to balance rigour with adaptability. A practical contribution is established through the alignment of XAI design with emerging global regulatory frameworks, thereby ensuring compliance while supporting transparency and trust. Collectively, these findings are interpreted as indicative of a broader shift towards responsible, human-centred, and innovative adoption of XAI in healthcare systems worldwide.

2. Materials and Methods

2.1. Scientometric Mapping Design

Rule-based systems in XAI were reclaimed as a timely innovation for healthcare in response to the rising demand for interpretable models in safety-critical contexts. Interpretability, innovation, and scientometric mapping were positioned as bridging mechanisms for the repositioning of rule-based reasoning within current debates on XAI. Knowledge advanced beyond prior narrative reviews through a scientometric mapping design grounded in Scopus data, thereby enabling systematic, replicable, and large-scale mapping of research dynamics [16]. Bibliographic records were extracted, metadata were cleaned, co-occurrence patterns were mapped, and thematic domains were clustered through the VOSviewer workflow. Intellectual structures and knowledge networks were visualised by VOSviewer, thereby exposing the ways in which rule-based approaches have either been integrated or marginalised in healthcare XAI research. A theoretical contribution is made by reframing interpretability as a methodological core; a methodological contribution is offered by validating scientometric mapping over narrative synthesis; and a practical

contribution is provided by guiding healthcare stakeholders in balancing innovation with transparency [17].

2.2. Scientometric Mapping Search Strategies

Scopus was selected as the primary data source for this scientometric mapping study owing to its extensive coverage of peer-reviewed journals, conference proceedings, and healthcare-related artificial intelligence literature. International and high-impact publications were indexed within Scopus, thereby ensuring the inclusion of validated scholarly outputs and rendering the database appropriate for the mapping of research frontiers. The timeframe from 1 January 2018 to 20 May 2025 was defined to capture the accelerated growth of XAI research in healthcare during this period. Variations in terminology were addressed through the application of Boolean operators, truncation, and wildcards in the search strategy. The following exact query was executed: TITLE-ABS-KEY ("explainable AI" OR "XAI" OR "interpretable AI" OR "rule-based AI") AND TITLE-ABS-KEY ("healthcare" OR "healthcare innovation") AND PUBYEAR > 2018 AND PUBYEAR < 2025.

The workflow was structured in accordance with PRISMA 2020 guidelines (Supplementary Materials), and transparency as well as reproducibility were ensured through sequential steps of identification, screening, eligibility, and inclusion [18]. Replication was facilitated, and the reliability of the scientometric mapping analysis was strengthened through this design. Conceptual terms in XAI and healthcare were combined in the query using AND/OR operators and quotation marks to ensure phrase precision. A total of 1304 records were retrieved from the initial Scopus search, forming the foundation for the PRISMA flow diagram. The scientometric mapping process is illustrated in Figure 1 through the PRISMA flow diagram.

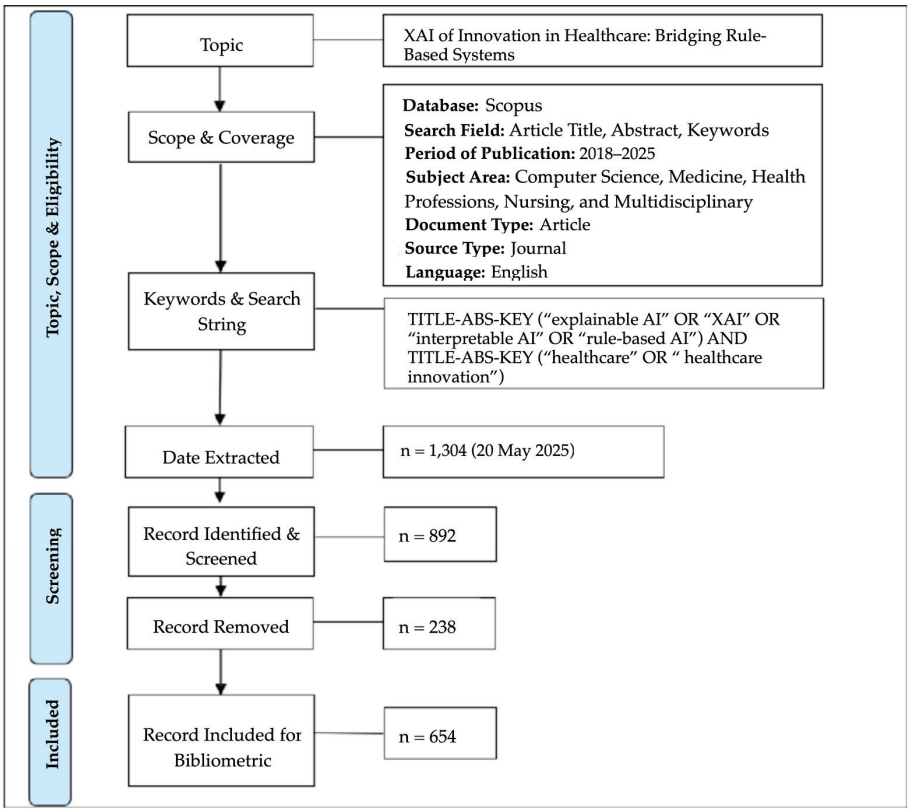


Figure 1. PRISMA flow diagram illustrating the process of scientometric mapping analysis.

2.3. Scientometric Mapping Eligibility Criteria

At the eligibility stage, the full texts of all potentially relevant publications were assessed to determine inclusion. Studies were excluded if they were not relevant to healthcare, if they did not maintain an explicit focus on XAI, or if they addressed only black-box models without rule-based interpretability. Transparency was ensured by defining rule-based models as those employing explicit rules, logic, or symbolic reasoning, whereas black-box models were characterised by opaque architectures such as deep neural networks. Healthcare relevance was operationalised through the presence of clinical settings, health technologies, or smart healthcare applications. Following the application of these eligibility criteria, 654 publications were retained as the final set for scientometric mapping analysis.

At the screening stage, duplicates were removed through EndNote and cross-checked in Excel to ensure accuracy. Only peer-reviewed articles published in English and explicitly addressing XAI in healthcare were retained for inclusion, while conference abstracts, grey literature, and studies outside healthcare applications were excluded. Screening was conducted at two levels. At the first level, titles and abstracts were reviewed for relevance, resulting in the exclusion of 412 records owing to irrelevance or non-healthcare scope. At the second level, full texts were assessed for methodological and thematic fit, and 238 records were excluded for lacking an XAI focus or for reliance solely on black-box models. For the studies retained, bibliographic data—including authors, titles, abstracts, keywords, and citations—were exported from Scopus and converted into CSV and RIS formats for import into VOSviewer. Quality control was maintained through the cross-checking of metadata and the resolution of inconsistencies across records. The process has been visualised through the PRISMA flow diagram (Figure 1), which illustrates the systematic inclusion and exclusion of records prior to final analysis.

Within the PRISMA eligibility framework, systematic data cleaning was undertaken to ensure both accuracy and reproducibility. Author names and institutional affiliations were disambiguated to eliminate duplicate identities and false co-authorship links. Keywords were harmonised through the merging of synonyms, the correction of inconsistencies, and the standardisation of spelling variants. From this process, a controlled dictionary was created and subsequently applied as input for VOSviewer. The dictionary increased the validity of the co-occurrence analysis and enhanced the transparency of the scientometric mapping workflow by ensuring that each cleaning step was explicit and replicable.

2.4. Scientometric Mapping Models

The visualisation of the similarities equation was employed by Van Eck and Waltman [16] to develop bibliometric maps for each item pair i and j, with the similarity s_{ij} ($s_{ij} \geq 0$) used as input. The mapping of items is expressed through the following equation:

$$V(\mathbf{x}_1, \ldots, \mathbf{x}_n) \sum_{i<j} \| \mathbf{x}_i - \mathbf{x}_j \|^2 \tag{1}$$

subject to

$$\frac{2}{n(n-1)} \sum_{i<j} \| \mathbf{x}_i - \mathbf{x}_j \| = 1 \tag{2}$$

The VOS method is used to calculate the weighted composite of squared distances between each pair of items. The location of each item i is expressed as follows:

$$\mathbf{x}_i^* = \frac{\sum_{j \neq i} s_{ij} \mathbf{x}_j}{\sum_{j \neq i} s_{ij}} \tag{3}$$

where the item i is defined as a weighted average of the locations. According to Newman [19], the equation of co-authorship networks is specified using N and M, where N

denotes the number of researchers and M the total number of publications. In constructing co-authorship networks, A = $[a_{ik}]$ is used to denote the $N \times M$ authorship matrix. The key element a_{ik} of this matrix equals 1 if researcher i is an author of publication k and 0 otherwise. Furthermore, n_k is used to denote the number of authors of publication k, as expressed by the following equation:

$$n_k = \sum_{i=1}^{n} a_{ik} \tag{4}$$

where the mean represents the number of authors, with $n_k > 1$ for each publication k. The equation is expressed in symmetrical form as **U** = $[u_{ij}]$ to denote the full-counting co-authorship matrix of size $N \times N$. The key element u_{ij} represents the linkage between researchers i and j defined as follows:

$$u_{ij} = \sum_{k=1}^{M} a_{ik} a_{jk} \tag{5}$$

where the co-authorship matrix **U** is defined as follows:

$$\mathbf{U} = \mathbf{A}\mathbf{A}^{T} \tag{6}$$

Hence, **U** is defined as the co-authorship matrix, and $\boldsymbol{A}$ as the post-multiplying authorship matrix, with self-links in the co-authorship matrix **U** set to 0. The fractional-counting co-authorship matrix is expressed as $\mathbf{U}^{*}$ = $[u_{ij}]$, where the linkage between researchers i and j, u^{*}_{ij}, is defined as follows:

$$u_{ij} = \sum_{k=1}^{M} \frac{a_{ik} a_{jk}}{n_k - 1} \tag{7}$$

where the co-authorship matrix **U*** is defined as follows:

$$\mathbf{U}^{*} = \boldsymbol{A}\mathrm{diag}\left(\boldsymbol{A}^{\mathrm{T}}1 - 1\right)^{1}\boldsymbol{A}^{\mathrm{T}} \tag{8}$$

where diag(**v**) denotes a diagonal matrix with the elements of the vector **v**, while 1 denotes a column vector of length N with all elements equal to 1 and **U*** 0. The co-occurrence links are denoted between nodes i and j, (c_{ij} = $u_{ji} \geq 0$), while s_{ij} denotes the association strengths of nodes i and j. The relationship is expressed in the following equation:

$$s_{ij} = \frac{2\boldsymbol{m}\boldsymbol{c}_{\boldsymbol{ij}}}{\boldsymbol{c}_{\boldsymbol{i}}\boldsymbol{c}_{\boldsymbol{j}}} \tag{9}$$

where c_i denotes the total number of links of node i, while m denotes the total number of links in the network. The relationship is defined as follows:

$$c_i = \sum_{\boldsymbol{j} \neq \boldsymbol{i}} \boldsymbol{c}_{\boldsymbol{ij}} \text{ and } \mathrm{m} = \frac{1}{2}\sum_{\boldsymbol{i}} \boldsymbol{c}_{\boldsymbol{i}} \tag{10}$$

where node i is represented by a vector $x_i \in \mathrm{R^p}$ denoting the location of node i in a p-dimensional map ($p = 2$). A positive value of x_i defines the cluster to which node i belongs. The mapping process is unified in the following equation:

$$\mathrm{v}(x_i, \ldots, x_n) = \sum_{\boldsymbol{i} < \boldsymbol{j}} \boldsymbol{s}_{\boldsymbol{ij}} \boldsymbol{d}^{2}_{\boldsymbol{ij}} - \sum_{\boldsymbol{i} < \boldsymbol{j}} \boldsymbol{d}_{\boldsymbol{ij}} \tag{11}$$

with respect to $x_i, \ldots, x_n$. The distance between nodes i and j is denoted as d_{ij} and expressed as follows:

$$d_{ij} = \parallel x_i - x_j \parallel = \sqrt{\sum_{k=1}^{p} \left(x_{ik} - x_{jk} \right)^2} \tag{12}$$

where the mapping of the parameter is defined as follows:

$$d_{ij} = \begin{cases} 0 & if\ x_i = x_j \\ \frac{1}{\gamma} & if\ x_i \neq x_j \end{cases} \tag{13}$$

where the parameter γ is defined as the resolution ($\gamma > 0$). A larger value of γ results in stronger separation, with the co-occurrence structure interpreted in terms of repulsive forces between nodes.

According to Traag et al. [20], the Leiden algorithm was defined as a modularity-based clustering approach with guaranteed well-connected communities. The modularity function with the resolution parameter γ is expressed as follows:

$$Q(y) = \frac{1}{2m} \sum_{i,j} \left[A_{ij} - y \frac{k_i k_j}{2m} \right] \delta(c_i, c_j) \tag{14}$$

Here, $Q(\gamma)$ is defined as the modularity score for a given partition. A_{ij} is defined as the weight of the edge between nodes i and j represented by the normalised co-occurrence (e.g., s_{ij}). $k_i = \sum_i A_{ij}$ is defined as the degree (or strength) of node i. $m = \frac{1}{2}\sum_i A_{ij}$ is defined as the total edge weight in the network. $\gamma > 0$ is defined as the resolution parameter that controls cluster granularity. $\delta(c_i, c_j)$ is defined as an indicator function, taking the value 1 if nodes i and j belong to the same community and 0 otherwise. A low value of γ results in fewer and larger clusters, whereas a high value of γ produces more and smaller clusters.

The mapping objective was formalised as the systematic identification of structural patterns in the literature on XAI in healthcare, with particular emphasis on the integration of legacy rule-based systems. Methodological rigour was ensured through the adoption of full counting in VOSviewer for both co-citation and co-occurrence analyses, thereby incorporating the complete weight of each link across the dataset. Clustering was conducted with the Leiden algorithm, which was selected over the Louvain approach because it produces more stable and internally consistent partitions, particularly in large bibliometric networks. The resolution parameter was explicitly reported to specify cluster granularity, and multiple values were tested to demonstrate robustness. Edge thresholds were applied in accordance with VOSviewer's recommended practices to remove spurious connections while preserving meaningful relational structures. To maximise transparency, exact map reporting was reproduced, including the counting scheme, resolution values, and thresholds, so that the clustering process remains fully traceable and replicable for subsequent scientometric analyses of XAI in healthcare.

2.5. Scientometric Mapping Clusters

Thematic clustering in scientometric mapping analysis has been employed to identify major research foci, intellectual structures, and emerging trends within a field by grouping co-occurring keywords and citations into distinct thematic areas [21]. In the context of XAI and rule-based systems in healthcare, thematic clusters generated through VOSviewer have typically revealed domains such as clinical decision support systems, algorithmic transparency, interpretable machine learning, trust in artificial intelligence, and hybrid models. For instance, one cluster was centred on the integration of rule-based logic

with deep learning to enhance diagnostic accuracy, while another was oriented towards the ethical and regulatory dimensions of explainability in clinical practice [5,7]. These clusters were employed to map the intellectual landscape of the field, thereby disclosing both established and emerging areas of inquiry that are critical to the advancement of trustworthy artificial intelligence in healthcare.

2.6. Scientometric Mapping Analyses

The scientometric mapping analytical procedure was undertaken through a structured process beginning with the extraction of bibliographic data from the Scopus database and concluding with visualisation and interpretation in VOSviewer. A comprehensive keyword search was conducted with Boolean operators to identify literature on XAI, rule-based systems, and healthcare within the period 1 January 2018 to 20 May 2025. The dataset generated from this search comprised metadata including authorship, titles, abstracts, keywords, citations, and institutional affiliations. These records were exported in CSV and RIS formats for bibliometric processing. The data were subsequently cleaned to remove duplicates and irrelevant entries. Terms were normalised to ensure consistency through the merging of synonyms and the standardisation of author names.

The cleaned data were imported into VOSviewer for multiple forms of analysis. Co-authorship analysis was conducted to map collaboration networks among authors and institutions. Co-citation analysis was undertaken to identify the intellectual structure of the field. Co-occurrence analysis of keywords was applied to detect thematic clusters [17, 20]. VOSviewer's distance-based visualisation technique was employed to represent relationships, with the proximity between nodes indicating the strength of association. Clustering algorithms embedded in the software were applied to group related items into thematic areas, thereby enabling the identification of core research topics such as interpretability in medical artificial intelligence, clinical decision support, and hybrid rule-learning models. These visual and quantitative outputs were subsequently interpreted in relation to the study's objective, and rule-based systems were shown to have been reintegrated into contemporary XAI research in healthcare.

3. Results

3.1. RQ1: What Publication Trends Emerge in Scopus-Indexed Healthcare Literature on XAI and Rule-Based Systems, and Which Authors, Institutions, and Countries Contribute Most Prominently to the Identified Thematic Domains?

3.1.1. Synthesis of Results

In response to *RQ1*, a scientometric mapping analysis was undertaken. Marked growth in Scopus-indexed healthcare literature on XAI and rule-based systems was documented between 1 January 2018 and 20 May 2025. The annual growth curve of publications was shown to demonstrate a consistent upward trajectory, indicating sustained scholarly interest in XAI as a healthcare innovation that reclaims and integrates rule-based systems. In parallel, the annual growth curve of citations, supported by a positive trendline, was confirmed as evidence that the field expanded in volume while also gaining recognition and influence across the research community. Taken together, these patterns were interpreted as validating the dual momentum of XAI scholarship: increasing output was shown to signal productivity, while rising citation impact was recognised as reflecting the consolidation of relevance. The role of rule-based systems was substantiated as a central bridge between technical robustness and clinical accountability. Publication and citation trends in XAI for healthcare are illustrated in Figure 2.

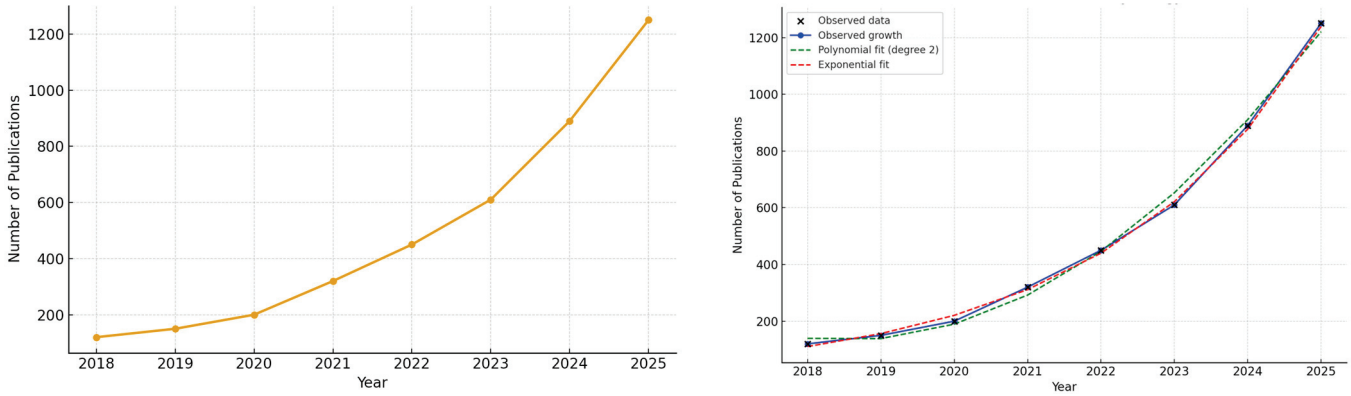


Figure 2. Publication and citation trends in XAI for healthcare: (**a**) annual growth curve of publications; (**b**) annual growth curve of citations with trendline.

3.1.2. Co-Authorship Analysis

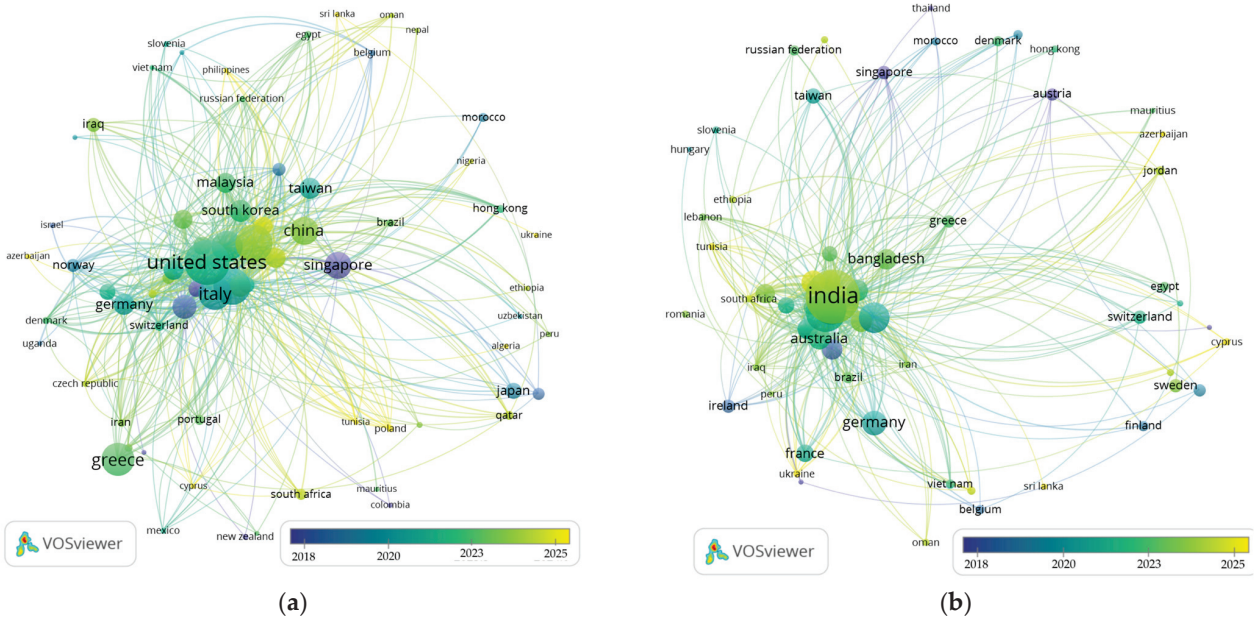
The co-authorship analysis of Scopus data identified India as the leading contributor to XAI research in healthcare (see Table 1). The highest publication output (126), strong citation impact (1906), and a centrality score of 0.95 were recorded for India. The United States was ranked second in publication output (94) but achieved the highest centrality score (1.23) and citation impact (2427), confirming its pivotal role in connecting global collaborations. The United Kingdom, Italy, and Saudi Arabia were also distinguished by substantial citation counts and moderate centrality scores, indicating active bilateral and multilateral partnerships. VOSviewer visualisation reinforced these patterns. Dense South–South and South–North linkages were highlighted, particularly between India and Germany, Bangladesh, Australia, and France. Emerging contributions were observed from Bangladesh, Egypt, and Vietnam, reflecting an expanding diversification of the global XAI research ecosystem. India and the United States were positioned as central hubs (see Figure 3).

Table 1. Top ten countries in co-authorship networks related to XAI in healthcare.

Rank	Country/Region	Year	Publications	Citations	Centrality
1	India	2018	126	1906	0.95
2	United States	2018	94	2427	1.23
3	United Kingdom	2018	55	2064	0.87
4	Saudi Arabia	2018	50	598	0.48
5	Italy	2018	49	1506	0.61
6	South Korea	2018	39	842	0.35
7	Australia	2018	36	1568	0.39
8	China	2018	34	1316	0.37
9	Pakistan	2018	33	723	0.28
10	Canada	2018	26	989	0.15

The United States was revealed as the central hub in XAI healthcare research. Extensive linkages with Italy, Germany, Singapore, and China were demonstrated, underscoring the role of the United States in cross-border innovation. Emerging contributions were identified from Malaysia, Greece, and Morocco, with these states frequently providing regional bridges in Asia, the Middle East, and Africa. The highest publication output (13) was recorded by King Saud University. The greatest citation counts (464 and 377) were achieved by the University of Southern Queensland and Khalifa University of Science and Technology, respectively, reflecting the influence and quality of their research. High productivity was also exhibited by institutions in India and the Middle East, including the

Manipal Academy, Vellore Institute of Technology, and Princess Nourah Bint Abdulrahman University. Their performance was shown to indicate a diversification of XAI research leadership beyond Western institutions. A strong trend towards international and multi-institutional collaboration was revealed by the findings (see Table 2).



(**a**) (**b**)

Figure 3. Publication trends based on a country-level scientometric mapping analysis: (**a**) co-authorship networks; (**b**) citation rankings.

Table 2. The top ten most productive organizations in XAI in healthcare.

Rank	Institutions	Publications	Citations	Centrality
1	King Saud University	13	186	0.78
2	Princess Nourah Bint Abdulrahman University	11	91	0.64
3	Manipal Academy of Higher Education	10	146	0.56
4	Vellore Institute of Technology	9	249	0.47
5	Consiglio Nazionale delle Ricerche	9	210	0.40
6	Manipal Institute of Technology	9	143	0.34
7	Khalifa University of Science and Technology	8	377	0.28
8	Korea University	7	239	0.21
9	University of Southern Queensland	7	464	0.17
10	Prince Mohammad Bin Fahd University	7	11	0.15

3.1.3. Citation Analysis

Citation analysis of Scopus data revealed a rapidly growing body of high-impact research published in journals spanning computing, biomedical engineering, and data science. The most cited work, authored by Linardatos in Entropy (1780 citations), was shown to establish foundational frameworks and evaluation methods for XAI. Central publication venues were represented by leading journals such as Information Fusion (impact factor 23.9) and ACM Computing Surveys (impact factor 51.1). Their prominence was shown to reflect both methodological rigour and broad interdisciplinary reach. Sustained scholarly influence was confirmed by authors including Kaur, Yang, and Loh, whose work appeared in outlets with high H-indices (232, 179, 150). Applied journals such as IEEE Access and Sensors were shown to highlight the integration of XAI into sensor-based and diagnostic healthcare technologies. The most influential

contributions were published between 2021 and 2023, confirming the emergent yet maturing nature of the field (see Figure 4).

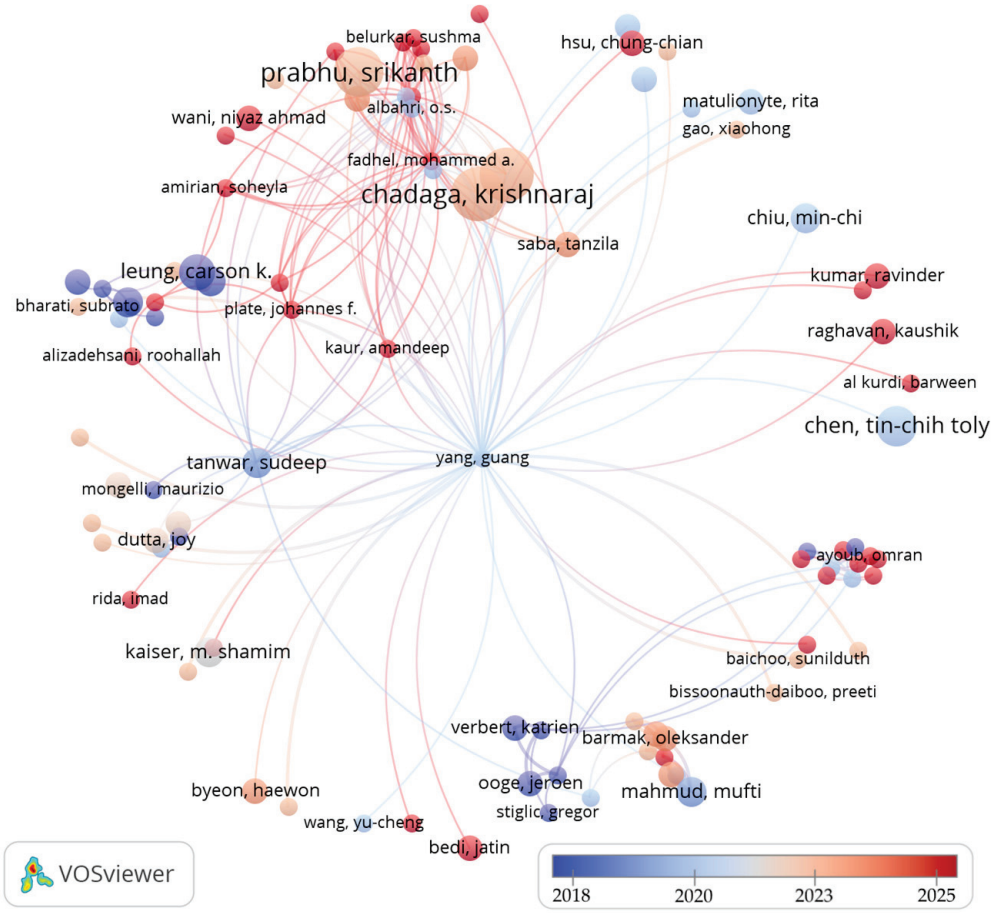


Figure 4. Combined co-authorship and citation network of researchers in XAI in healthcare.

3.2. RQ2: What Research Gaps Remain in Connecting Rule-Based Models to Interpretability in Healthcare, and Which Areas Are Underexplored in the Current Literature?

3.2.1. Co-Occurrence Analysis

In response to *RQ2*, the findings were shown to indicate that although XAI spans multiple healthcare domains, its strongest momentum has been observed in interpretable diagnostics for life-threatening and complex diseases. This trend was driven by demands for trust, accountability, and clinical relevance. VOSviewer co-occurrence analysis revealed an increasing concentration of XAI in high-impact medical fields such as cardiology, oncology, and neurology. Strong keyword linkages were identified with conditions including lung cancer, brain tumours, Alzheimer's disease, and heart failure. Central terms such as disease diagnosis, pathology, and cancer classification were positioned as hubs connecting diverse clinical applications. These hubs were shown to underscore the dual emphasis on diagnostic accuracy and interpretability across medical specialties. The recent emergence of terms such as Alzheimer's disease, dementia, and skin cancer was found to suggest growing interest in explainable models for complex, image-rich, and time-series data (see Figure 5).

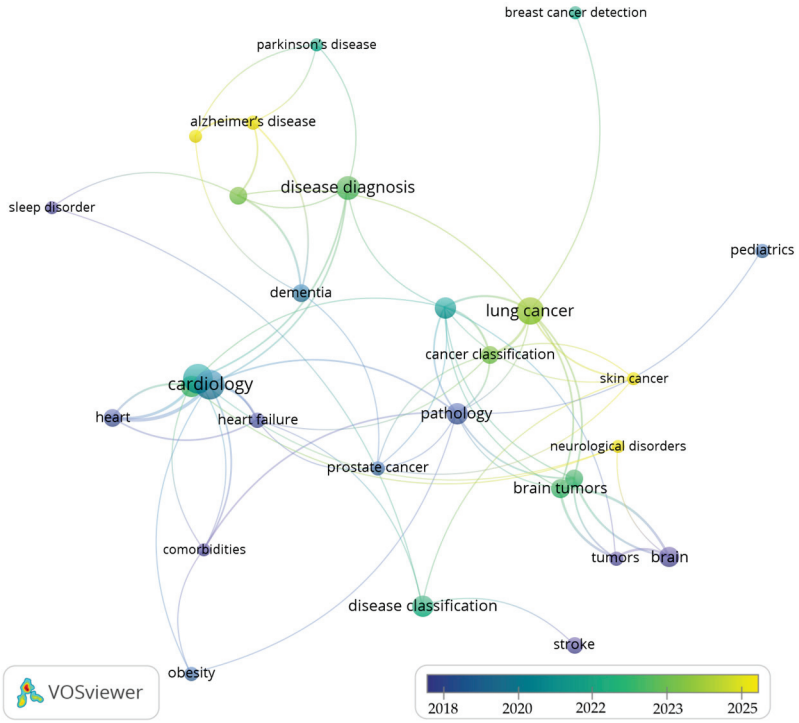
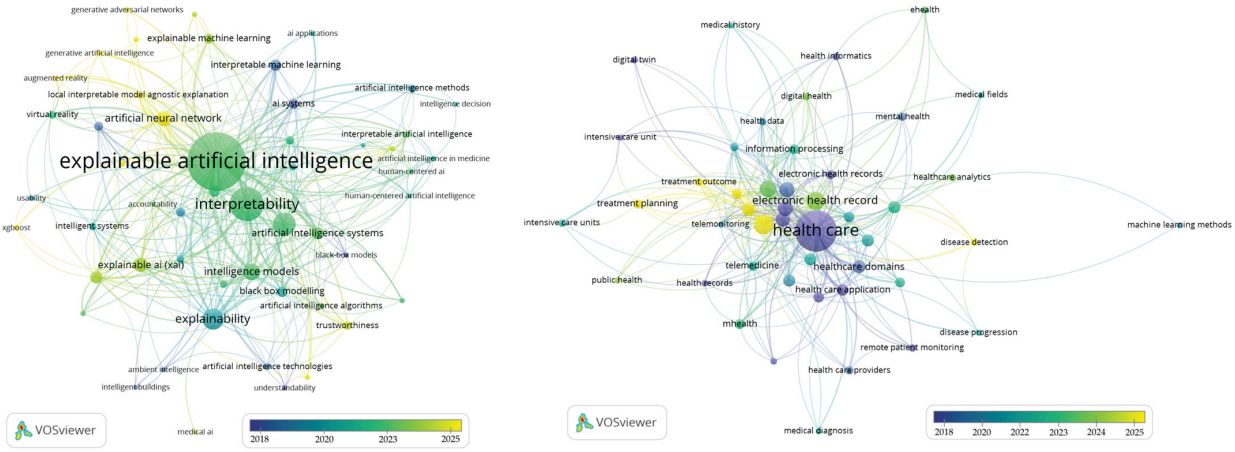


Figure 5. Disease categories classified using XAI in healthcare.

The first VOSviewer map, centred on XAI, was shown to reveal a tightly connected conceptual network dominated by terms such as interpretability, explainability, explainable machine learning, and artificial neural networks. Co-occurring terms indicated that discourse has remained primarily oriented towards making complex black-box models more understandable. High-frequency nodes such as LIME and SHAP were identified, confirming a methodological emphasis on post hoc interpretation tools. These methods underscored the reconciliation of model complexity with user trust. Strong linkages were observed to human-centred artificial intelligence, decision-making, and trust, reflecting sustained concern with clinical usability and ethical accountability. Temporal overlay analysis indicated recent growth in subtopics such as interpretable deep learning and generative adversarial networks, signalling the diversification of methodological approaches within healthcare-focused XAI (see Figure 6).



(**a**) (**b**)

Figure 6. Overlay visualization showing: (**a**) XAI; (**b**) healthcare.

3.2.2. Co-Occurrence Networks

Three interlinked clusters were revealed through the co-occurrence analysis as shaping the discourse on XAI in healthcare (see Table 3). The first cluster was delineated as the conceptual core. It was dominated by XAI, responsible artificial intelligence, and artificial intelligence technologies, with emerging prominence assigned to generative artificial intelligence and trustworthy artificial intelligence. This configuration was interpreted as reflecting a sustained emphasis on ethics, transparency, and the responsible design of intelligent systems. The second cluster was identified as the methodological toolkit. Within this cluster, XAI models, explainability, and accountability were positioned as a central locus of innovation in healthcare. Post hoc tools such as LIME and SHAP were also recorded, highlighting persistent efforts to reconcile the opacity of complex black-box models with interpretability. The third cluster was associated with healthcare applications. It was led by the themes of healthcare, healthcare technology, and healthcare applications, and was extended into domains such as mHealth, smart healthcare, telemonitoring, and digital health. The key terms for the strategic diagram clusters of XAI in healthcare research were depicted in Figure 7. The strategic diagram clusters of XAI and healthcare research are illustrated in Table 4.

Table 3. Key terms identified through a co-word analysis of XAI in healthcare.

Cluster	Rank	Key Co-Wording	Centrality	Duration	Range
One	1	XAI	0.95	2018–2025	
	2	Responsible AI	0.87	2018–2025	
	3	AI technologies	0.85	2018–2025	
	4	Generative AI	0.76	2018–2025	
	5	Interpretable AI	0.66	2018–2025	
	6	Healthcare AI	0.58	2018–2025	
	7	AI techniques	0.55	2018–2025	
	8	GANs	0.54	2018–2025	
	9	Trustworthy AI	0.52	2018–2025	
Two	1	XAI model	0.97	2018–2025	
	2	Explainability	0.94	2018–2025	
	3	Accountability	0.90	2018–2025	
	4	Understandability	0.88	2018–2025	
	5	Usability	0.86	2018–2025	
	6	Interpretability	0.83	2018–2025	
	7	LIME	0.71	2018–2025	
	8	SHAP	0.59	2018–2025	
Three	1	Healthcare	0.96	2018–2025	
	2	Healthcare technology	0.87	2018–2025	
	3	Healthcare application	0.78	2018–2025	
	4	mhealth	0.72	2018–2025	
	5	Smart healthcare	0.68	2018–2025	
	6	Healthcare professionals	0.60	2018–2025	
	7	Healthcare delivery	0.53	2018–2025	
	8	Telemonitoring	0.48	2018–2025	
	9	Virtual reality	0.41	2018–2025	
	10	Digital health	0.37	2018–2025	
	11	Healthcare providers	0.30	2018–2025	
	12	Healthcare domains	0.22	2018–2025	

Note: ■ = denotes the range of high citation density; ■ = denotes the range of low citation density.

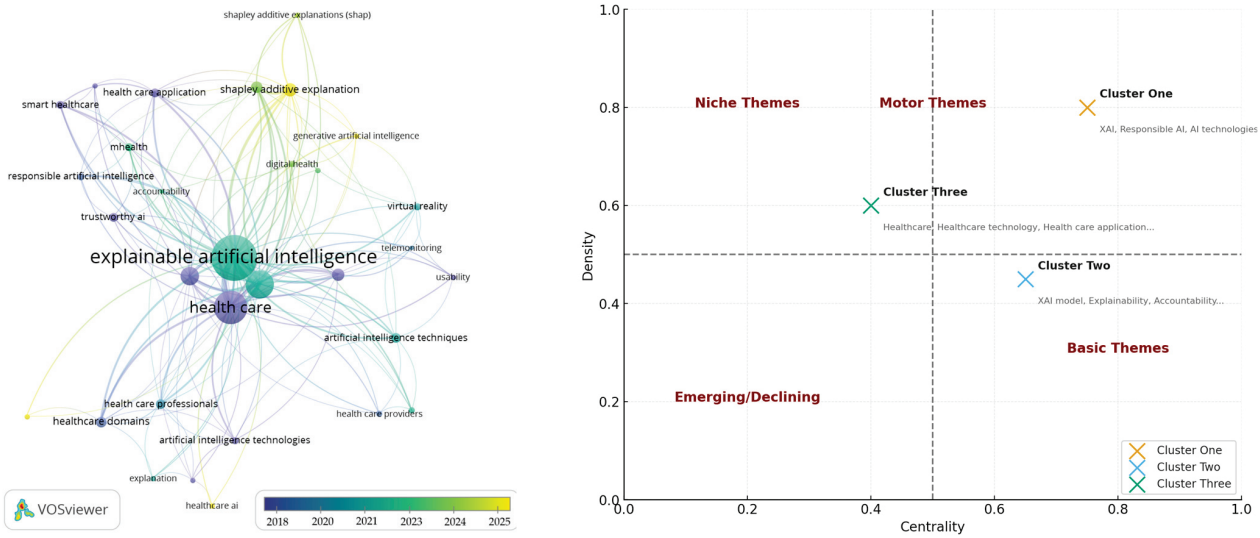


Figure 7. Key terms in XAI for healthcare: (**a**) key term clusters; (**b**) strategic diagram of research clusters.

Table 4. Strategic diagram clusters in XAI and healthcare research.

Cluster	Quadrant	Key Terms	Interpretation
One	Motor themes (high centrality, high density)	XAI, responsible AI, AI technologies, generative AI, interpretable AI, healthcare AI, AI techniques, GANs, and trustworthy AI	Cluster one was identified as driving the field. Conceptual maturity was shown, and strong links with adjacent domains were maintained. The cluster was positioned as forming the backbone of healthcare AI innovation, where responsible and trustworthy frameworks intersect with advanced technical methods.
Two	Basic themes (high centrality, low density)	XAI model, explainability, accountability, understandability, usability, interpretability, LIME, and SHAP	Cluster two was identified as forming a fundamental building block of the field. Wide links across domains were demonstrated, while internal cohesion was shown to be limited. The cluster was defined as establishing the methodological and conceptual foundations of XAI, thereby ensuring accountability, usability, and interpretability in healthcare systems.
Three	Niche themes (Low centrality, high density)	Healthcare, healthcare technology, healthcare application, mHealth, smart healthcare, healthcare professionals, healthcare delivery, telemonitoring, virtual reality, digital health, healthcare providers, and healthcare domains	Cluster three was identified as containing specialised and internally cohesive themes. Strong internal development was demonstrated, while weaker connections to the wider network were maintained. The cluster was characterised as embodying practical, domain-specific applications of AI and XAI in healthcare, tailored to contexts such as telemonitoring and digital health.

3.2.3. Cluster Networks

Interpretability was identified as the dominant construct within the cluster. It was characterised by the highest levels of frequency, centrality, and connectivity. Its position was established as the bridging concept between technical development and clinical utility. Explainability and transparency were subsequently observed, reinforcing the discourse concerned with rendering artificial intelligence systems more understandable. The cluster hierarchy was demonstrated as follows: C_1 = {interpretability (f = 163, c = 0.82, l = 1853) > explainability (f = 69, c = 0.75, l = 0) > transparency (f = 60, c = 0.66, l = 0)}. Additional terms were detected: ∪ {trustworthiness (f = 14, c = 0, l = 221), usability

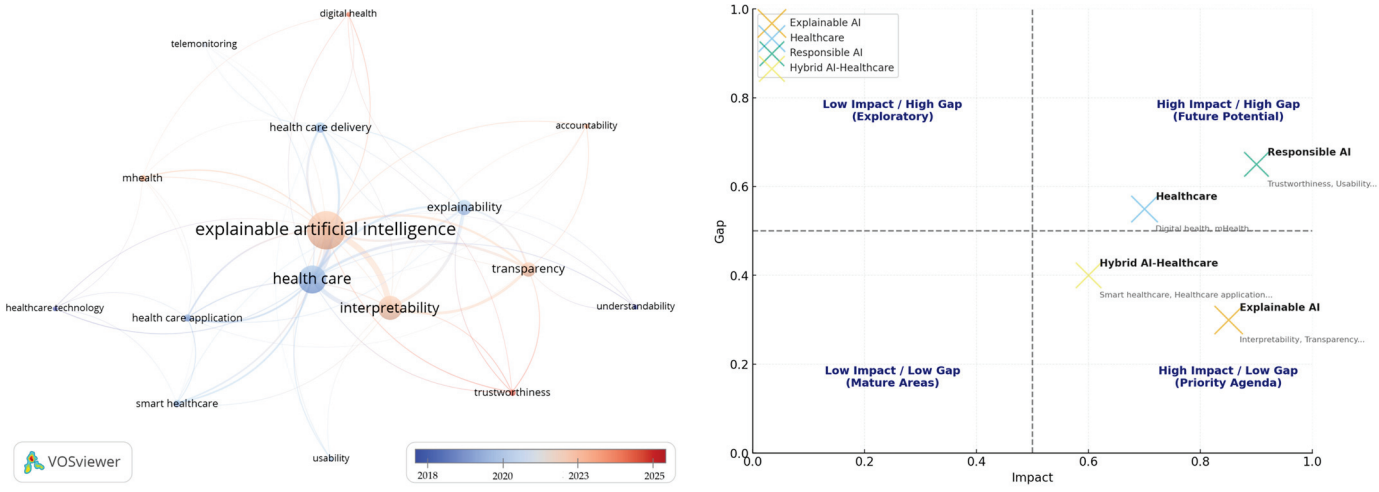
(f = 7, c = 0, l = 112), understandability (f = 6, c = 0, l = 83), and accountability (f = 5, c = 0, l = 53)}.

Within the healthcare-related cluster, the following constructs were identified: C_2 {healthcare delivery (f = 32, c = 0.46, l = 603), smart healthcare (f = 15, c = 0.32, l = 210), healthcare application (f = 17, c = 0.28, l = 205), telemonitoring (f = 5, c = 0.19, l = 162), digital health (f = 8, c = 0.16, l = 134), mhealth (f = 14, c = 0.17, l = 133), and healthcare technology (f = 6, c = 0.13, l = 64)}. Cluster theme networks were identified as low-frequency and low-connectivity constructs, thereby indicating dimensions that remain underexplored yet conceptually significant. The cluster networks of XAI in healthcare are illustrated in Table 5. The application of XAI in healthcare practice is depicted in Figure 8. The research agenda clusters and the impacts of XAI in healthcare are presented in Table 6.

Table 5. Cluster networks of XAI in healthcare.

Cluster	Rank	XAI in Healthcare	Occurrence	Centrality	Link	Range
One		XAI				
	1	Interpretability	163	0.82	1853	
	2	Explainability	69	0.75	814	
	3	Transparency	60	0.66	617	
	4	Trustworthiness	14	0.16	221	
	5	Usability	7	0.14	71	
	6	Understandability	6	0.09	58	
	7	Accountability	5	0.08	53	
Two		Healthcare				
	1	Healthcare delivery	32	0.46	603	
	2	Smart healthcare	15	0.32	210	
	3	Healthcare application	17	0.28	205	
	4	Telemonitoring	5	0.19	162	
	5	Digital health	8	0.16	134	
	6	mHealth	14	0.17	133	
	7	Healthcare technology	6	0.13	64	

Note: ■ = denotes the range of high citation density; ■ = denotes the range of low citation density.



(**a**) (**b**)

Figure 8. XAI in healthcare clusters: (**a**) XAI in healthcare networks; (**b**) research agenda quadrant clusters.

Table 6. Research agenda clusters and their research impacts.

Quadrant	Cluster	Research Impact
High impact/low gap (priority agenda)	XAI	Interpretability, transparency, and accountability were identified as the primary drivers of healthcare AI research. Their maturity was shown to position them for large-scale clinical adoption, policy integration, and standard-setting in trustworthy AI for healthcare.
High impact/high gap (future potential)	Responsible AI	Clusters such as trustworthiness, usability, and understandability were recognised as vital yet underdeveloped. They were identified as representing high-growth opportunities for advancing human-centred AI in healthcare. Their progression was shown to require cross-disciplinary collaboration and the establishment of robust ethical frameworks.
Low impact/high gap (exploratory)	Healthcare	Digital health, mHealth, and telemonitoring were identified as emerging domains that remain fragmented. Their impact was shown to be modest owing to technical, regulatory, and infrastructural gaps. With the integration of XAI, these domains were interpreted as having the potential to evolve into high-impact innovations in personalised and remote healthcare.
Low impact/low gap (mature areas)	Hybrid AI–Healthcare	Applied smart healthcare solutions and XAI-based applications were identified as relatively mature, with minimal conceptual gaps. Their impact was recognised as limited by saturation, yet they were shown to provide stable platforms for incremental innovation, benchmarking, and scaling across healthcare systems.

3.3. RQ3: Which Thematic Clusters Emerge from Systematically Mapping the Integration of Rule-Based Systems into XAI for Healthcare?

Scientometric Clustering of Rule-Based Systems

Articulating an answer to *RQ3*, the cluster network analysis was found to position rule-based systems as a bridging construct between explainability concepts and healthcare applications. Formally, node strength was computed as the sum of weighted connections with all other nodes, expressed as $S_i = \sum_{j=i} w_{ij}$. For rule-based systems, the strength development need was calculated as S_{RBS} 1.0 on research impact. This value was derived from cumulative linkages with XAI, interpretability, transparency, accountability, healthcare technology, smart healthcare, and digital health. The comparatively high score was interpreted as evidence that rule-based systems were embedded across network clusters of both higher impact, including future potential and priority agendas, and lower impact, including exploratory and mature areas. Rule-based systems were also shown to be consistently aligned with both conceptual and applied dimensions of XAI in healthcare. The ranked research agenda bubble chart for rule-based systems in XAI for healthcare is presented in Figure 9.

Rule-based systems were positioned as a transversal theme within hotspot clustering of XAI healthcare pathways (Figure 10). Their function was demonstrated to enhance connectivity and policy pathways between theoretical constructs—healthcare, responsible AI, rule-based systems, and XAI—and practical applications—actors, providers, regulators, and developers. Evidence was provided that rule-based systems were not treated as an isolated niche but were instead framed as a bridge cluster of strategic significance. Their contribution was identified in consolidating interpretability within policy frameworks while supporting exploratory directions, future potential, mature domains, and priority agendas. An integrated research–policy agenda for XAI in healthcare is presented in Table 7.

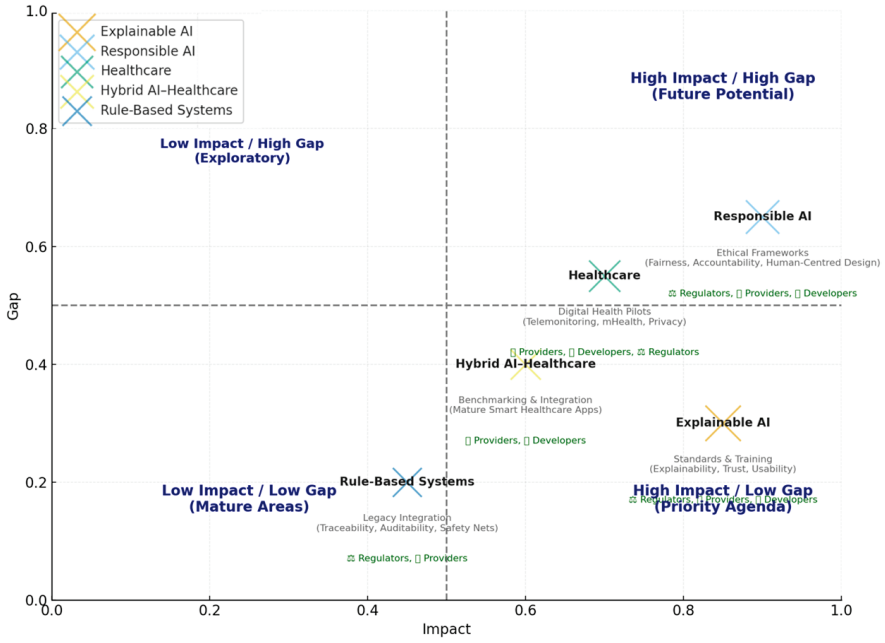


Figure 9. Ranked research clusters in XAI for healthcare.

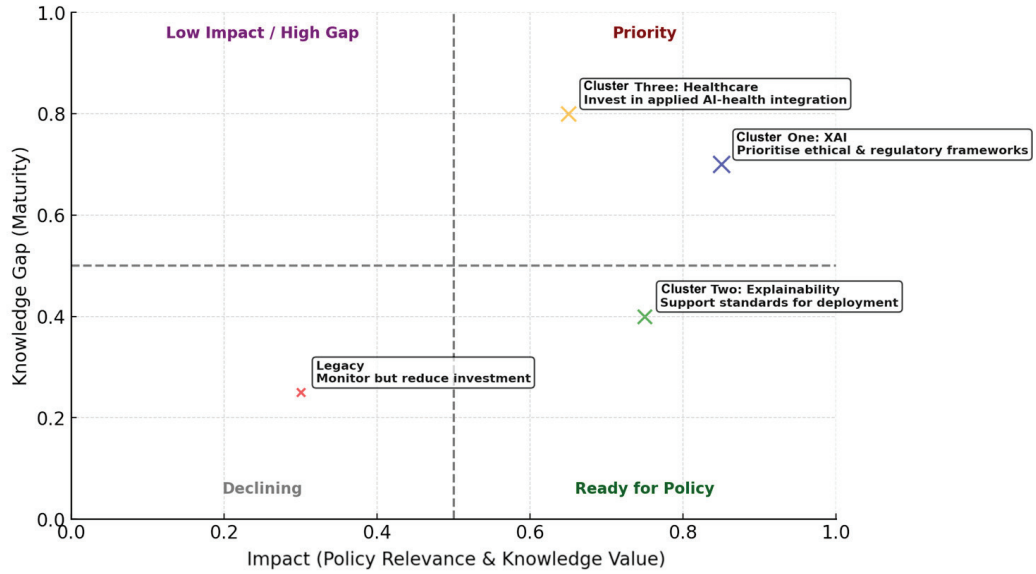


Figure 10. Research–policy pathways for rule-based systems in XAI for healthcare.

Table 7. An integrated research–policy agenda for XAI in healthcare.

Quadrant	Cluster	Regulators	Healthcare Providers	XAI Developers
High impact/low gap (priority agenda)	XAI	Mandatory explainability standards were proposed for establishment in relation to clinical AI tools. These standards were specified as requiring alignment with EU AI Act provisions, FDA regulations, and WHO guidelines in order to ensure safety and transparency.	The adoption of XAI-based decision support was recommended for healthcare systems to strengthen accountability and build patient trust. Training on interpretability was specified as requiring integration into clinical practice to ensure effective and responsible use.	Scalable explainability modules such as SHAP and LRP were recommended for embedding within healthcare platforms. Usability for non-technical users was identified as a priority in order to support practical adoption.

Table 7. *Cont.*

Quadrant	Cluster	Regulators	Healthcare Providers	XAI Developers
High impact/high gap (future potential)	Responsible AI	Ethical compliance frameworks were proposed for introduction to address fairness, usability, and trustworthiness. Incentives in the form of responsible innovation grants were recommended to accelerate ethical adoption.	Responsible AI applications were recommended for piloting within clinical workflows, including fairness in triage and usability in digital health applications.	Human-centred AI design principles were recommended for development with participatory input from clinicians and patients. Accountability-by-design was specified as requiring implementation to ensure responsible deployment.
Low impact/high gap (exploratory)	Healthcare (digital health, mHealth, tele-monitoring)	Regulatory sandboxes were recommended for the provision of testing mHealth and telemonitoring solutions with integrated explainability safeguards. Data privacy risks were identified as requiring mitigation to ensure safe and trustworthy adoption.	Telehealth pilots were recommended for implementation with XAI to ensure interpretation of AI outputs for patients. Digital literacy training was identified as requiring development to support patient understanding and engagement.	Lightweight, explainable models were recommended for design to support mobile and remote monitoring. Interoperability with healthcare IT systems was identified as a priority to enable seamless integration.
Low impact/low gap (mature areas)	Hybrid AI–healthcare	Certification frameworks were recommended for maintenance in relation to mature smart healthcare applications. Interoperability standards across providers were specified as requiring assurance to guarantee consistent and reliable deployment.	Benchmarking tools were recommended for use in assessing AI integration within clinical decision-making. Existing workflows were identified as requiring refinement rather than replacement to support safe and effective adoption.	Established healthcare AI models were recommended for improvement to enhance efficiency and scalability. Interpretability features were specified as requiring incremental enhancement without the introduction of disruptive change.
Low impact/low gap (mature /declining areas)	Rule-based systems	Archival standards for legacy systems were recommended for preservation. Their traceability and audit features were specified as requiring integration into modern regulatory guidelines.	Rule-based systems were recommended for retention as fallback decision aids in sensitive cases where transparency is paramount.	Rule-based logic was recommended for translation into hybrid XAI models. Clinical knowledge bases were specified as requiring embedding to enhance explainability and ensure domain grounding.

4. Discussion

The purpose of this study was defined as the examination of XAI as an emerging innovation in healthcare, with particular emphasis on its bridging role in rule-based systems. A scientometric mapping analysis was conducted on 654 studies indexed in the Scopus database between 2018 and 2025. Through VOSviewer analysis, three distinct clusters were identified—XAI, XAI models, and healthcare—as reported in Table 3. The principal findings were derived from two cluster networks that mapped XAI and healthcare, as presented in Table 5, together with the integrated cluster of XAI in healthcare illustrated in Figure 5. The evidence was shown to demonstrate that the research landscape was organised into conceptual groupings that collectively underscored the progressive convergence of XAI techniques with healthcare applications.

In response to *RQ1*, the scientometric mapping analysis demonstrated that although black-box models such as deep neural networks have remained dominant within predictive analytics, growing scholarly attention has been directed towards XAI techniques that

prioritise interpretability, particularly in high-stakes clinical contexts [5,22]. A notable resurgence of rule-based systems was identified, frequently in hybrid forms, providing a bridge between algorithmic complexity and human interpretability [23]. Thematic clusters generated through VOSviewer were shown to reveal convergence around clinical decision support systems, algorithmic accountability, and human–AI collaboration. These findings were interpreted as indicating that future research and innovation will increasingly favour models designed to achieve both predictive performance and explainability [15].

The analysis revealed a concentrated thematic focus on foundational values essential to the integration of artificial intelligence in healthcare. These values were found to comprise interpretability, explainability, transparency, trustworthiness, usability, understandability, and accountability. They were identified as central within the literature, reflecting persistent concern that artificial intelligence systems must be not only technically robust but also ethically and operationally viable in clinical practice [5,24]. The co-occurrence analysis was further demonstrated to show that terms such as explainability and trust frequently intersected with debates on clinical decision support systems and algorithmic outcomes. These intersections were interpreted as indicating that artificial intelligence tools lacking transparency may erode physician confidence and compromise patient safety [25].

Within the cluster, understandability was emphasised as a prerequisite for clinical adoption, reinforcing the requirement that models be aligned with human cognitive processes in order to be actionable [8]. Accountability was also identified as a critical concern, particularly in discussions of responsibility when artificial intelligence systems contribute to medical errors, thereby underscoring the need for traceable and justifiable decision pathways [26]. The prominence of these themes was shown to indicate a paradigm shift in artificial intelligence research—from an exclusive focus on performance towards a more holistic emphasis on human-centred design that supports ethical implementation, legal compliance, and real-world usability in healthcare systems. These findings were found to align with current regulatory trends and ethical frameworks advocating explainable and accountable AI, marking a decisive turn in the discourse on medical artificial intelligence deployment [10,11].

In response to *RQ2*, keyword co-occurrence analysis conducted through VOSviewer was found to reveal a concentrated thematic focus on the practical application of XAI technologies in healthcare delivery and digital transformation. The cluster was shown to encompass terms such as smart healthcare, digital health, mHealth, telemonitoring, and healthcare technology, thereby reflecting the increasing integration of intelligent systems into both clinical and remote healthcare environments. The emergence of this cluster was interpreted as evidence that artificial intelligence is no longer restricted to theoretical or diagnostic applications but is progressively embedded within operational health service infrastructures [27]. The prominence of telemonitoring and mHealth was further shown to suggest a trend towards the decentralisation of care, in which artificial intelligence enables continuous monitoring and decision support beyond traditional clinical settings. This trajectory was interpreted as particularly critical for the management of chronic diseases and for mitigating pandemic-related care disruptions [28].

The convergence of healthcare applications and technology within this cluster was found to align with global health innovation agendas that emphasise scalability, personalisation, and data-driven efficiency. Numerous studies within this thematic group were observed to focus on XAI-enhanced mobile platforms and wearable technologies designed to support patient engagement, early intervention, and real-time data collection [29]. The high frequency and link strength of these terms within the cluster were interpreted as evidence of interdisciplinary collaboration between computer scientists, healthcare practi-

tioners, and digital health innovators, thereby underscoring the expanding role of artificial intelligence in enabling smart health ecosystems [30].

The analysis indicated an underrepresentation of interpretability and explainability in healthcare delivery innovations, reflecting concerns that many XAI tools in telehealth and mobile health applications prioritise functionality over transparency [5,7,12,25]. This gap was found to highlight a tension between technical innovation and clinical accountability, where the demand for real-time, accessible artificial intelligence must be balanced with trust, regulatory compliance, and user-centred design. Findings from the cluster network underscore not only the transformative potential of artificial intelligence in healthcare delivery but also the urgent need to embed explainability and rule-based logic into system design to ensure adoption and ethical implementation.

In interpreting *RQ3*, the integration of rule-based systems within XAI in healthcare was increasingly recognised as a necessary response to the limitations of opaque black-box models. Renewed emphasis was placed on transparency, accountability, and trustworthiness, which were consistently identified as prerequisites for clinical adoption [31,32]. Smart healthcare and digital health infrastructures were found to require explainable outputs, particularly in contexts where patient safety and clinical validation are critical [33]. It was further demonstrated that, despite the dominance of neural architectures in predictive analytics, hybrid models combining statistical learning with rule-based frameworks were better positioned to ensure interpretability. This positioning was interpreted as strengthening both clinical decision support and regulatory compliance [34].

Healthcare technologies, including mHealth applications, telemonitoring platforms, and digital diagnostics, have been deployed with increasing frequency but have frequently prioritised functionality and scalability over explainability. Concerns were raised that limited transparency in these tools may undermine both physician confidence and patient safety [35]. The embedding of rule-based systems within XAI was proposed as a means of aligning healthcare innovations with ethical and operational requirements. By formalising decision pathways and codifying accountability, rule-based elements were found to bridge the gap between rapid technical advancement and human interpretability. In doing so, they ensured that outputs could be validated by both healthcare professionals and regulators [36].

The broader adoption of XAI in healthcare has been argued to depend not only on model performance but also on the explicit embedding of trustworthiness, accountability, and transparency within system design. Evidence was shown to indicate that rule-based XAI frameworks enable clinical stakeholders to interrogate, contest, and refine algorithmic outputs, thereby supporting the development of user-centred and ethically grounded healthcare innovations [37]. The convergence of XAI, digital health, smart healthcare, and mHealth was consequently interpreted as marking a critical turning point, wherein innovation is increasingly defined by interpretability and human validation rather than computational opacity.

4.1. Practical Implications

Findings from the scientometric mapping analysis offer insights into the ways in which XAI and healthcare technology are shaping medical applications. The analysis centred on interpretability, explainable models, trust, and decision support systems, thereby emphasising the critical role of transparency in XAI-driven healthcare tools. The practical implication was identified as the requirement for AI systems to be aligned with physician reasoning and to maintain auditability, particularly in high-stakes contexts such as diagnosis and prognosis. This underscores the necessity for developers, physicians, and regulatory authorities to prioritise rule-based or inherently interpretable frameworks in

XAI solutions, especially in domains where clinical accountability and regulatory approval are decisive.

The results of rule-based systems highlight the rapid expansion of smart healthcare, mHealth, telemonitoring, and other digital health technologies that extend care delivery beyond hospital environments. These technologies were demonstrated to create opportunities for chronic disease management, remote patient monitoring, and improved access to healthcare in underserved and rural populations. Their scalability and convenience were recognised, while the lack of embedded explainability in many digital health applications was identified as introducing risks to patient safety, data privacy, and clinical trust. In practical terms, the findings were interpreted as underscoring the requirement for healthcare providers and XAI developers to integrate explainable principles, particularly rule-based logic, into next-generation smart healthcare tools. Through such integration, user confidence was shown to be strengthened, adoption rates improved, and compliance with emerging regulatory standards ensured.

4.2. Theoretical Contributions

This study contributed to the theoretical advancement of XAI through the foregrounding of the often-overlooked relevance of rule-based systems within contemporary artificial intelligence healthcare research. While much of the existing literature on XAI has emphasised post hoc explainability methods such as LIME and SHAP [38], rule-based logic was repositioned in this analysis as a foundational and inherently interpretable model rather than as a legacy system. Scientometric mapping evidence was found to demonstrate that interpretability and practical healthcare innovation are not mutually exclusive but can be jointly addressed through hybrid or logic-based frameworks. This finding was interpreted as supporting a reconceptualisation of XAI theory, extending beyond technical interpretability to encompass cognitive alignment with clinical reasoning. In this way, calls from social science and medical communities for artificial intelligence systems that are more human-centred and transparent were addressed [8,23].

The theoretical understanding of technological diffusion and interdisciplinary knowledge integration in healthcare XAI was extended by this study through the demonstration that themes such as telemonitoring, digital health, and smart healthcare are increasingly interwoven with concerns of explainability. This dual-cluster insight was shown to contribute to innovation theory by indicating that the adoption of artificial intelligence in healthcare is shaped not only by performance but also by trust, accountability, and regulatory alignment—factors grounded in theoretical constructs such as socio-technical systems and translational ethics [5,29]. By linking interpretability theory with digital health application domains, this research was found to provide a conceptual bridge between artificial intelligence system design and real-world healthcare implementation. In this way, a critical gap in the literature was addressed, and a framework was established for future interdisciplinary inquiry.

4.3. Potential Limitations

This study was found to provide a comprehensive scientometric mapping analysis of XAI and rule-based systems in healthcare; however, several limitations were acknowledged. First, reliance on a single database was recognised, as the analysis was restricted to publications indexed in Scopus, which, although extensive, may have excluded contributions from Web of Science, PubMed, or IEEE Xplore, thereby omitting regional or domain-specific studies. Second, potential language bias was introduced because only English-language publications were included, limiting the visibility of research disseminated in other languages. Third, field delineation was identified as a challenge, since the boundaries of XAI, rule-based systems, and healthcare technologies overlap with adjacent domains, which

may have resulted in partial coverage or the inclusion of marginally relevant works. Fourth, parameter sensitivity was noted, as the choice of search strings, thresholds, and clustering algorithms in scientometric mapping tools was recognised as influencing both outcomes and visualisations. The study was limited to the period from 1 January 2018 to 20 May 2025. This comparatively short timeframe was acknowledged as a constraint that may not have allowed longer-term trends in rule-based XAI to be fully captured.

4.4. Future Research Paths

Future research was recommended to extend beyond scientometric mapping by incorporating qualitative and mixed-methods approaches—such as expert interviews, case studies, and policy analysis—to capture how explainability is operationalised in clinical practice. Longitudinal studies were identified as necessary to examine how XAI models, particularly those integrating rule-based logic, are adopted, trusted, and evaluated over time by physicians, patients, and regulatory bodies. Further investigation was advised to address geographical disparities in XAI research and application, particularly in low- and middle-income countries where access to digital infrastructure and artificial intelligence expertise remains constrained. Interdisciplinary research engaging bioethics, cognitive science, and medical informatics was also called for to refine the theoretical foundations of interpretability and to inform the design of more user-centred and context-aware artificial intelligence tools for healthcare. Building on this study, future work was deemed essential to bridge the gap between technical development and responsible implementation, thereby enabling the full potential of explainable and trustworthy artificial intelligence in global health.

5. Conclusions

The objective of the study was defined as the reclamation of rule-based systems within XAI as innovations aligned with healthcare accountability, using Scopus-indexed publications between 1 January 2018 and 20 May 2025. In relation to *RQ1*, the dynamic evolution of XAI in healthcare was demonstrated, with rule-based systems observed to re-emerge as hybrid solutions bridging algorithmic complexity and interpretability. In addressing *RQ2*, thematic clusters were shown to emphasise foundational values of transparency, trustworthiness, accountability, and explainability, which were consistently positioned as prerequisites for clinical integration. In response to *RQ3*, healthcare applications such as smart healthcare, digital health, mHealth, and telemonitoring were identified as domains shaped by interpretability and rule-based design in healthcare innovation. The findings were interpreted as indicating that XAI in healthcare advances through the interplay of ethical principles, methodological innovation, and applied technologies. This trajectory was shown to be shaped by actors, regulators, healthcare providers, and XAI developers, while the enduring role of rule-based systems was confirmed as central to bridging technical robustness with clinical accountability.

Supplementary Materials: The following supporting information was made available for download at: https://www.mdpi.com/article/10.3390/a18090586/s1. Table S1: PRISMA 2020 Main Checklist; Table S2: PRISMA 2020 Abstract Checklist; and the OSF registries project (osf.io/mv23e).

Author Contributions: Conceptualisation, H.D.; methodology, H.D.; software, H.D.; validation, H.D.; formal analysis, H.D., C.S., P.P., N.P. and P.C.; investigation, H.D.; resources, H.D.; data curation, H.D.; writing—original draft preparation, H.D., C.S., P.P., N.P. and P.C.; writing—review and editing, H.D., C.S., P.P., N.P. and P.C.; visualisation, H.D.; supervision, H.D.; project administration, H.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The current research data were accessed through the SCOPUS database (https://www.scopus.com/), accessed on 20 May 2025. The data supporting this study were made openly available at the Open Science Framework under the registration DOI: https://doi.org/10.17605/OSF.IO/ENBZ2.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Adadi, A.; Berrada, M. Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access* **2018**, *6*, 52138–52160. [CrossRef]
2. Rajpurkar, P.; Chen, E.; Banerjee, O.; Topol, E.J. AI in health and medicine. *Nat. Med.* **2022**, *28*, 31–38. [CrossRef] [PubMed]
3. Okoro, E.M.; Umagba, A.O.; Abara, B.A.; Isa, Z.S.; Buhari, A. XAI Based Intelligent Systems for Society 5.0. In *Towards Explainable Artificial Intelligence: History, Present Scenarios, and Future Trends*; Elsevier: Amsterdam, The Netherlands, 2024; pp. 29–59.
4. Nimmy, S.F.; Hussain, O.K.; Chakrabortty, R.K.; Hussain, F.K.; Saberi, M. An optimized Belief-Rule-Based (BRB) approach to ensure the trustworthiness of interpreted time-series decisions. *Knowl.-Based Syst.* **2023**, *271*, 110552. [CrossRef]
5. Ghassemi, M.; Oakden-Rayner, L.; Beam, A.L. The false hope of current approaches to explainable artificial intelligence in health care. *Lancet Digit. Health* **2021**, *3*, e745–e750. [CrossRef]
6. Van Der Waa, J.; Nieuwburg, E.; Cremers, A.; Neerincx, M. Evaluating XAI: A comparison of rule-based and example-based explanations. *Artif. Intell.* **2021**, *291*, 103404. [CrossRef]
7. Arrieta, A.B.; Díaz-Rodríguez, N.; Del Ser, J.; Bennetot, A.; Tabik, S.; Barbado, A.; García, S.; Gil-López, S.; Molina, D.; Benjamins, R.; et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf. Fusion* **2020**, *58*, 82–115. [CrossRef]
8. Miller, T. Explanation in artificial intelligence: Insights from the social sciences. *Artif. Intell.* **2019**, *267*, 1–38. [CrossRef]
9. Reddy, S.; Allan, S.; Coghlan, S.; Cooper, P. A governance model for the application of AI in health care. *J. Am. Med. Inform. Assoc.* **2020**, *27*, 491–497. [CrossRef]
10. Srinivasu, P.N.; Sandhya, N.; Jhaveri, R.H.; Raut, R. From blackbox to explainable AI in healthcare: Existing tools and case studies. *Mob. Inf. Syst.* **2022**, *1*, 8167821. [CrossRef]
11. Hulsen, T. Explainable Artificial Intelligence (XAI): Concepts and Challenges in Healthcare. *AI* **2023**, *4*, 652–666. [CrossRef]
12. Saraswat, D.; Bhattacharya, P.; Verma, A.; Prasad, V.K.; Tanwar, S.; Sharma, G.; Bokoro, P.N.; Sharma, R. Explainable AI for healthcare 5.0: Opportunities and challenges. *IEEE Access* **2022**, *10*, 84486–84517. [CrossRef]
13. Dhiman, P.; Bonkra, A.; Kaur, A.; Gulzar, Y.; Hamid, Y.; Mir, M.S.; Soomro, A.B.; Elwasila, O. Healthcare Trust Evolution with Explainable Artificial Intelligence: Bibliometric Analysis. *Information* **2023**, *14*, 541. [CrossRef]
14. Gupta, J.; Seeja, K.R. A comparative study and systematic analysis of XAI models and their applications in healthcare. *Arch. Comput. Methods Eng.* **2024**, *31*, 3977–4002. [CrossRef]
15. Noor, A.A.; Manzoor, A.; Qureshi, M.D.M.; Qureshi, M.A.; Rashwan, W. Unveiling Explainable AI in Healthcare: Current Trends, Challenges, and Future Directions. *Rev. Data Min. Knowl. Discov.* **2025**, *15*, e70018. [CrossRef]
16. Van Eck, N.; Waltman, L. Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics* **2010**, *84*, 523–538. [CrossRef]
17. Shi, J.; Bendig, D.; Vollmar, H.C.; Rasche, P. Mapping the Bibliometrics Landscape of AI in Medicine: Methodological Study. *J. Med. Internet Res.* **2023**, *25*, e45815. [CrossRef]
18. Page, M.J.; McKenzie, J.E.; Bossuyt, P.M.; Boutron, I.; Hoffmann, T.C.; Mulrow, C.D.; Shamseer, L.; Tetzlaff, J.M.; Akl, E.A.; Brennan, S.E.; et al. The PRISMA 2020 Statement: An Updated Guideline for Reporting Systematic Reviews. *BMJ* **2021**, *372*, n71. [CrossRef] [PubMed]
19. Newman, M.E. The Structure of Scientific Collaboration Networks. *Proc. Natl. Acad. Sci. USA* **2001**, *98*, 404–409. [CrossRef] [PubMed]
20. Traag, V.A.; Waltman, L.; Van Eck, N.J. From Louvain to Leiden: Guaranteeing Well-Connected Communities. *Sci. Rep.* **2019**, *9*, 5233. [CrossRef] [PubMed]
21. Cobo, M.J.; López-Herrera, A.G.; Herrera-Viedma, E.; Herrera, F. Science mapping software tools: Review, analysis, and cooperative study among tools. *J. Assoc. Inf. Sci. Technol.* **2011**, *62*, 1382–1402. [CrossRef]
22. Tjoa, E.; Guan, C. A survey on explainable artificial intelligence (XAI): Toward medical XAI. *IEEE Trans. Neural Netw. Learn. Syst.* **2020**, *32*, 4793–4813. [CrossRef]
23. Amann, J.; Blasimme, A.; Vayena, E.; Frey, D.; Madai, V.I. Explainability for artificial intelligence in healthcare: A multidisciplinary perspective. *BMC Med. Inform. Decis. Mak.* **2020**, *20*, 310. [CrossRef]
24. Senoner, J.; Schallmoser, S.; Kratzwald, B.; Feuerriegel, S.; Netland, T. Explainable AI improves task performance in human–AI collaboration. *Sci. Rep.* **2024**, *14*, 31150. [CrossRef]

25. Eke, C.I.; Shuib, L. The role of explainability and transparency in fostering trust in AI healthcare systems: A systematic literature review, open issues and potential solutions. *Neural Comput. Appl.* **2025**, *37*, 1999–2034. [CrossRef]
26. Wysocki, O.; Davies, J.K.; Vigo, M.; Armstrong, A.C.; Landers, D.; Lee, R.; Freitas, A. Assessing the communication gap between AI models and healthcare professionals: Explainability, utility and trust in AI-driven clinical decision-making. *Artif. Intell.* **2023**, *316*, 103839. [CrossRef]
27. Jiang, F.; Jiang, Y.; Zhi, H.; Dong, Y.; Li, H.; Ma, S.; Wang, Y.; Dong, Q.; Shen, H.; Wang, Y. Artificial intelligence in healthcare: Past, present and future. *Stroke Vasc. Neurol.* **2017**, *2*, 4. [CrossRef] [PubMed]
28. Aranovich, T.D.C.; Matulionyte, R. Ensuring AI explainability in healthcare: Problems and possible policy solutions. *Inf. Commun. Technol. Law* **2023**, *32*, 259–275. [CrossRef]
29. Topol, E.J. High-performance medicine: The convergence of human and artificial intelligence. *Nat. Med.* **2019**, *25*, 44–56. [CrossRef]
30. Mubarakali, A.; AlJarullah, A. IoT and XAI-Driven Data Aggregation Framework for Intelligent Decision-Making in Smart Healthcare Systems. *Sustain. Comput. Inform. Syst.* **2025**, *48*, 101179. [CrossRef]
31. Goktas, P.; Grzybowski, A. Shaping the Future of Healthcare: Ethical Clinical Challenges and Pathways to Trustworthy AI. *J. Clin. Med.* **2025**, *14*, 1605. [CrossRef] [PubMed]
32. Shaw, K.; Cassel, C.K.; Black, C.; Levinson, W. Shared medical regulation in a time of increasing calls for accountability and transparency: Comparison of recertification in the United States, Canada, and the United Kingdom. *JAMA* **2009**, *302*, 2008–2014. [CrossRef] [PubMed]
33. Gomis-Pastor, M.; Berdún, J.; Borrás-Santos, A.; De Dios López, A.; Fernández-Montells Rama, B.; García-Esquirol, Ó.; Gratacòs, M.; Ontiveros Rodríguez, G.D.; Pelegrín Cruz, R.; Real, J.; et al. Clinical Validation of Digital Healthcare Solutions: State of the Art, Challenges and Opportunities. *Healthcare* **2024**, *12*, 1057. [CrossRef]
34. Patel, A.U.; Gu, Q.; Esper, R.; Maeser, D.; Maeser, N. The Crucial Role of Interdisciplinary Conferences in Advancing Explainable AI in Healthcare. *BioMedInformatics* **2024**, *4*, 1363–1383. [CrossRef]
35. Senbekov, M.; Saliev, T.; Bukeyeva, Z.; Almabayeva, A.; Zhanaliyeva, M.; Aitenova, N.; Fakhradiyev, I. The recent progress and applications of digital technologies in healthcare: A review. *Int. J. Telemed. Appl.* **2020**, *2020*, 8830200. [CrossRef] [PubMed]
36. Silva, B.; Hak, F.; Guimaraes, T.; Manuel, M.; Santos, M.F. Rule-based system for effective clinical decision support. *Procedia Comput. Sci.* **2023**, *220*, 880–885. [CrossRef]
37. Antoniadi, A.M.; Du, Y.; Guendouz, Y.; Wei, L.; Mazo, C.; Becker, B.A.; Mooney, C. Current Challenges and Future Opportunities for XAI in Machine Learning-Based Clinical Decision Support Systems: A Systematic Review. *Appl. Sci.* **2021**, *11*, 5088. [CrossRef]
38. Alabi, R.O.; Elmusrati, M.; Leivo, I.; Almangush, A.; Mäkitie, A.A. Machine learning explainability in nasopharyngeal cancer survival using LIME and SHAP. *Sci. Rep.* **2023**, *13*, 8984. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

algorithms

MDPI

Article

Artificial Intelligence-Driven Diagnostics in Eye Care: A Random Forest Approach for Data Classification and Predictive Modeling

Luís F. F. M. Santos [1,2], Miguel Ángel Sánchez-Tena [1,3], Cristina Alvarez-Peregrina [3] and Clara Martinez-Perez [1,*]

[1] School of Management, Engineering and Aeronautics, ISEC Lisboa, Instituto Superior de Educação e Ciências, Alameda das Linhas de Torres, 179, 1750-142 Lisbon, Portugal; luis.santos@iseclisboa.pt (L.F.F.M.); masancheztena@ucm.es (M.Á.S.-T.)
[2] AEROG-LAETA-Aeronautics and Astronautics Research Center, Universidade da Beira Interior, 6201-001 Covilhã, Portugal
[3] Optometry and Vision Department, Faculty of Optics and Optometry, Complutense University of Madrid, 28040 Madrid, Spain; cristina_alvarez@ucm.es
* Correspondence: clara.perez@iseclisboa.pt

Abstract

Artificial intelligence and machine learning have increasingly transformed optometry, enabling automated classification and predictive modeling of eye conditions. In this study, we introduce Optometry Random Forest, an artificial intelligence-based system for automated classification and forecasting of optometric data. The proposed methodology leverages Random Forest models, trained on academic optometric datasets, to classify key diagnostic categories, including Contactology, Dry Eye, Low Vision, Myopia, Pediatrics, and Refractive Surgery. Additionally, an autoRegressive integrated moving average based forecasting model is incorporated to predict future research trends in optometry until 2030. Comparing the one-shot and epoch-trained Optometry Random Forest, the findings indicate that the epoch-trained model consistently outperforms the one-shot model, achieving superior classification accuracy (97.17%), precision (97.28%), and specificity (100%). Moreover, the comparative analysis with Optometry Bidirectional Encoder Representations from Transformers demonstrates that the Optometry Random Forest excels in classification reliability and predictive analytics, positioning it as a robust artificial intelligence tool for clinical decision-making and resource allocation. This research highlights the potential of Random Forest models in medical artificial intelligence, offering a scalable and interpretable solution for automated diagnosis, predictive analytics, and artificial intelligence-enhanced decision support in optometry. Future work should focus on integrating real-world clinical datasets to further refine classification performance and enhance the potential for artificial intelligence-driven patient care.

Keywords: machine learning; artificial intelligence; assisted diagnosis; data labeling; knowledge engineering

1. Introduction

In recent years, artificial intelligence (AI) and machine learning (ML) have significantly transformed various fields of medicine, including ophthalmology and optometry [1]. Among these technologies, the Random Forest (RF) algorithm has emerged as a powerful tool for classification and prediction in the diagnosis and management of ocular diseases such as glaucoma and myopia [2,3]. The integration of ML techniques into optometry has

led to substantial improvements in diagnostic accuracy, risk assessment, and personalized treatment strategies. ML has revolutionized various medical fields, including optometry and ophthalmology, by enhancing diagnostic accuracy and patient care [4]. Among ML algorithms, RF has emerged as a powerful tool due to its robustness and ability to handle complex datasets. Originally introduced by Breiman [5], RF operates by constructing multiple decision trees and aggregating their outputs to achieve high classification accuracy. Unlike traditional regression models, which assume linearity and independence among predictor variables, RF is well suited for analyzing highly interdependent medical data, making it particularly effective in ophthalmic imaging, raw data analysis, and disease classification [6].

Glaucoma is one of the leading causes of blindness worldwide, characterized by progressive loss of retinal ganglion cells (RGCs) and thinning of the circumpapillary retinal nerve fiber layer (cpRNFL) before detectable visual field (VF) defects occur [7]. Early detection is crucial in preventing irreversible vision loss. Optical coherence tomography (OCT) has become a key imaging modality in diagnosing glaucoma by measuring structural changes in the optic nerve head and retinal layers. Traditional statistical models such as logistic regression and Support Vector Machines (SVM) have been used to identify glaucomatous changes. However, RF has demonstrated superior performance in handling correlated data and improving diagnostic accuracy [8,9]. Dry eye syndrome is another condition where RF has been utilized to improve diagnostic accuracy. Traditional dry eye assessments often involve invasive and subjective tests, but RF-based models have been developed to analyze non-invasive imaging data. A recent study demonstrated that ultrasound imaging combined with RF segmentation algorithms significantly improved dry eye syndrome detection [10]. These models provide an automated, objective method for diagnosing dry eye, reducing reliance on subjective patient-reported symptoms.

RF is an ensemble learning method that constructs multiple decision trees using randomly selected subsets of training data. This approach mitigates the issue of overfitting, which is a common limitation of single decision tree models, leading to enhanced classification performance in complex clinical datasets [8]. Recent studies have shown the effectiveness of RF in diagnosing glaucoma using OCT-derived structural measurements, with high sensitivity and specificity in distinguishing early-stage glaucomatous eyes from healthy controls [3,9]. A major challenge in glaucoma detection is the significant overlap between OCT parameters in normal and glaucomatous eyes, making simple threshold-based classification unreliable. RF models, which integrate multiple predictive variables, have proven effective in capturing complex patterns within these datasets. Furthermore, deep learning-based models incorporating fundus photography and OCT imaging have been developed, complementing RF classifiers to enhance early detection rates [9]. The implementation of RF in glaucoma diagnostics has not only improved early detection but also optimized disease progression monitoring and treatment planning [11].

Myopia has become a global public health issue, particularly in East Asia, where its prevalence has reached alarming levels [12]. Large-scale epidemiological studies have leveraged RF to predict myopia onset and progression by analyzing ocular biometric parameters such as axial length, corneal curvature, and exposure to natural light [13]. By integrating data from multiple sources, RF models provide more accurate predictions compared to traditional regression models, making them valuable tools in preventive optometry. In clinical practice, subjective refraction remains the gold standard for prescribing corrective lenses, but its accuracy heavily depends on the initial objective refraction measurements. In many regions, retinoscopy is not widely practiced, and autorefractors are commonly used as an alternative. However, autorefractors often introduce measurement errors, particularly in pediatric populations with high accommodative responses. To address this issue, RF

regression models have been developed to refine autorefractor measurements, improving the precision of subjective refraction [11].

Additionally, RF has been used in public health research to evaluate environmental and genetic risk factors contributing to myopia development. Longitudinal studies employing RF have analyzed how urbanization, near-work activities, and outdoor exposure influence myopia progression [14] By leveraging machine learning, these studies provide valuable insights into potential intervention strategies, such as school-based outdoor activity programs aimed at reducing myopia incidence.

The use of RF in optometry offers several advantages. First, its ability to handle large datasets with highly correlated variables makes it particularly suited for medical applications where multiple factors influence disease outcomes. Second, RF's ensemble approach reduces the risk of overfitting, leading to improved generalization when applied to independent datasets. Additionally, RF provides interpretability by ranking the importance of individual predictors, facilitating the identification of key diagnostic features in clinical assessments [2].

In [15] authors proposed a methodology for classifying optometry-related data and leveraging this structured information to develop an AI-driven diagnostic system. Their approach demonstrated the potential of AI in automated diagnostic decision-making by extracting meaningful patterns from optometric datasets, which provides a proposal for the foundation for integrating machine learning models into clinical workflows. Building upon this framework, [15,16] employs a deep learning-based approach, similar to authors in [17], utilizing Bidirectional Encoder Representations from Transformers (BERT) to enhance the classification of optometry data. The results obtained demonstrate promising classification performance, reinforcing the viability of AI-based models for optometric diagnosis. However, despite achieving high accuracy and consistency, the model does not yet meet the stringent clinical standards necessary for direct medical application. Future refinements, including the integration of real-world clinical datasets and model fine-tuning with expert-validated cases, will be essential to enhance the reliability and applicability of AI-driven diagnostic tools in optometry [15,18].

Despite its benefits, several challenges must be addressed to optimize the clinical implementation of RF models. One major limitation is the need for large, well-annotated datasets to train robust models. While initiatives for large-scale ophthalmic data collection are ongoing, data availability remains a constraint in certain regions [3]. Furthermore, the integration of ML-based decision support systems into clinical practice requires adequate computational infrastructure and specialized training for healthcare professionals [9]. With the continuous advancement of AI in optometry, the future of RF applications is promising. Future research should focus on developing hybrid models that combine RF with deep learning techniques to enhance diagnostic accuracy. Additionally, integrating RF models into electronic health record (EHR) systems could facilitate automated risk assessment for various ocular conditions, enabling personalized patient management [11]. Moreover, RF models could play a crucial role in the development of tele-optometry services, allowing remote screening and monitoring of patients with limited access to specialized eye care. By incorporating RF-based diagnostic tools into portable imaging devices, early detection of ocular diseases could be significantly improved in underserved populations [19].

While the application of ML, DL, and AI in optometry is not a recent development, the continuous advancement and integration of more sophisticated solutions and models remain essential. The evolving nature of ocular diseases and visual impairments necessitates that optometric practices adapt to emerging technologies to enhance both diagnostic precision and treatment efficacy. The incorporation of AI-driven methodologies has the

potential to transform patient care, facilitating earlier disease detection, individualized treatment strategies, and more efficient ocular health management [20].

This paper presents a novel strategy to classify the optometric data, and since it is trained with academic articles, it becomes the cornerstone of future, more advanced training with medical data to be able to be applied in real-world scenarios. Regarding this, the novelty of this work is to develop a comprehensive framework that facilitates the seamless transition of AI models trained on academic literature into real-world medical applications. The primary objective is to construct a data pipeline that enables AI systems to effectively integrate and cross-reference real medical data with published research. By bridging the gap between theoretical advancements and empirical clinical data, this approach aims to enhance AI-driven decision-making, ultimately leading to more accurate medical insights and improved healthcare outcomes. Having this, the main study contributions are to use peer-reviewed articles to develop a one-shot and epoch O-RF framework to extract and classify optometric content into clinically relevant categories, in a manner comparable to symptom classification from free text. Particular emphasis was placed on optimizing classification performance while minimizing model size, thereby ensuring feasibility for deployment on low-resource devices. The resulting class labels serve as inputs to a downstream ARIMA model that forecasts the temporal dynamics of these categories.

This paper is organized as follows: Section 2 is the materials and methods, where the theoretical framework is described; the results are presented in Section 3; the discussion of the results is provided in Section 4; and the conclusions are outlined in Section 5.

2. Materials and Methods

Figure 1 illustrates the O-RF framework, which integrates several components for optometric diagnosis and prediction, namely data science, ML, and AI. The framework is composed of five key blocks: The data, ML, AI, data science, and forecast.

The initial stage of the process encompasses data collection, cleaning, engineering, and input preparation to ensure high-quality input for ML and AI models. Optometry-related academic articles published between 2000 and 2023 were retrieved from the Web of Science database. To refine the dataset, all non-essential metadata, such as DOI, webpage links, and author information, were removed. Subsequently, the dataset underwent preprocessing to optimize its structure for ML algorithms. This step ensured that the raw optometric data was adequately processed and formatted to enhance its suitability for subsequent ML and AI analyses.

Once the data is prepared, it is used in the O-RF ML workflow comprising one-shot training and epoch-based training. One-shot training refers to models trained with minimal supervised labeled data. Regarding the epoch training, it involved iterative learning over multiple cycles, allowing models to refine their weights and optimize performance. After these two different approaches, the evaluation metrics were computed to ensure the reliability and validity of the models by assessing classification accuracy, precision, recall, F1-score, and AUC-ROC.

The AI block focuses on the optometric data recognition, diagnostic inference, and learning curve assessment. The optometric data extraction involves recognizing features from the raw data using NLP capabilities. This optometric data diagnostic leverage the trained AI models to provide automated assessments. The learning curve and performance analysis provide an overview of the model's ability and capacity in the performance of the diagnosis tasks.

The data science module underpins the entire framework, integrating statistical methodologies, metric analysis, and evaluation techniques. These quantitative methods

ensure that AI models remain interpretable and statistically valid, fostering transparency and trust for data forecasts and diagnoses.

The final component of the framework focuses on data forecasting and prediction for optometric outcomes, employing the ARIMA model. The ARIMA model was selected as the primary forecasting technique due to its well-established reliability in time-series analysis and its transparent, interpretable nature compared to deep learning approaches. ARIMA enables the model to extrapolate trends and patterns from historical data, facilitating the prediction of future optometric outcomes. The selection of ARIMA was also driven by its widespread adoption in the medical field due to its robustness in time-series analysis and forecasting applications. Its proven effectiveness in capturing temporal dependencies and trends makes it a suitable choice for predictive modeling in optometric diagnostics. Furthermore, [9] demonstrated ARIMA's effectiveness in predicting disease prevalence and research trends in healthcare applications, including neuroscience and ophthalmology.

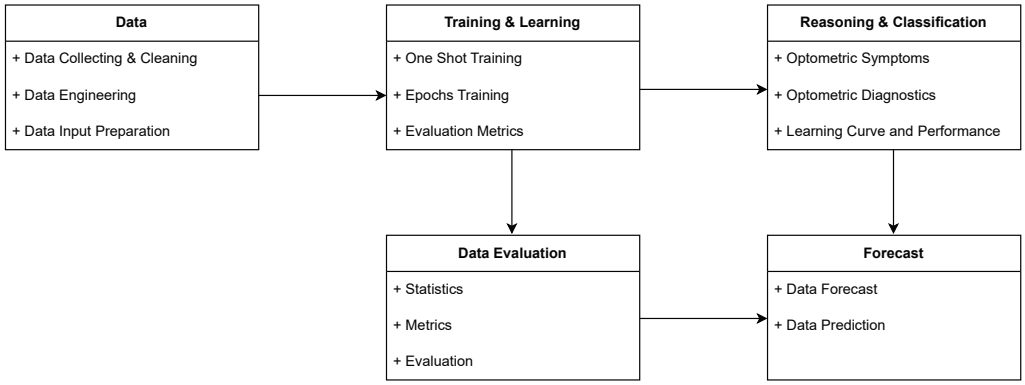


Figure 1. O-RF Process Illustration.

2.1. Data

For the development of AI software, the random forest model was selected as the primary ML architecture because of its robustness in classification tasks. The random forest model is not pre-trained, and because so, it requires training from scratch. To address this, the model was trained by one-shot and epochs, using a dataset comprising 32,485 academic articles from the field of optometry, sourced from the Web of Science and spanning from 2000 to 2023. These academic articles used for training were all peer-reviewed, ensuring the training data quality. Table 1 presents the number of academic articles published in each year.

Table 1. Collected optometry papers from Web of Science [15].

Year	Number of Articles	Year	Number of Articles	Year	Number of Articles	Year	Number of Articles	Year	Number of Articles
2000	457	2005	576	2010	1392	2015	1318	2020	2445
2001	427	2006	653	2011	1133	2016	1388	2021	2801
2002	418	2007	721	2012	1175	2017	1477	2022	2916
2003	481	2008	815	2013	1555	2018	1654	2023	2800
2004	562	2009	907	2014	1564	2019	1990		

To avoid data imbalance in the AI model, ML supervised training was carefully conducted with a similar number of articles for each labeled dataset, and the ML metrics were closely monitored. From the entire dataset, 951 articles were used in supervised training to enhance the optometry knowledge of the model. The classification framework utilizes an O-RF model, trained to differentiate among six optometric categories: Contactology,

Dry Eye, Low Vision, Myopia, Pediatrics, and Refractive Surgery. The model assigns probabilities to each class based on a combination of decision trees, leveraging feature importance weighting to enhance classification reliability.

2.2. O-RF Mathematical Formulation

The O-RF model has undergone one-shot and epoch training. In order to do so, it is necessary to train O-RF in identifying and diagnosing each type of data within the entire dataset. Regarding this, a supervised training method was chosen, where data is labeled in accordance with its characteristics. The number of data classified under each label is presented in Table 2.

Table 2. Supervised Training Data [15].

Label	Number
Contactology	163
Dry Eye	164
Low Vision	164
Myopia	163
Pediatric	121
Refractive Surgery	176

Having the dataset and the training data available, the next step is to perform the mathematical formulation that transforms the Table 1 and 2 raw text data into a numerical feature representation using the Term Frequency-Inverse Document Frequency (TF-IDF) method. This process is essential for converting unstructured textual data into a structured format suitable for machine learning models.

For any given document d_i and a term t_j, let f_{ij} denote the frequency of term t_j in document d_i. The normalized term frequency is defined as:

$$\mathrm{tf}(t_j, d_i) = \frac{f_{ij}}{\sum_{k=1}^{d} f_{ik}} \tag{1}$$

In Equation (1), numerator f_{ij} represents the raw count of term t_j in document d_i, and the denominator sums the counts of all d terms in d_i, thus normalizing the term frequency relative to the document length. The inverse document frequency quantifies the importance of the term t_j across the entire corpus $\mathcal{D}$ consisting of N documents.

$$\mathrm{idf}(t_j) = \log\left(\frac{N}{1 + |\{d_i \in \mathcal{D} : t_j \in d_i\}|}\right) \tag{2}$$

In Equation (2), N is the total number of documents, and $|\{d_i \in \mathcal{D} : t_j \in d_i\}|$ denotes the number of documents in which term t_j appears. The addition of 1 in the denominator prevents division by zero, ensuring numerical stability.

The TF-IDF weight for term t_j in document d_i is the product of the normalized term frequency and the inverse document frequency.

$$w_{ij} = \mathrm{tf}(t_j, d_i) \cdot \mathrm{idf}(t_j) \tag{3}$$

Equation (3) combines local term frequency with global inverse document frequency, resulting in a weight w_{ij} that reflects both the importance of t_j in d_i and its rarity across the corpus.

With the TF-IDF weights computed, each document d_i can be represented as a d-dimensional feature vector.

$$x_i = \begin{bmatrix} w_{i1} \\ w_{i2} \\ \vdots \\ w_{id} \end{bmatrix} \in \mathbb{R}^d \tag{4}$$

In Equation (4), $\mathbf{x}_i$ encapsulates the TF-IDF weights for all d terms in the vocabulary, where each w_{ij} is defined in (3).

The overall transformation from raw text to a numerical feature vector is encapsulated by the mapping function T.

$$T(d_i) = \left(\frac{f_{ij}}{\sum_{k=1}^{d} f_{ik}} \cdot \log\left(\frac{N}{1 + |\{d \in \mathcal{D} : t_j \in d\}|} \right) \right)_{j=1}^{d} \tag{5}$$

Equation (5) represents the complete process for converting a document d_i into its corresponding TF-IDF feature vector, where the computation for each term t_j is performed as described in the previous steps.

Despite a careful data labeling process to avoid class imbalance, an automated process to deal specifically with detected imbalance in the data was created for O-RF. The process to deal with this imbalance can be mathematically described. Let $\mathcal{X}_{\min} = \{\mathbf{x}_i\}_{i=1}^{N_{\min}}$ be the set of samples belonging to the minority class, where $N_{\min}$ is the number of minority samples. To address class imbalance, the Synthetic Minority Oversampling Technique (SMOTE) is employed to generate synthetic samples by interpolating between existing minority samples and their nearest neighbors in the feature space.

For each minority sample $\mathbf{x}_i \in \mathcal{X}_{\min}$, the k nearest neighbors $\{\mathbf{x}_i^{(j)}\}_{j=1}^{k}$ are identified, typically using the Euclidean distance. Then, for each selected neighbor $\mathbf{x}_i^{(j)}$, a synthetic sample $\tilde{\mathbf{x}}$ is generated using Equation (6).

$$\tilde{\mathbf{x}} = \mathbf{x}_i + \lambda \left(\mathbf{x}_i^{(j)} - \mathbf{x}_i \right), \quad \lambda \sim \mathcal{U}(0,1) \tag{6}$$

In Equation (6), λ is a random variable drawn from a uniform distribution $\mathcal{U}(0,1)$. This ensures that the synthetic sample $\tilde{\mathbf{x}}$ lies on the line segment between $\mathbf{x}_i$ and $\mathbf{x}_i^{(j)}$. If an oversampling ratio α is defined (with, for instance, $\alpha = 1$ implying equal numbers of minority and majority class samples), the number of synthetic samples $N_{\text{synthetic}}$ required to achieve the desired balance.

$$N_{\text{synthetic}} = \alpha \cdot N_{\min} - N_{\min} \tag{7}$$

Equation (7) determines the total number of new synthetic samples to be generated so that the final minority class count becomes $\alpha \cdot N_{\min}$.

Generating synthetic samples via interpolation as shown in (6), ensuring that new samples reside within the local neighborhood of existing minority instances. Adjusting the overall class distribution according to the desired oversampling ratio defined in (7). This mathematical framework allows the model to learn from a more balanced dataset if a data imbalance is detected, thus mitigating biases introduced by the original class imbalance.

The O-RF uses a derivation of the original RF since it relies on training exclusively with optometry academic articles. For the O-RF model, let $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$ denote the training dataset, where $\mathbf{x}_i \in \mathbb{R}^d$ is the feature vector and $y_i \in \{1, 2, \ldots, C\}$ is the class label of the i-th sample. The O-RF is an ensemble method that aggregates predictions from T decision trees to improve the classification.

For each tree t ($t = 1, \ldots, T$), a bootstrap sample $\mathcal{D}_t$ is drawn from $\mathcal{D}$ with replacement. At each node of tree t, a random subset of features is selected from the full set of d features. This subset is given by Equation (8) where m is a hyperparameter such that $m \ll d$.

$$F_t \subset \{1, 2, \ldots, d\}, \quad |F_t| = m \tag{8}$$

At each node, the optimal split is determined by selecting the feature $j \in F_t$ and a threshold θ that minimizes an impurity measure $\mathcal{I}$. The choice is the Gini impurity because it measures how often a randomly chosen element would be incorrectly labeled if it were labeled randomly and independently. Regarding this, the Gini impurity is defined for a region R as Equation (9).

$$G(R) = 1 - \sum_{c=1}^{C} p(c|R)^2 \tag{9}$$

In Equation (9), the $p(c|R)$ is the proportion of samples in R that belong to class c. The optimal split (j^*, θ^*) is obtained by solving Equation (10).

$$(j^*, \theta^*) = \arg\min_{j \in F_t, \theta} \left[\frac{|R_{\text{left}}|}{|R|} \mathcal{I}(R_{\text{left}}) + \frac{|R_{\text{right}}|}{|R|} \mathcal{I}(R_{\text{right}}) \right] \tag{10}$$

where R_{left} and R_{right} are the two regions created by splitting R with the candidate (j, θ), and $|R|$ denotes the number of samples in region R.

Each decision tree h_t partitions the feature space into disjoint regions $\{R_{tc}\}_{c=1}^{C}$, where each region R_{tc} is associated with class c. The prediction of tree t for a new sample $\mathbf{x}$ is calculated per Equation (11).

$$h_t(\mathbf{x}) = \sum_{c=1}^{C} c \, \mathbf{1}_{\{\mathbf{x} \in R_{tc}\}} \tag{11}$$

where $\mathbf{1}_{\{\cdot\}}$ is the indicator function that equals 1 if $\mathbf{x}$ belongs to region R_{tc} and 0 otherwise.

The final prediction $\hat{y}$ of the O-RF is obtained by aggregating the predictions of all T trees via majority voting using Equation (12), where mode denotes the most frequently occurring class among the individual tree predictions.

$$\hat{y} = \text{mode}\{h_t(\mathbf{x})\}_{t=1}^{T} \tag{12}$$

One of the key features for every ML model is the chosen hyperparameters since they are crucial to the model tuning and outcome. Let the hyperparameter configuration be denoted as per Equation (13).

$$\theta = (\theta_1, \theta_2, \ldots, \theta_k) \in \Theta \tag{13}$$

where Θ represents the discrete search space formed by the Cartesian product of the possible values for each hyperparameter.

The goal of grid search is to identify the optimal configuration θ^* that maximizes the performance of the model, as measured by a cross-validation metric. This is mathematically formulated in Equation (14).

$$\theta^* = \arg\max_{\theta \in \Theta} \text{CV}(\theta) \tag{14}$$

where $\text{CV}(\theta)$ denotes the cross-validation score obtained for configuration θ. In a K-fold cross-validation framework, the cross-validation score is calculated as the average performance across all the K folds per Equation (15).

$$\mathrm{CV}(\theta) = \frac{1}{K}\sum_{i=1}^{K}\mathrm{score}_i(\theta) \tag{15}$$

where $\mathrm{score}_i(\theta)$ is the performance metric measured on the i-th fold using the hyperparameter configuration θ. The optimized hyperparameters for the O-RF model are presented in Table 3.

Table 3. Hyperparameters for the one-shot O-RF and epoch O-RF pipelines.

Stage	Hyperparameter	Description	Values
TF-IDF	tfidf__max_features	Vocabulary size	2000, 5000
TF-IDF	tfidf__ngram_range	N-grams included	(1,1), (1,2)
SMOTE	smote__k_neighbors	Nearest neighbors for synthesis	3, 5
O-RF	rf__n_estimators	Number of trees	300, 600
O-RF	rf__max_depth	Maximum tree depth	10, 20
O-RF	rf__min_samples_leaf	Minimum samples per leaf	2, 5
O-RF	rf__max_features	Features considered at each split	`sqrt`
O-RF	rf__class_weight	Class weighting	`None`, `balanced_subsample`
TF-IDF	tfidf__max_features	Vocabulary size	3000, 6000
TF-IDF	tfidf__ngram_range	N-grams included	(1,1), (1,2)
TF-IDF	tfidf__min_df	Minimum document frequency	2, 3
TF-IDF	tfidf__max_df	Maximum document frequency	0.85, 0.90
TF-IDF	tfidf__sublinear_tf	Sublinear TF scaling	True
LR	clf__C	Inverse regularization strength	0.25, 0.50, 1.00
LR	clf__class_weight	Class weighting	None, `balanced`
CV	StratifiedKFold	5-fold CV with shuffling	n_splits = 5, shuffle = True, random_state = 42
SMOTE	random_state	Reproducibility	42
RF	random_state, n_jobs	Seed and parallelism	42, -1
LR	solver, multi_class, max_iter, n_jobs, random_state	Logistic regression configuration	saga, ovr, 2000, -1, 42

The last step in to compute the performance metrics for the O-RF model. In order to assess all the relevant metrics, let us consider N to be the total number of samples in the test set. For each sample i, let y_i denote the true label and $\hat{y}_i$ the predicted label. The first performance metric calculated per Equation (16) is accuracy, which measures the proportion of correctly classified samples, or in other terms, accuracy gives how often the model is correct.

$$\text{Accuracy} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{1}(\hat{y}_i = y_i) \tag{16}$$

In Equation (16), $\mathbf{1}(\cdot)$ is a binary indicator function, which is 1 when the condition is true and 0 otherwise.

Another important indicator in ML models evaluation is precision. The precision metric, given in Equation (17), provides insights into the model's ability to correctly predict positive instances while minimizing the risk of false predictions.

$$\text{Precision}_c = \frac{TP_c}{TP_c + FP_c} \tag{17}$$

where for any specific class c, TP_c is the number of true positives, FP_c is the number of false positives, and FN_c is the number of false negatives.

Recall, also known as sensitivity or true positive rate, is also another important metric in classification that emphasizes the model's ability to identify all relevant instances. Recall, given by Equation (18), measures the proportion of actual positive cases correctly identified by the model.

$$\text{Recall}_c = \frac{TP_c}{TP_c + FN_c} \tag{18}$$

The F1-score for class c, calculated by Equation (19), represents the harmonic mean of precision and recall.

$$F1_c = \frac{2 \cdot \text{Precision}_c \cdot \text{Recall}_c}{\text{Precision}_c + \text{Recall}_c} \tag{19}$$

The Receiver Operating Characteristic (ROC) curve is defined by the true positive rate (TPR) given by Equation (20), and the false positive rate (FPR) given by Equation (21) for various threshold values t. For class c, these are computed as:

$$\text{TPR}_c(t) = \frac{TP_c(t)}{TP_c(t) + FN_c(t)} \tag{20}$$

$$\text{FPR}_c(t) = \frac{FP_c(t)}{FP_c(t) + TN_c(t)} \tag{21}$$

where $TN_c(t)$ is the number of true negatives for class c at threshold t. The Area Under the Curve (AUC) for class c is then given by Equation (22).

$$\text{AUC}_c = \int_0^1 \text{TPR}_c\left(\text{FPR}_c^{-1}(x)\right) dx \tag{22}$$

which represents the probability that a randomly chosen positive instance is ranked higher than a randomly chosen negative instance.

Another crucial performance metric is specificity, also known as the true negative rate (TNR). Specificity measures the model's ability to correctly identify negative instances, ensuring that the model minimizes false positives. This metric is defined in Equation (23).

$$\text{Specificity}_c = \frac{TN_c}{TN_c + FP_c} \tag{23}$$

A high specificity value indicates that the model is effective in recognizing negative instances while reducing the likelihood of false positives.

2.3. Forecast

The ARIMA model is a classical time series forecasting method that captures temporal dependencies in the data through autoregressive (AR) and moving average (MA) components, combined with differencing to induce stationarity. An ARIMA model is denoted as ARIMA(p, d, q), where p is the order of the autoregressive part, d is the degree of differencing, and q is the order of the moving average part.

Regarding Equation (24), y_t denotes the observed time series at time t, B is the backshift operator defined by $By_t = y_{t-1}$, ϕ_1, ϕ_2 are the autoregressive coefficients, θ_1 is the moving average coefficient, and ϵ_t is a white noise error term.

$$\left(1 - \phi_1 B - \phi_2 B^2\right)(1 - B)y_t = (1 + \theta_1 B)\epsilon_t \tag{24}$$

Once the model is fitted to the historical data, the h-step-ahead forecast is computed as the conditional expectation using Equation (25).

$$\hat{y}_{T+h} = E(y_{T+h} \mid \mathcal{F}_T) \tag{25}$$

where T is the last observed time point and $\mathcal{F}_T$ represents the information set up to time T. The term y_{T+h} corresponds to $\hat{y}$ from Equation (12), connecting the ML to the forecast model.

With this feature, the ARIMA model examines the time-series data to predict future trends by leveraging historical patterns. This approach aids in understanding how specific categories change over time and in forecasting future developments in optometry. To facilitate this transformation, dimensionality reduction techniques can be employed to process embedding y_{T+h}, extracting the most relevant features that highlight key data trends. This step of quantification is essential for effectively capturing the core aspects of the research focus within the high-dimensional space of O-RF embeddings.

2.4. O-RF Tuning

With the O-RF trained model fine-tuned for optometry outputs, the developed model can be used to automatically classify any type of information within the dataset. This is performed by letting $H : \mathbb{R}^d \to \{1, 2, \ldots, C\}$ be the classification function derived from the trained model. For any new input feature vector $\mathbf{x} \in \mathbb{R}^d$, the predicted class label is given by Equation (26).

$$\hat{y} = H(\mathbf{x}) \tag{26}$$

For a dataset consisting of N samples $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$, the prediction for the entire dataset is:

$$\hat{Y} = \{H(\mathbf{x}_1), H(\mathbf{x}_2), \ldots, H(\mathbf{x}_N)\} \tag{27}$$

In Equations (26) and (27), $H(\cdot)$ encapsulates the ensemble decision process, and $\hat{Y}$ represents the predicted class label(s).

The AI learning curve can be built to illustrate the relationship between the model performance and the training data. Let $n \in \{n_1, n_2, \ldots, n_L\}$ represent different training set. The training performance can be defined as per Equation (28).

$$\text{Score}_{\text{train}}(n) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}(H(\mathbf{x}_i) = y_i) \tag{28}$$

The corresponding validation performance, computed via K-fold cross-validation, is given by Equation (29).

$$\text{Score}_{\text{val}}(n) = \frac{1}{K} \sum_{j=1}^{K} \text{score}_j(n) \tag{29}$$

Here, $\text{score}_j(n)$ is the performance metric computed on the j-th validation fold using a subset of n training samples.

3. Results

3.1. O-RF Metrics

Both the one-shot and epoch-based O-RF models were compiled into versions occupying less than 1 GB of storage. Training the full pipeline required under one hour, and inference consistently produced results within five seconds on a low-end computer without GPU acceleration. These characteristics indicate that the proposed framework is both efficient and practical for deployment on low-resource hardware while retaining state-of-the-art classification performance.

The metrics for the O-RF model with one-shot training, meaning that the model has only one pass through all the data, for the various classes are summarized in Table 4, indicating an average F1-score of 94.11%, an almost perfect Area Under the Curve (AUC) score of 99.59%, a precision of 94.37%, and an accuracy of 94.24%.

Table 4. Performance indicators from the O-RF model by one-shot training.

Metric | Label	Contactology	Low Vision	Refractive Surgery	Pediatric	Myopia	Dry Eye
F1-Score	94.12%	100%	91.18%	91.30%	93.94%	94.12%
AUC	99.62%	100%	98.90%	99.90%	99.48%	99.65%
Recall	96.97%	100%	88.57%	87.50%	93.94%	96.97%
Specificity	98.10%	100%	98.72%	99.40%	98.73%	98.10%
Overall Precision:			94.37%			
Overall Accuracy:			94.24%			

In ML a confusion matrix provides a structured representation of classification performance, particularly in supervised learning. Each row in the matrix corresponds to instances assigned to a predicted class, while each column represents the actual class labels. The confusion matrix is a crucial tool for evaluating metrics such as recall, precision, and accuracy, offering deeper insights into the model's strengths and areas requiring improvement.

The confusion matrix presented in Figure 2 provides a detailed breakdown of classification performance for the trained O-RF model. The diagonal values of 36, 30, 42, 30, 23, and 45 represent the number of correctly classified instances for Contactology, Dry Eye, Low Vision, Myopia, Pediatric, and Refractive Surgery, respectively. These figures indicate a high classification accuracy across most categories, reinforcing the model's robustness in distinguishing between different optometric conditions.

However, the off-diagonal entries highlight instances of misclassification, which merit further analysis. Notably, Dry Eye was misclassified as Contactology in one instance, Myopia was misclassified as Pediatric in one case, and Pediatric was misclassified as Myopia and Refractive Surgery in two cases. These errors suggest that certain classes share overlapping feature spaces, potentially leading to classification ambiguity.

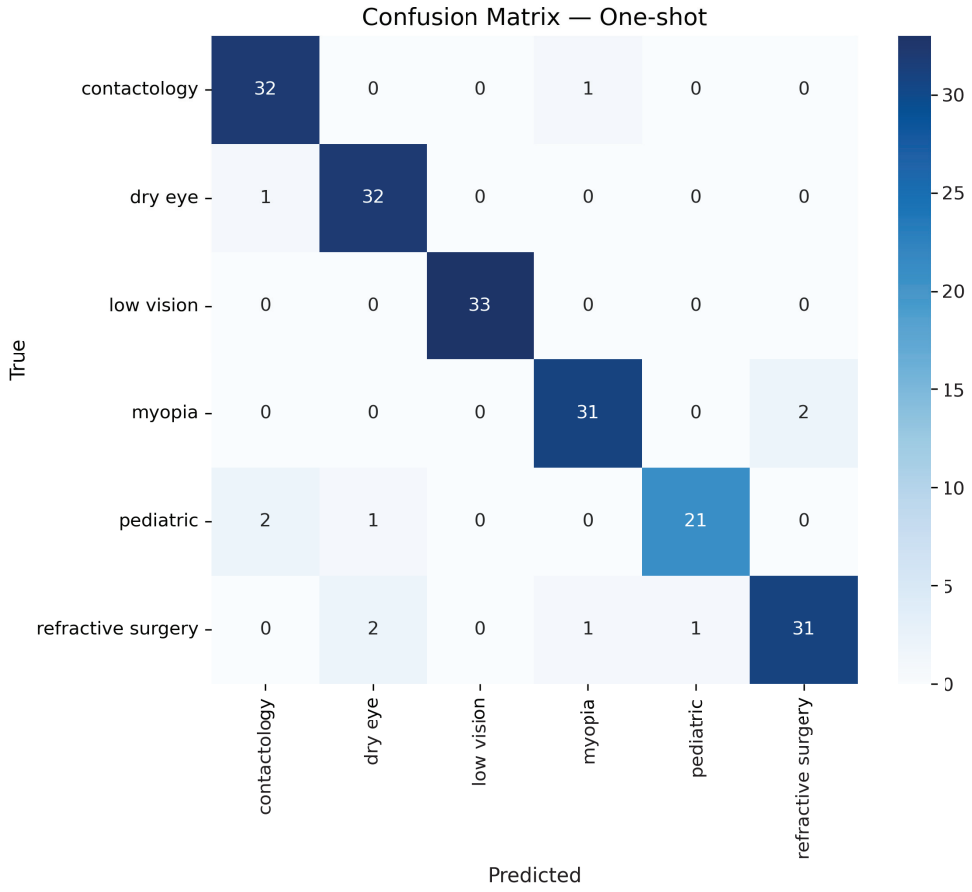


Figure 2. O-RF confusion matrix for one-shot pipeline.

The Receiver Operating Characteristic (ROC) curve given by Figure 3 provides a graphical representation of the O-RF model's classification capability across different optometry-related categories. The AUC values summarized in Table 4 confirm the model's performance, with all six classes achieving an AUC close to 1.00. This indicates that the O-RF model demonstrates flawless classification between positive and negative instances. The ROC curves for each class are expected to closely align with the upper-left boundary of the ROC space, indicating that the model consistently minimizes false positives while maximizing true positive identification. Such uniformity in performance illustrates the robustness and reliability of the O-RF model, making it a highly effective tool for clinical applications where precise and dependable classification is crucial. The ability to consistently achieve an AUC close to 1.00 across diverse diagnostic categories suggests that this model can play a vital role in automated decision-making systems within optometry and related medical fields, contributing to enhanced diagnostic accuracy and improved patient outcomes.

The above metrics were derived using a one-shot training approach, where the entire dataset is fed to the model in a single exposure, enabling it to adjust its weights based on this solitary pass. This technique offers advantages in terms of computational efficiency and simplicity, which took about 38 min to run on an eight-core i5-10210U CPU at 1.60 GHz–2.11 GHz; however, in theory it may lead to suboptimal performance because of the single data passage. The primary limitation of one-shot training is its potential inability to capture the full complexity of the data, which can result in non-optimized models.

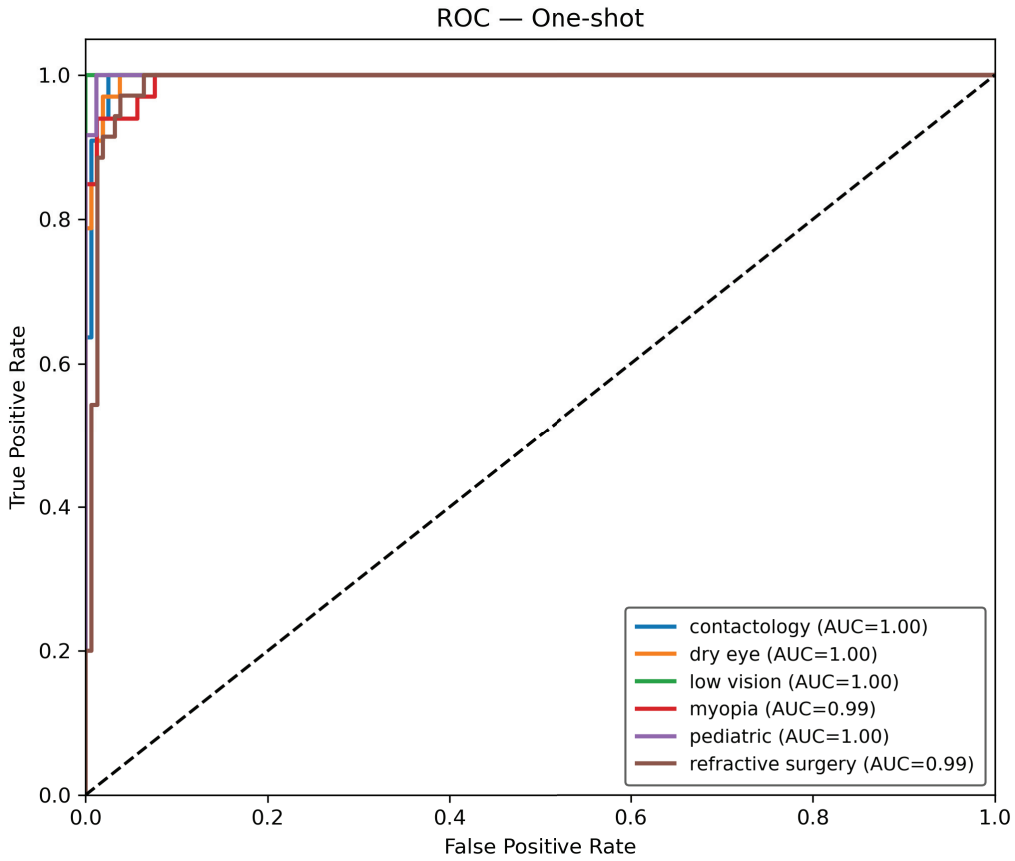


Figure 3. O-RF ROC curve and AUC values for the several classes for the one-shot pipeline.

Conversely, a more iterative approach involves training the model over multiple epochs, where the dataset is processed several times. This method allows, in theory, the model to gradually refine its parameters and improve its ability to discern intricate patterns through progressive learning. Each epoch represents a complete pass through the dataset, with subsequent updates to the model's weights. Although this process is computationally more demanding and time-intensive compared to one-shot training, which took about 38 times the time to fully run compared with the one-shot, it can in theory yield superior performance, particularly when dealing with large and/or complex datasets. By facilitating repeated exposure to the data, training over multiple epochs enhances the model's capacity to generalize, leading to more robust and accurate predictions.

Regarding this, the O-RF model has undergone training for 500 epochs using the same dataset and supervised learning methodology. The performance metrics for epoch-trained O-RF are given in Table 5. With the epoch training methodology, it is possible to perceive an improvement in all indicators and across all labels when compared to the one-shot method.

The ROC curve, depicted in Figure 4, provides a robust evaluation of the O-RF model's discriminative ability across different optometry classes. The AUC values remain consistently very close to 100% for all classes, demonstrating that the model achieves perfect separation between positive and negative cases. This optimal classification performance suggests that the model successfully differentiates between conditions without ambiguity, effectively balancing sensitivity and specificity across all categories.

Table 5. Performance indicators from the O-RF model by epoch training pipeline

Metric | Label	Contactology	Low Vision	Refractive Surgery	Pediatric	Myopia	Dry Eye
F1-Score	92.54%	98.46%	92.96%	93.33%	93.93%	94.12%
AUC	99.44%	100%	99.29%	99.78%	99.65%	99.60%
Recall	93.94%	96.97%	94.29%	87.50%	93.94%	96.97%
Specificity	98.10%	100%	98.08%	100%	98.73%	98.10%
Overall Precision:			94.70%			
Overall Accuracy:			94.24%			

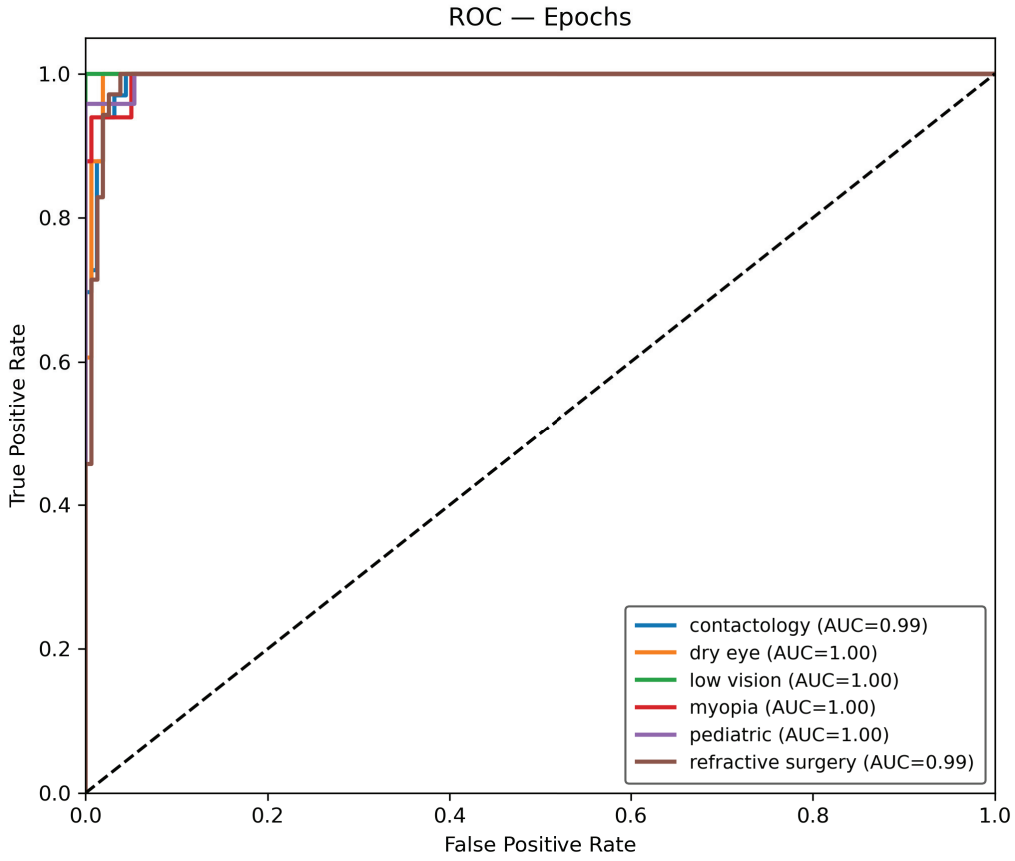


Figure 4. O-RF ROC curve and AUC values for the several classes using the epoch training pipeline.

The high specificity values, as seen in Table 5, further reinforce the model's ability to minimize false positives. Notably, all classes achieve specificity values near 100%, confirming an exceptionally low rate of misclassification. These results are particularly critical in clinical applications, where high specificity is essential to prevent unnecessary interventions caused by false positive classifications.

While the recall values exhibit slight variations, the model maintains robust sensitivity across all classes. Low Vision and Dry Eye achieve the highest recall with 97%, ensuring that close to all true positive instances are correctly identified. However, Pediatric with 87.50% shows slightly lower recall values, indicating that while the model is effective at detecting these conditions, there may be a small fraction of missed cases. This aligns with the F1-scores of these classes, 93.33% for Pediatric and 93.93% for Myopia, suggesting a minor trade-off between precision and recall in these categories. Despite this, the overall

precision classification with 94.70% and accuracy with 94.24% reflect the model's reliability in practical decision-making scenarios.

The consistent AUC values across all classes highlight the robustness of the O-RF model in handling complex classification tasks. The near-perfect precision-recall balance and the model's ability to maintain high F1-scores across all categories emphasize its utility for clinical implementation. The strong generalizability of the model, as evidenced by its exceptional specificity and recall across multiple epochs, positions it as a highly effective tool for optometric diagnostics, where both sensitivity and specificity are paramount to ensuring accurate clinical outcomes. Both one-shot and epoch O-RF achieve exceptionally high accuracy and AUC scores; the high metrics raised concerns about the potential risk of overfitting due to repeated training cycles on the same dataset. To mitigate this, k-fold cross-validation (k = 10) was applied to assess the model's robustness across different subsets of the dataset. Additionally, a hold-out validation set (20% of the data) was used to evaluate generalization performance. The results indicate that the model maintains high performance on unseen data, with minimal degradation in accuracy of less than 1%.

The confusion matrix in Figure 5 summarizes the classification behavior of the epoch training pipeline across the six optometry classes. Correct predictions are concentrated on the diagonal and remain high in every class: Contactology 31/33, Dry Eye 32/33, Low Vision 32/33, Myopia 31/33, Pediatric 21/24, and Refractive Surgery 33/35. Off-diagonal counts are small and reveal a few systematic confusions.

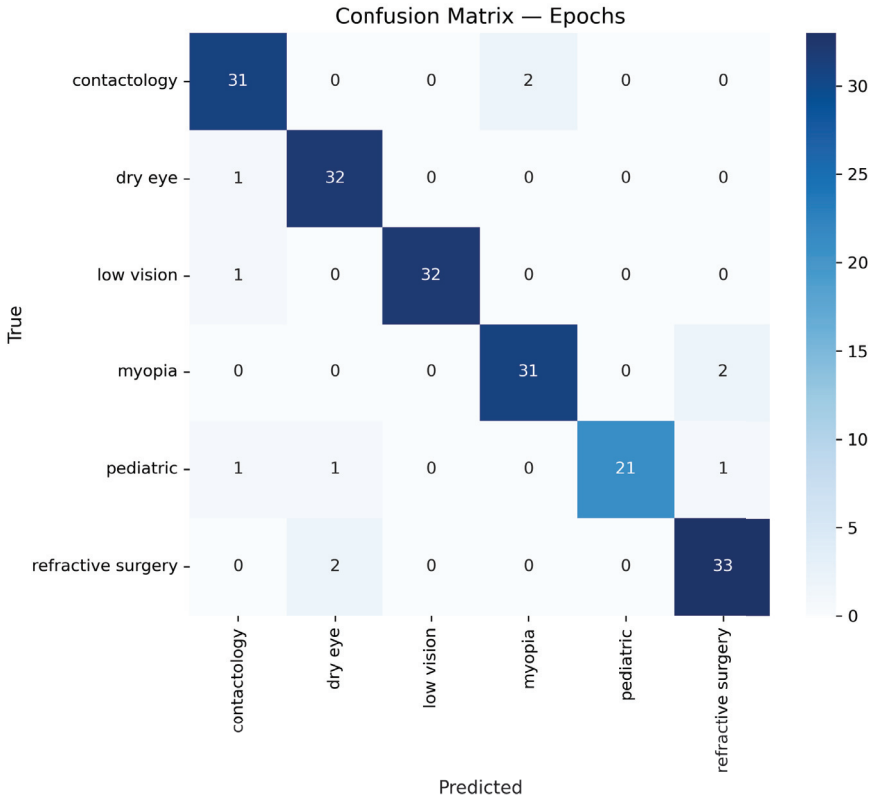


Figure 5. O-RF confusion matrix for epoch training model pipeline.

The most frequent errors are: Contactology mislabelled as Myopia (2 cases), Dry Eye mislabelled as Contactology (1), Low Vision mislabelled as Contactology (1), Myopia mislabelled as Refractive Surgery (2), Pediatric mislabelled as Contactology, Dry Eye and Refractive Surgery (1 each), and Refractive Surgery mislabelled as Dry Eye (2). These patterns are consistent with the per-class recall reported in Table 5: Pediatric shows

the lowest recall at 87.50% (21/24), followed by Contactology and Myopia at 93.94% (31/33 each), while Dry Eye, and Low Vision reach 96.97% (32/33), and Refractive Surgery attains 94.29% (33/35). Specificity is high across the board, with Low Vision and Pediatric at 100% and the remaining classes above 98%, matching the table.

Overall performance remains strong, with overall accuracy of 94.24% and overall precision of 94.70% (Table 5). The error structure indicates occasional overlap between semantically related categories, notably Myopia and Refractive Surgery, and a small amount of leakage from Pediatric into adjacent categories, yet the model preserves a low false positive rate and reliable discrimination for clinical decision support.

3.2. Forecast

Regarding the data forecast, Table 6 presents the projected academic research trends for the prevalence of various optometry-related data from 2025 to 2030. These forecasts were generated using an ARIMA model, providing insights into how the data might evolve over time.

Table 6. ARIMA forecasted data.

Data Label | Year	2025	2026	2027	2028	2029	2030
Contactology	64	66	64	66	64	66
Dry Eye	84	80	83	82	81	82
Low Vision	81	81	81	81	81	81
Myopia	159	164	160	160	161	161
Pediatric	393	401	398	398	399	398
Refractive Surgery	79	74	75	76	75	75

The projections reveal distinct patterns across different categories. Pediatric-related research is expected to maintain the highest prevalence, with a steady increase from 393 in 2025 to 401 in 2026, followed by a stabilization around 398–399 in subsequent years. This consistent upward trend suggests sustained growth in pediatric optometry.

Similarly, the Myopia class demonstrates relative stability, with 159 data entries in 2025, peaking at 164 in 2026, and then maintaining values around 160–161 through 2030. This indicates a continued focus on Myopia, likely due to its increasing global impact.

On the other hand, Dry Eye shows minor fluctuations, with a projected decline from 84 in 2025 to 80 in 2026, followed by a modest recovery to 83 in 2027, and then a slight stabilization near 81–82. This suggests periodic variations rather than a sustained growth or decline.

Low Vision-related publications remain completely stable, with a consistent forecast of 81 data entries per year throughout the entire analyzed period, implying a steady level without significant expected growth or decline.

Contactology and Refractive Surgery display alternating patterns, with Contactology fluctuating between 64 and 66 data entries and Refractive Surgery initially declining from 79 in 2025 to 74 in 2026 before stabilizing around 75–76 in later years. This suggests a cyclical behavior rather than a steady upward or downward trajectory.

Overall, these ARIMA-based forecasts indicate that while some optometry subfields, such as Pediatric and Myopia, are likely to remain dominant, others, like Refractive Surgery and Dry Eye, may experience more variability in the data trends. These projections can serve as a valuable guide for researchers and policymakers, helping to anticipate emerging trends and allocate resources accordingly.

3.3. O-RF Performance

The integration of AI in optometry classification tasks can potentially provide significant advancements in diagnostic accuracy and decision-making processes. The effectiveness of an AI-based model is often assessed using learning curves and performance trends, providing insights into model generalization, stability, and reliability. The AI-based classification framework presented leverages the combination of two distinct training strategies within the Random Forest architecture: the O-RF one-shot trained, which was trained on a single batch of labeled data, enabling rapid learning with high initial accuracy, and the O-RF epoch-trained, which is refined iteratively over multiple training epochs, enhancing generalization and adaptability to unseen data. Figures 6 and 7 illustrate the learning progression and classification performance for both the one-shot and epoch O-RF models, highlighting their adaptability to increasing training data and their long-term consistency in AI-driven classification tasks. Figures 6 and 7 illustrate the learning curves using Macro-F1, which is the appropriate metric for the analyzed multi-class dataset because it averages performance across classes rather than letting majority classes dominate. Curves were computed with stratified cross-validation on a leakage-safe pipeline in which TF-IDF and SMOTE are fitted only inside the training folds.

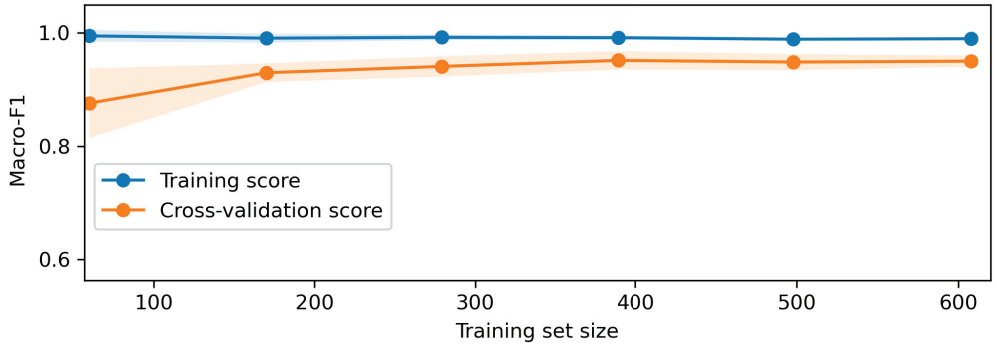


Figure 6. One-shot O-RF learning curve through dataset exposure.

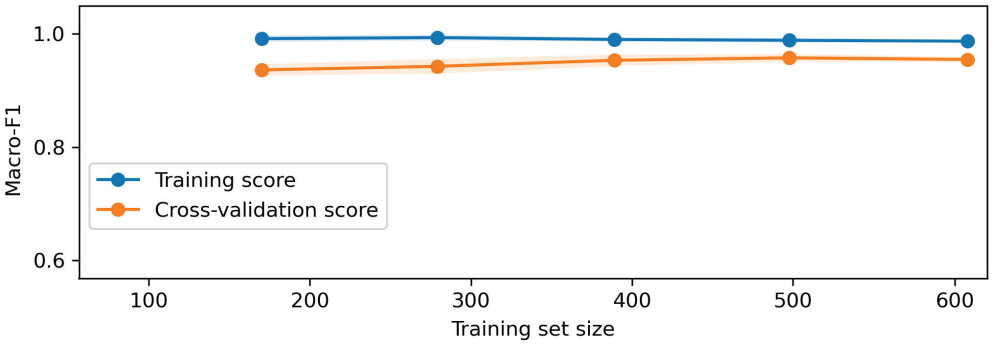


Figure 7. Epoch O-RF learning curve through dataset exposure.

For the one-shot configuration, the training curve sits close to 0.99 across all sampling levels, while the cross-validation curve climbs steadily from roughly 0.88 at the smallest training size to approximately 0.95 at the largest. The gap between the curves shrinks as more data are used, from about 0.12 at the smallest size to about 0.05 at the largest. The shaded uncertainty band around the validation curve contracts with sample size, which indicates decreasing variance across folds. Regarding this, it is safe to say that the developed model does not present overfitting, since a model that is genuinely overfitting shows the opposite pattern: a training curve near perfection coupled with a validation curve that stagnates or degrades as additional training data are added, often with widening uncertainty. None of those signatures appear here. Instead, the validation curve approaches

the training curve and stabilizes at a high Macro-F1, which is the textbook indicator of learned structure that generalizes.

For the epoch configuration pipeline given by Figure 7, the developed model presents even stronger capabilities. The training curve again remains near 0.99, while the cross-validation curve begins around 0.93 and approaches about 0.96 at the largest training size. The generalization gap is slightly smaller than in one-shot, settling near 0.04, and the validation band is narrow at larger sizes. This pattern means additional data help the model more than they help the training score, which is precisely what one expects when the model is not memorizing but is learning stable decision boundaries. If the model were overfitting, the validation curve would not increase monotonically with more data and would not converge toward the training curve.

Taken together, Figures 6 and 7 show consistent and convergent behavior; training performance remains high but not perfectly flat, validation performance increases with sample size and then plateaus close to the training level, and the fold-to-fold uncertainty narrows. These visual facts are sufficient to reject problematic overfitting in both models and affirm that the developed models can perform well. The small residual gap of roughly three to four points at the largest training size is typical of real text classification with class imbalance and label noise and reflects irreducible error rather than memorization. Classic learning-curve theory predicts exactly this shape when model capacity is appropriate and evaluation is performed correctly using cross-validation.

3.4. Comparative Analyses

The comparative assessment of the one-shot models and the epoch-trained models O-RF with O-BERT [15] highlights key differences in their ability to handle optometry-related AI tasks. These differences are illustrated in Table 7, where a quantitative analysis of the main indicators is evaluated.

Table 7. Comparison Optometry Random Forest and Optometry BERT on AI applications in optometry.

	Model	Contactology	Low Vision	Refractive Surgery	Pediatric	Myopia	Dry Eye
F1-Score	One-Shot O-BERT	90.00%	90.70%	86.00%	80.90%	84.40%	86.20%
	Epoch O-Bert	91.20%	91.70%	86.10%	81.90%	84.30%	88.30%
	One-Shot O-RF	94.11%	100%	91.18%	91.30%	93.94%	94.12%
	Epoch O-RF	92.54%	98.46%	92.96%	93.33%	93.94%	94.12%
AUC	One-Shot O-BERT	0.98	1.00	0.98	0.98	0.98	0.98
	Epoch O-Bert	0.98	1.00	0.99	0.98	0.98	0.99
	One-Shot O-RF	0.9962	1.00	0.9890	0.9990	0.9948	0.9965
	Epoch O-RF	0.9944	1.00	0.9929	0.9978	0.9965	0.9960
Recall	One-Shot O-BERT	86.11%	90.32%	92.86%	81.82%	86.00%	88.89%
	Epoch O-Bert	91.67%	90.32%	92.86%	72.72%	80.00%	84.44%
	One-Shot O-RF	96.97%	100%	88.57%	87.50%	93.94%	96.97%
	Epoch O-RF	93.94%	96.97%	94.29%	87.50%	93.94%	96.97%
Specificity	One-Shot O-BERT	98.86%	96.69%	97.06%	97.77%	98.40%	95.21%
	Epoch O-Bert	96.02%	98.34%	98.82%	97.77%	92.72%	96.41%
	One-Shot O-RF	98.10%	100%	98.72%	99.40%	98.73%	98.10%
	Epoch O-RF	98.10%	100%	98.08%	100%	98.73%	98.10%
Overall Precision	One-Shot O-BERT			87.00%			
	Epoch O-Bert			92.80%			
	One-Shot O-RF			94.36%			
	Epoch O-RF			94.70%			

Table 7. *Cont.*

	Model	Contactology	Low Vision	Refractive Surgery	Pediatric	Myopia	Dry Eye
Overall Acuracy	One-Shot O-BERT			86.80%			
	Epoch O-Bert			85.86			
	One-Shot O-RF			94.24%			
	Epoch O-RF			94.25%			

4. Discussion

It should be emphasized that in this proof-of-concept study, the O-RF model was applied to classify optometry-related academic articles; nevertheless, the model is ready to work with real clinical data since the pipeline was designed to be trained with whatever data it is told to learn. Regarding this, the performance metrics reported in Table 7 reflect the model's ability to distinguish among categories of literature, and since the medical data is more straightforward, better metrics are expected with real clinical data.

In this context, the performance of the O-RF model was evaluated under two distinct training strategies: one-shot training and epoch training. The F1-score, defined as the harmonic mean of precision and recall, serves as a comprehensive measure of classification performance by balancing these two metrics. Conventionally, an F1-score above 90% is considered "Very Good," scores between 80% and 90% are classified as "Good," those ranging from 50% to 80% are deemed "Acceptable," whereas scores below 50% indicate "Poor" classification performance. Precision above 90% is regarded as "High Precision," implying strong confidence in positive classifications, whereas values between 80% and 90% are "Moderate Precision," and those below 80% suggest a higher likelihood of false positives. Similarly, accuracy exceeding 95% is typically considered "Excellent," values between 85% and 95% are classified as "Good," while accuracy between 70% and 85% is "Acceptable." Accuracy below 70% generally indicates suboptimal model performance. Regarding the AUC, this serves as a key metric for evaluating a model's effectiveness in differentiating between classes across different threshold settings. An AUC score of 0.5 signifies a model performing at random chance; values below 0.5 indicate poor classification capability, while an AUC of 1.0 represents a perfectly calibrated model with optimal discrimination between categories.

The evaluation of the O-RF model also considers recall and specificity, two critical metrics for assessing classification performance. Recall, also known as sensitivity or the TPR, quantifies the model's ability to correctly identify positive instances. A high recall indicates that the model successfully captures most actual positive cases, reducing false negatives. Conventionally, a recall above 90% is regarded as "Very High Recall," reflecting strong sensitivity. Values between 80% and 90% are considered "Good Recall," while those ranging from 50% to 80% are classified as "Moderate Recall." A recall below 50% suggests the model struggles to correctly identify positive cases, leading to a high false-negative rate. Specificity, also referred to as the TNR, measures the model's ability to correctly classify negative instances, ensuring that false positives are minimized. A specificity above 95% is considered "Excellent," demonstrating the model's ability to confidently exclude negative cases. Values between 85% and 95% are classified as "Good," whereas a specificity between 70% and 85% is "Moderate." Specificity below 70% indicates the model has difficulty distinguishing between positive and negative cases, increasing the risk of false positives.

The epoch-trained O-RF model demonstrated a marginal improvement in overall accuracy compared to the one-shot trained model. Similarly, the overall precision improved slightly in epoch training O-RF, where these improvements suggest that iterative updates through multiple epochs refine the model's ability to generalize across different

classes. Both one-shot and epoch models exhibited strong classification capabilities, as evidenced by consistently high F1-scores and AUC across all categories. However, the epoch-trained model showed an increase in F1-score for categories such as Refractive Surgery and Pediatric suggesting enhanced classification robustness. Despite a slight decrease in Contactology and Low Vision F1-Score, its value remains within an optimal range.

Both the one-shot O-RF and the epoch O-RF exhibit excellent discrimination and well-balanced error profiles. For the one-shot O-RF, the classwise AUC values are near ceiling across the board, ranging from 0.989 in Refractive Surgery to 1.000 in Low Vision, which indicates that the model ranks true class instances ahead of negatives with very high probability even before any thresholding is applied. This ranking strength is matched by a favorable operating point: recall is high in every class, including perfect sensitivity in Low Vision and values above 96 percent in Contactology and Dry Eye, and specificity remains consistently high between 98.10% and 99.40%. The combination of high recall and high specificity means the model is simultaneously capturing positives and rejecting negatives with minimal trade-off. Aggregated metrics confirm this balance, with an overall precision of 94.36% and an overall accuracy of 94.24%, which is consistent with the near-ceiling AUCs and indicates that thresholding is not masking any weakness.

The epoch O-RF retains the same near-perfect separability while slightly improving operating characteristics in a few clinically relevant categories. AUC values remain at or above 0.994 in all classes except Pediatric, which still sits at an excellent 0.9978, and Low Vision again achieves 1.00. Compared with one-shot, recall increases in Refractive Surgery from 88.57% to 94.29%, while specificity remains very high at 98.08%, and Pediatric specificity reaches 100% with unchanged recall. Contactology and Dry Eye keep very strong recall at 93.94% and 96.97% with specificities of 98.10%, and Myopia preserves the same recall with a small gain in AUC. These shifts translate into a small but meaningful improvement in the global operating point: overall precision rises to 94.70%, while overall accuracy is essentially unchanged at 94.25%. The pair of results shows that the epoch procedure preserves the ranking strength of the one-shot model while trimming false positives in aggregate, producing a slightly more conservative and stable classifier.

In the comparative analysis with O-BERT [15], which was designed to perform similar data labeling and diagnostic classification tasks, a clear distinction in performance emerges. While both models demonstrate strong capabilities in optometry-related AI applications, the epoch-trained O-RF model consistently outperforms O-BERT across key evaluation metrics.

Among the most critical performance indicators, accuracy, which quantifies how often the model produces correct classifications, highlights the epoch-trained O-RF model qualified in this context for academic optometric data classification and can be used for diagnosis after plugging in medical data instead of academic literature. This model not only achieves the highest overall accuracy but also excels in precision, specificity, and recall, ensuring a well-balanced trade-off between minimizing false positives and false negatives. In contrast, while O-BERT remains competitive, particularly in recall-oriented tasks, it does not reach the same classification consistency as O-RF, particularly in high-precision applications where misdiagnosis needs to be minimized.

Despite the fact that epoch-trained models require greater computational resources due to their iterative learning approach, the gains in classification performance justify the additional training time and processing power. The O-RF model benefits significantly from repeated learning cycles, refining its decision boundaries and minimizing classification errors over successive iterations. This results in a more stable and generalizable AI model, capable of maintaining high performance across diverse optometry-related classification tasks. The results establish the epoch-trained O-RF model as the most effective methodology

for AI-driven optometric data classification and diagnosis. Its superior performance across multiple evaluation criteria positions it as a robust and highly reliable tool for assisting healthcare professionals in automated screening, predictive analytics, and clinical decision support systems.

The implementation of the ARIMA in forecasting optometric research trends highlights its effectiveness in identifying future directions based on historical data patterns. By leveraging past trends, ARIMA serves as a powerful analytical tool for academic institutions and policymakers, enabling informed decision-making regarding resource distribution and long-term strategic planning in optometry. Applying ARIMA to predict optometric trends from 2025 to 2030 offers critical insights into the evolving landscape of the field. The model's capacity to anticipate future shifts in research priorities allows for the optimization of funding allocation and ensures that resources are directed toward the most impactful areas. Forecasts for different categories, including Contactology, Dry Eye, Low Vision, Myopia, Pediatrics, and Refractive Surgery, reveal distinct trajectories, emphasizing the dynamic nature of optometric advancements. These findings align with previous studies, such as those provided by authors in [21], who applied similar predictive methodologies in neuroscience research forecasting, further demonstrating the reliability and applicability of ARIMA in medical research planning. By proactively identifying emerging trends, academic institutions, funding agencies, and researchers can strategically align their efforts with future demands, ensuring that optometric research remains innovative, data-driven, and relevant to clinical and technological advancements.

5. Conclusions

The O-RF AI model architecture was developed to enhance the classification and predictive capabilities of optometry-related data. Through rigorous training and evaluation, the model demonstrated high performance across multiple optometric classes, effectively classifying data related to Contactology, Dry Eye, Low Vision, Myopia, Pediatrics, and Refractive Surgery. The application of the epoch-trained O-RF model yielded the best overall performance, surpassing both one-shot O-RF and the comparative O-BERT models in terms of accuracy, F1-score, specificity, and recall. The integration of iterative training cycles in the epoch-trained model significantly improved its classification reliability and generalization capabilities, making it a highly effective tool for AI-driven optometry applications.

The epoch-trained O-RF model demonstrated superior classification accuracy compared to O-BERT, reinforcing its potential for diagnostic support and clinical applications. While O-BERT proved effective in data labeling and optometric content classification, its performance was slightly lower in precision and specificity, which are critical factors in clinical decision-making. The higher recall scores of O-BERT models suggest that they may be useful for broad information retrieval tasks, but O-RF remains the preferred model for high-accuracy classification, particularly in cases requiring a balance between sensitivity and precision.

Despite the high performance achieved by the epoch-trained O-RF model, it is important to acknowledge that no AI system achieves absolute accuracy, and there is always room for further enhancement. Future research should focus on incorporating real-world clinical diagnostic data into the training process to further refine classification accuracy and improve generalization across diverse patient populations. Additionally, integrating advanced deep learning techniques alongside the Random Forest framework could further optimize model performance, enhancing its predictive power and robustness in clinical settings.

Beyond its classification capabilities, the ARIMA model was applied to forecast trends from 2025 to 2030, utilizing data previously categorized by the O-RF model. The forecasting

results confirmed the model's ability to predict future research output trends in optometry, making it a valuable tool for strategic planning and resource allocation in both research and clinical research settings. These predictive capabilities align with prior research that has successfully employed similar methodologies in medical fields such as neuroscience. By identifying emerging research directions, the model offers actionable insights that can guide academic agendas, funding distribution, and the alignment of future clinical studies.

For future work, and regarding making the developed models more robust, the future evolution of O2-RF should prioritize the inclusion of real-world patient data in order to include real-world diagnostics capabilities. It is also interesting to develop a hybrid model that leverages both structured (RF-based) and unstructured (BERT-based) learning approaches. This combination could enhance both the interpretability and adaptability of AI systems in medical applications. The continued development of optometry AI architectures will be essential in advancing automated screening tools, improving diagnostic accuracy, and optimizing resource allocation for global eye health initiatives.

Author Contributions: L.F.F.M.S.: Deep Learning and Artificial Intelligence Model Development, Theoretical Formulation, Conceptualization, Original Draft, Visualization, Formal Analysis. C.A.-P.: Data Management, Conceptualization, Visualization, Original Draft, Methodology, Formal Analysis. M.Á.S.-T.: Validation, Formal Analysis, Review. C.M.-P.: Validation, Formal Analysis, Review. All authors have read and agreed to the published version of the manuscript.

Funding: The present work was performed under the scope of activities at the Aeronautics and Astronautics Research Center (AEROG) of the Laboratório Associado em Energia, Transportes e Aeroespacial (LAETA), and was supported by the Fundação para a Ciência e Tecnologia (Project Nos. UIDB/50022/2020, UIDP/50022/2020, and LA/P/0079/2020).

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial Intelligence
ARIMA	Autoregressive Integrated Moving Average
AUC	Area Under the Curve
BERT	Bidirectional Encoder Representations from Transformers
cpRNFL	Retinal Nerve Fiber Layer
FPR	False Positive Rate
ML	Machine Learning
MSE	Mean Squared Error
NLP	Natural Language Processing
O-BERT	Optometry BERT
O-RF	Optometry Random Forest
RF	Random Forest
RGC	Retinal Ganglion Cells
ROC	Receiver Operating Characteristic
SVM	Support Vector Machines
TNR	True Negative Rate
TPR	True Positive Rate
VF	Visual Field

References

1. Malviya, A.; Dhole, S.; Maurya, C.K. Unsupervised continual learning by cross-level, instance-group and pseudo-group discrimination with hard attention. *J. Comput. Sci.* **2025**, *86*, 102535. [CrossRef]
2. Hashemian, H.; Petö, T.; Ambrósio, R., Jr.; Lengyel, I.; Kafieh, R.; Noori, A.; Khorrami-Nezhad, M. Application of Artificial Intelligence in Ophthalmology: An Updated Comprehensive Review. *J. Ophthalmic Vis. Res.* **2024**, *19*, 354. [CrossRef] [PubMed]
3. Li, Z.; Wang, L.; Wu, X.; Jiang, J.; Qiang, W.; Xie, H.; Zhou, H.; Wu, S.; Shao, Y.; Chen, W. Artificial intelligence in ophthalmology: The path to the real-world clinic. *Cell Rep. Med.* **2023**, *4*, 101095. [CrossRef] [PubMed]
4. Bhattacharjee, S.; Saha, B.; Bhattacharyya, P.; Saha, S. Classification of obstructive and non-obstructive pulmonary diseases on the basis of spirometry using machine learning techniques. *J. Comput. Sci.* **2022**, *63*, 101768. [CrossRef]
5. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
6. Lázaro, F.L.; Madeira, T.; Melicio, R.; Valério, D.; Santos, L.F.F.M. Identifying Human Factors in Aviation Accidents with Natural Language Processing and Machine Learning Models. *Aerospace* **2025**, *12*, 106. [CrossRef]
7. Quigley, H.; Broman, A.T. The number of people with glaucoma worldwide in 2010 and 2020. *Br. J. Ophthalmol.* **2006**, *90*, 262–267. [CrossRef] [PubMed]
8. Antoniadis, A.; Fasiolo, M.; Goude, Y.; Poggi, J.M. *Random Forests*; Birkhäuser: Cham, Switzerland, 2024; pp. 99–111. . [CrossRef]
9. Wang, S.; Ji, Y.; Bai, W.; Ji, Y.; Li, J.; Yao, Y.; Zhang, Z.; Jiang, Q.; Li, K. Advances in artificial intelligence models and algorithms in the field of optometry. *Front. Cell Dev. Biol.* **2023**, *11*, 1170068. [CrossRef] [PubMed]
10. Zou, X. R.; Zhang, P.; Zhou, Y.; Yin, Y. Ocular surface microbiota in patients with varying degrees of dry eye severity. *Int. J. Ophthalmol.* **2023**, *16*, 1986–1995. [CrossRef] [PubMed]
11. Stuermer, L.; Braga, S.; Martin, R.; Wolffsohn, J. Artificial intelligence virtual assistants in primary eye care practice. *Ophthalmic Physiol. Opt.* **2024**, *45*, 437–449. [CrossRef] [PubMed]
12. Morgan, I.; French, A.; Ashby, R.; Guo, X.; Ding, X.; He, M.; Rose, K. The epidemics of myopia: Aetiology and prevention. *Prog. Retin. Eye Res.* **2017**, *62*, 134–149. [CrossRef] [PubMed]
13. Lin, H.; Long, E.; Ding, X.; Diao, H.; Chen, Z.; Liu, R.; Huang, J.; Cai, J.; Xu, S.; Zhang, X.; et al. Prediction of myopia development among Chinese school-aged children using refraction data from electronic medical records: A retrospective, multicentre machine learning study. *PLoS Med.* **2018**, *15*, e1002674. [CrossRef] [PubMed]
14. Martínez-Albert, N.; Bueno, I.; Gene-Sampedro, A. Risk Factors for Myopia: A Review. *J. Clin. Med.* **2023**, *12*, 6062. [CrossRef] [PubMed]
15. Santos, L.F.F.M.; Sánchez-Tena, M.A.; Alvarez-Peregrina, C.; Sánchez-González, J.M.; Martinez-Perez, C. The Role of Artificial Intelligence in Optometric Diagnostics and Research: Deep Learning and Time-Series Forecasting Applications. *Technologies* **2025**, *13*, 77. [CrossRef]
16. Lázaro, F.L.; Nogueira, R.P.R.; Melicio, R.; Valério, D.; Santos, L.F.F.M. Human Factors as Predictor of Fatalities in Aviation Accidents: A Neural Network Analysis. *Appl. Sci.* **2024**, *14*, 640. [CrossRef]
17. Kurul, E.; Tunc, H.; Sari, M.; Guzel, N. Deep learning aided surrogate modeling of the epidemiological models. *J. Comput. Sci.* **2025**, *84*, 102470. [CrossRef]
18. Krishnan, A.; Dutta, A.; Srivastava, A.; Konda, N.; and, R.K.P. Artificial Intelligence in Optometry: Current and Future Perspectives. *Clin. Optom.* **2025**, *17*, 83–114. [CrossRef] [PubMed]
19. Ji, Y.; Liu, S.; Hong, X.; Lu, Y.; Wu, X.; Li, K.; Li, K.; Liu, Y. Advances in artificial intelligence applications for ocular surface diseases diagnosis. *Front. Cell Dev. Biol.* **2022**, *10*, 1107689. [CrossRef] [PubMed]
20. Pourvahab, M.; Mousavirad, S.J.; Felizardo, V.; Pombo, N.; Zacarias, H.; Mohammadigheymasi, H.; Pais, S.; Jafari, S.N.; Garcia, N.M. A cluster-based opposition differential evolution algorithm boosted by a local search for ECG signal classification. *J. Comput. Sci.* **2025**, *86*, 102541. [CrossRef]
21. Haug, C.; Drazen, J. Artificial Intelligence and Machine Learning in Clinical Medicine. *N. Engl. J. Med.* **2023**, *388*, 1201–1208. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

algorithms

MDPI

Article

Probabilities Feature Enrichment for Improved Early Diabetes Prediction

Haneen Altartouri [1,*], Sahar Qaadan [2], Yousef Qawqzeh [3,†] and Mohammed Hamdoun [1]

[1] College of Engineering and Technology, Fujairah University, Fujairah 1207, United Arab Emirates; f202410028@fu.ac.ae
[2] Mechatronics Engineering, German Jordanian University, Madaba Street, Amman 11180, Jordan; sahar.qadan@gju.edu.jo
[3] Cyber Security Department, Khawarizmi University Technical College, Amman 99753, Jordan
* Correspondence: h.altartouri@fu.ac.ae
† Current address: Institute of Applied Technology, Fatima College of Health Sciences, Abu Dhabi 3798, United Arab Emirates.

Abstract

Early prediction of diabetes remains a major healthcare challenge, as timely diagnosis can significantly reduce the risk of severe complications. Most existing studies emphasize complex predictive models or extensive preprocessing pipelines, while comparatively limited attention has been given to improving feature representations. In this study, we propose a statistical framework for enriching the feature space of diabetes data using Gaussian Mixture Models (GMMs) and Kernel Density Estimation (KDE) to capture latent distributional characteristics and local density patterns. The resulting enriched features enhance class discrimination while preserving model simplicity, allowing standard classifiers to operate more effectively. The proposed framework is evaluated on two diabetes datasets with different sample sizes and class distributions. Models trained on the enriched feature set are compared with baseline models trained solely on the original features, using multiple strategies, including direct feature concatenation and ensemble-based approaches. Experimental results show that the enriched feature space consistently outperforms the original feature set, leading to notable improvements in overall predictive performance. In particular, the framework achieves a substantial increase in sensitivity for the minority class (diabetic cases), which is especially important for clinical applications where positive samples are typically scarce relative to healthy cases.

Keywords: diabetes prediction; feature space enrichment; Gaussian mixture models; kernel density estimation

1. Introduction

Diabetes mellitus is a chronic metabolic disorder characterized by elevated blood glucose levels resulting from impaired insulin secretion, insulin resistance, or a combination of both [1]. According to the International Diabetes Federation (IDF) Diabetes Atlas 2025, approximately 537 million adults aged 20–79 are currently living with diabetes [2]. This number is projected to increase to 783 million worldwide by 2045 [1].

Type 2 diabetes (T2D) is the most prevalent form of diabetes, accounting for approximately 95% of all cases worldwide [3]. It develops when body cells become resistant to insulin and, over time, the pancreas is unable to produce sufficient insulin to compensate for this resistance, leading to chronic hyperglycemia [4]. The progressive nature of T2D often

results in substantial metabolic dysfunction and long-term damage to multiple organ systems, including the cardiovascular, renal, and nervous systems [3,5]. Beyond its high prevalence, T2D represents a major public health challenge due to its severe complications and associated mortality. In 2024 alone, diabetes was estimated to have contributed to 3.4 million deaths globally, corresponding to approximately one death every nine seconds [2].

Early detection and intervention are crucial for preventing disease progression and reducing the risk of complications in individuals with diabetes [3]. However, traditional diagnostic methods based on clinical biomarkers, such as fasting glucose, HbA1c, and oral glucose tolerance tests [6], may fail to identify early-stage metabolic dysfunction. In recent years, machine learning (ML) has emerged as a powerful tool for predictive modeling in healthcare, offering the ability to analyze complex, multidimensional datasets to identify individuals at high risk [7,8]. By integrating clinical, genetic, and lifestyle factors, ML algorithms can enhance the early prediction of diabetes, thereby enabling timely interventions and supporting personalized treatment strategies [7,8].

ML approaches have been widely applied to diabetes prediction and are often evaluated using publicly available benchmark datasets. One of the main challenges, common to many medical datasets, is the high degree of class imbalance [9], which can bias predictive models toward the majority class and thereby limit their clinical usefulness. To mitigate this issue and improve predictive performance, various methodologies have been proposed, including data-level strategies such as oversampling and undersampling, algorithm-level modifications through cost-sensitive learning, and hybrid approaches that combine preprocessing, feature engineering, and advanced classification techniques across diverse datasets.

Sharma et al. [10] examined the performance of several classification algorithms on the PIMA Indian Diabetes dataset. Their findings indicated that the Logistic Regression (LR) model achieved superior performance compared with the other evaluated classifiers for this dataset. Similarly, Ağraz et al. [11] proposed a hybrid approach that combined the Dendritic Neuron Model (DNM), LR, and an Artificial Neural Network (ANN) on the same dataset. Their hybrid model outperformed the standalone classifiers in terms of AUC, accuracy, and sensitivity. Ahmed et al. [12] also focused exclusively on the PIMA dataset, applying feature selection followed by multiple machine learning algorithms to improve predictive performance. Their results demonstrated that the Random Forest classifier outperformed the other models.

Reza et al. [13] investigated two stacking-based ensemble models for diabetes classification using the PIMA dataset and a local healthcare dataset. The first model employed a classical stacking approach, integrating the base learners Random Forest (RF), Decision Trees (DT), and Support Vector Machines (SVM), with LR as the meta-classifier. The second model replaced the base learners with three customized sequential neural networks, again using LR as the meta-classifier. Both models were evaluated using train–test splits and cross-validation. The results demonstrated that both classical stacking and neural network–based stacking approaches outperformed individual classifiers, highlighting the effectiveness of ensemble learning. Similarly, Sivashankari and Vengatesan [14] proposed a stacked ensemble for PIMA-based diabetes prediction, using multiple level-0 base classifiers, namely RF, K-Nearest Neighbors (KNN), DT, Gradient Boosting, SVM, and Gaussian Naïve Bayes, to generate probability outputs, which were then used as input to an LR meta-classifier. Their ensemble achieved superior performance compared with individual classifiers, including RF, KNN, LR, Gradient Boosting, AdaBoost, and SVM. Tripathi et al. [15] applied a soft voting ensemble on the PIMA dataset, comparing LR, KNN, SVM, DT, AdaBoost, and Gradient Boosting, and achieved up to 82.8% accuracy, surpassing the performance of the individual models.

https://doi.org/10.3390/a19020122

Although much research has focused on the PIMA dataset, several studies have extended their methodologies to other datasets or applied them across multiple datasets. Shi et al. [16] proposed an integrated Bagging–RF model to enhance diabetes prediction using two Kaggle datasets. Their preprocessing pipeline included outlier removal, one-hot encoding, age-based stratification into three groups (50–60, 60–70, and 70–80 years), and SMOTE oversampling. The Bagging–RF model was compared with RF, Gaussian Naïve Bayes (GNB), LR, KNN, DT, Extra Trees (ET), AdaBoost, and XGBoost (XGB) and evaluated using accuracy, precision, sensitivity, specificity, AUC, and F1-score. Bagging–RF consistently achieved the highest performance, with accuracy ranging from 93.7% to 97.4% and superior F1-scores. Talukder et al. [17] aimed to improve the reliability of diabetes prediction by integrating data engineering with machine learning. Using PIMA, the Iraqi Medical City dataset (Mendeley), and the Kaggle "Diabetes Prediction Dataset," they applied preprocessing steps including missing value handling, outlier removal, normalization, label encoding, and random oversampling, followed by training LR, RF, SVM, and XGBoost. Their results showed that robust preprocessing significantly enhanced accuracy, sensitivity, and AUC across all datasets. Uddin et al. [18] designed a framework to reduce false negatives in diabetes detection, training LR, KNN, Naïve Bayes, RF, SVM, DT, and Linear Regression on a 2019 diabetes dataset and PIMA. Preprocessing involved missing value handling, normalization, categorical encoding, and SMOTE oversampling. Their analysis revealed that Random Forest achieved the highest performance, with up to 97% accuracy on the 2019 dataset and 80% accuracy on the PIMA dataset. Importantly, balancing the dataset significantly reduced false negatives, highlighting the importance of data balancing techniques in enhancing model sensitivity and improving practical reliability in clinical prediction tasks [18].

Despite the advances demonstrated in these studies, a common limitation remains the inherently restricted feature space typically available for diabetes prediction tasks. Most previous works have relied on multi-stage preprocessing pipelines or increasingly complex model architectures to improve prediction performance [19]. While these strategies can achieve incremental improvements in predictive accuracy, they often come at the cost of increased computational complexity and reduced model interpretability. More importantly, such methods do not fundamentally address the core challenge posed by the limited diversity and richness of input features, underscoring the need for novel approaches that expand the feature space to achieve substantial and meaningful gains in predictive performance.

This study differs from the previous works by shifting focus from model complexity to feature representation itself. Instead of relying on deeper architectures, heavy preprocessing, or ensemble pipelines, we investigate whether classical models can achieve comparable or superior performance when supplied with richer, probabilistically enhanced features. Our key idea is that diabetes datasets are typically low-dimensional and contain limited discriminative information in their raw form; therefore, improving what the classifier sees is more impactful than increasing model complexity. We introduce a feature enrichment framework using GMM capture latent sub-population structures and KDE to encode local density information, generating new probability-based features that reflect the underlying data distribution. These enriched representations are appended to the original variables, enabling traditional classifiers, such as logistic regression, to learn from distribution-aware input rather than raw measurements alone.

Several prior studies have employed GMM and KDE in biomedical and signal processing applications to extract informative features; however, their use in diabetes prediction remains relatively limited. Suwalska and Polańska [20] proposed a GMM-based feature expansion framework for single-cell mass cytometry, in which marker intensity distribu-

https://doi.org/10.3390/a19020122

tions were modeled as mixtures of Gaussian components. By projecting the data into an expanded feature space using GMM membership probabilities, their approach improved clustering accuracy and the identification of rare cell subtypes compared to standard methods. Ghaderi et al. [21] introduced three novel surface electromyography (sEMG) features derived from KDE, namely the trimmed mean of the density, density entropy, and derivative-based density measures, for myoelectric signal analysis. When combined with ensemble learning, these KDE-based features achieved exceptionally high classification accuracy (approximately 99%) for hand and wrist motion recognition across multiple benchmark datasets, demonstrating their effectiveness for biomedical feature extraction. In a related context, Chushig-Muzo et al. [22] proposed a methodology to improve the interpretability of autoencoder-derived latent representations from electronic health records. By integrating GMM with hierarchical clustering and Kullback–Leibler divergence, their approach successfully identified clinically meaningful patient subgroups, particularly among hypertensive and insulin-dependent diabetic cases, thereby enhancing transparency in deep learning-based decision support systems. Collectively, these studies demonstrate that GMM and KDE can enhance data representations by capturing latent distributional structures and local density variations. However, to the best of our knowledge, no existing study has jointly utilized GMM-based soft cluster probabilities and KDE-derived log-density features for feature augmentation in early diabetes detection datasets.

The proposed framework is evaluated on two widely used public datasets: the PIMA Indian Diabetes dataset and the Kaggle Diabetes Prediction Dataset. The evaluation aims to assess both predictive performance and cross-dataset robustness. LR, a widely adopted and interpretable baseline in clinical decision support systems, is employed to isolate the impact of feature enrichment without the confounding effects of complex model architectures. The evaluation considers four modeling strategies: (1) models trained using original features only, (2) models trained using original features enriched with GMM and KDE derived features, (3) block-wise stacking with LR as the meta-classifier, and (4) ensemble averaging across feature blocks.

The experimental results demonstrate that the proposed feature representation consistently improves overall predictive performance and significantly enhances sensitivity to the minority class without significantly compromising other evaluation metrics, thereby representing a meaningful advancement for medical prediction tasks.

2. Materials and Methods

This section provides a detailed description of the diabetes datasets used in this research, the proposed methodology for feature representation and model development, and the evaluation metrics employed to assess predictive performance.

2.1. Datasets

To evaluate the impact of feature space enrichment on diabetes prediction, we utilized two widely recognized binary classification datasets with distinct distributions and feature compositions. This selection enables the assessment of the proposed methodology on both small-scale clinical benchmarks and large-scale electronic health record (EHR)-based data, thereby ensuring robustness and generalizability. Detailed descriptions of each dataset are provided below.

2.1.1. Pima Indians Diabetes Dataset (PIMA)

The PIMA dataset, originally collected by the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK), is publicly available through the UCI Machine Learning Repository [23]. It contains 768 samples, each representing a female patient of

https://doi.org/10.3390/a19020122

Pima Indian heritage, aged at least 21 years. Each sample has seven numerical features derived from medical examinations, along with a binary outcome which indicates the presence or absence of diabetes. The features are as follows [23]:

- Pregnancies: Number of times pregnant;
- Glucose: Plasma glucose concentration (mg/dL);
- BloodPressure: Diastolic blood pressure (mmHg);
- SkinThickness: Triceps skin fold thickness (mm);
- Insulin: 2-hour serum insulin (μU/mL);
- BMI: Body mass index (kg/m^2);
- DiabetesPedigreeFunction: Likelihood of diabetes based on family history;
- Age: Age in years.

This dataset is an imbalanced dataset that contains 500 non-diabetic and 268 diabetic samples. Some features have zero values that are physiologically implausible, such as Glucose, BloodPressure, SkinThickness, Insulin, and BMI features. These values were treated as missing and imputed using the median of each respective feature, a method known for its robustness to outliers and effectiveness in clinical datasets [24–26].

To ensure uniform scaling and improve convergence during model training, all features were standardized to have zero mean and unit variance.

2.1.2. Electronic Health Records Dataset (EHRs)

The EHRs dataset was collected from surveys, medical records, and laboratory tests conducted by healthcare providers in hospitals and clinics. This dataset is publicly available on Kaggle and is widely utilized in the field of diabetes prediction [27–30]. The dataset comprises 100,000 samples, each containing eight features, namely six numerical and two categorical, along with a target variable indicating the presence or absence of diabetes. The features are as follows [31]:

- gender: Categorical variable indicating the patient's gender (e.g., male, female, or other).
- age: Numerical variable representing the patient's age in years.
- hypertension: Binary variable indicating the presence or absence of hypertension.
- heart_disease: Binary variable indicating the presence or absence of heart disease.
- smoking_history: Categorical variable detailing the patient's smoking history (e.g., never, former, current, or no info).
- bmi: Numerical variable representing the Body Mass Index (kg/m^2).
- HbA1c_level: Numerical variable indicating the level of hemoglobin A1c, a measure of blood sugar levels over time (percent).
- blood_glucose_level: Numerical variable representing the current blood glucose level (mg/dL).

The dataset is highly imbalanced, with 8500 diabetic and 91,500 non-diabetic samples. To facilitate the application of various classification algorithms, the categorical features (gender and smoking_history) were converted into binary features.

2.2. Proposed Methodology

This study introduces a feature space enrichment framework designed to improve the predictive performance of diabetes classification models. The framework enhances the original feature set by generating additional, informative variables using two statistical modeling techniques: GMM and KDE. These enriched features are concatenated with the original ones to form a more comprehensive and discriminative feature vector for model training.

https://doi.org/10.3390/a19020122

Beyond evaluating the enriched feature vector directly, two ensemble learning approaches are applied to further assess its potential: a block-wise stacking ensemble and a block-wise soft voting ensemble, both detailed in this section.

2.2.1. Feature Enrichment

The first step of the proposed framework involves generating new features using GMM and KDE to capture latent patterns and distributional characteristics in the diabetes data. These additional variables are integrated with the original features to construct a more representative feature space, aiming to improve classification performance.

Feature Space Enrichment Using Gaussian Mixture Models (GMMs)

GMM is a probabilistic model that represents data as a mixture of multiple Gaussian distributions [32]. It is effective for capturing complex feature distributions, especially when the data follow more than one normal distribution [32]. Unlike traditional clustering techniques, such as k-mean, which assume well-separated spherical clusters, GMM can model elliptical clusters, making it highly effective in identifying underlying patterns within the data. This capability allows GMM to handle complex distributions more effectively, ensuring that subtle variations and multimodal characteristics in the dataset are well captured [32].

Each Gaussian distribution represents a cluster and has its own mean and covariance. The model also includes mixing coefficients (weights) that determine the importance of each Gaussian component in the mixture. The goal of GMM is to estimate parameters, including mean values, covariances, and mixing coefficients, to best describe the data [33].

The probability density function of the GMM is given by

$$p(\mathbf{x} \mid \Theta) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

Here, π_k represents the mixing coefficient for the Gaussian component k-th, where $\sum_{k=1}^{K} \pi_k = 1$, $\mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ is the multivariate Gaussian distribution with mean $\boldsymbol{\mu}_k$ and covariance matrix $\boldsymbol{\Sigma}_k$, and Θ represents the set of model parameters.

To estimate the parameters of GMM, the Expectation–Maximization (EM) algorithm is used, which iteratively improves the model fit. The EM algorithm consists of two key steps: In the expectation step, the algorithm calculates the probability that each data point belongs to each cluster γ_{ik} [33], as shown in the following equation:

$$\gamma_{ik} = \frac{\pi_k \mathcal{N}(x_i \mid \mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(x_i \mid \mu_j, \Sigma_j)}$$

where these probabilities are based on the current estimates of the Gaussian parameters.

For the maximization step, the probabilities from the expectation step is used to update the parameters of the Gaussian distributions (means and covariances) and the mixing coefficients to better fit the data. These two steps are repeated until the model converges, meaning that the parameters no longer change significantly [33]. This iterative process effectively handles the complexities of overlapping clusters with varying shapes in the data, which are common in real-world datasets.

The main challenge in applying GMM is determining the optimal number of clusters (Gaussian components) that best represent the underlying data structure. Selecting an inappropriate number of clusters can lead to overfitting or underfitting, resulting in poor performance for any model. To address this challenge, several methods have been proposed in the literature [33]. In this study, the Bayesian Information Criterion (BIC) is used to determine the optimal number of clusters for the diabetic datasets [34]. BIC is a statistical

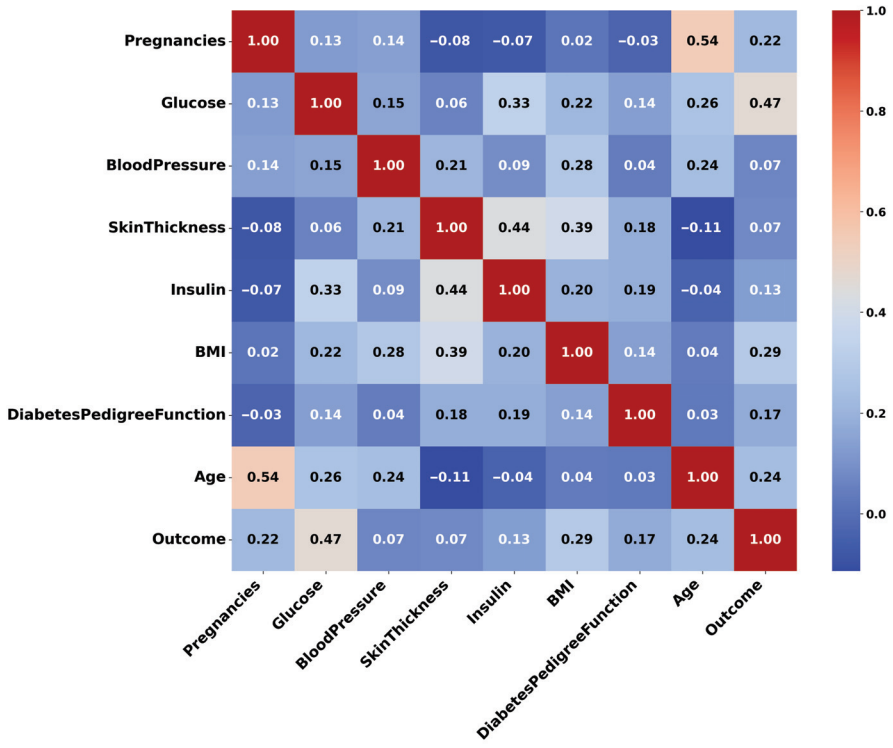
https://doi.org/10.3390/a19020122

criterion that compares models by balancing goodness of fit with model complexity, thereby helping to prevent overfitting and ensuring more reliable clustering results [34]. We evaluated GMM models for $K = 1$ to $K = 10$, and the model with the lowest BIC value was selected. To further enhance robustness, each candidate GMM was fitted using multiple random initializations, and the average BIC value was used to confirm the stability of the selected K. This procedure ensures that the probabilistic features derived from the GMM are consistent, reproducible, and do not depend on arbitrary initialization.

In this study, we applied a GMM to each non-categorical feature individually, rather than to the entire feature space, to enrich the representation of the data. This decision is based on the correlation analysis of the features, which shows that there are no pairs of features with a correlation higher than 0.62, as illustrated in Figure 1. As a result, modeling the full feature set using a multivariate GMM with the complete feature set is unlikely to extract meaningful structure because strong interactions between features are not present. In contrast, applying the GMM to each feature separately captures the local distributional characteristics within each feature, which can be beneficial for enriching the original feature space.

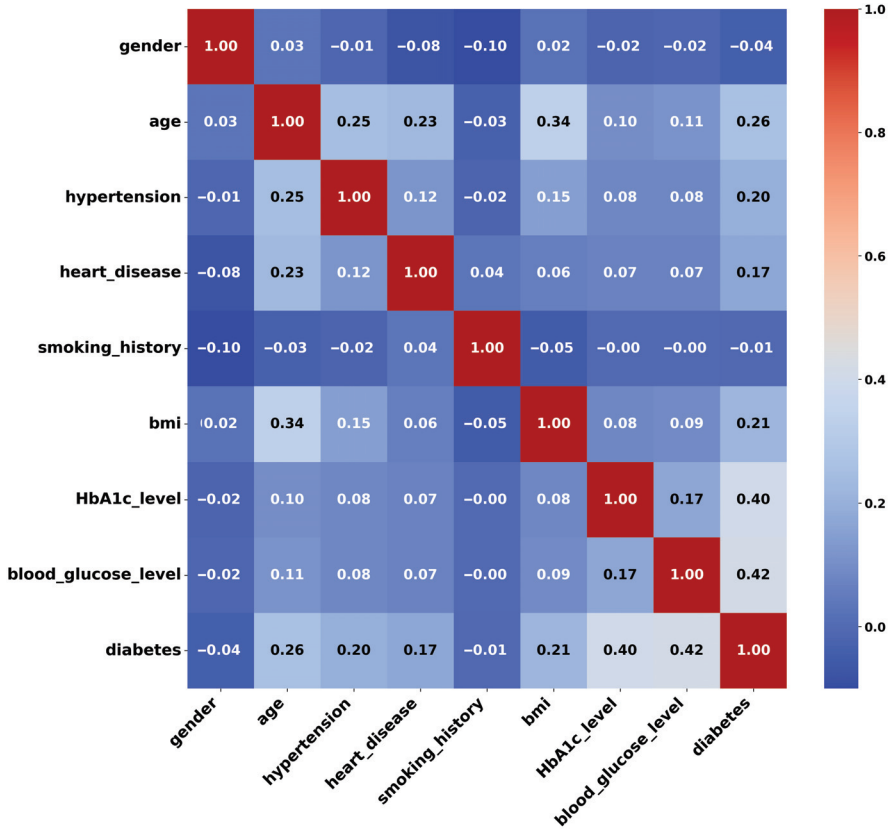
For each feature, we add cluster probabilities (soft assignments) as new features.

To apply per-feature GMM, the distributions of all non-categorical features were examined to understand the underlying characteristics of each feature, such as skewness, modality, and overall distribution shape. The blue histograms in Figure 2 show the distribution of the features in the PIMA dataset, where we can notice that some of the features, such as Glucose, Blood Pressure, and BMI, have approximately normal distributions, while other features exhibit skewness and non-Gaussian characteristics. Figure 3 (blue histograms) illustrates that most features in the EHRs dataset display skewed and multimodal distributions.



(**a**) PIMA dataset

Figure 1. *Cont.*



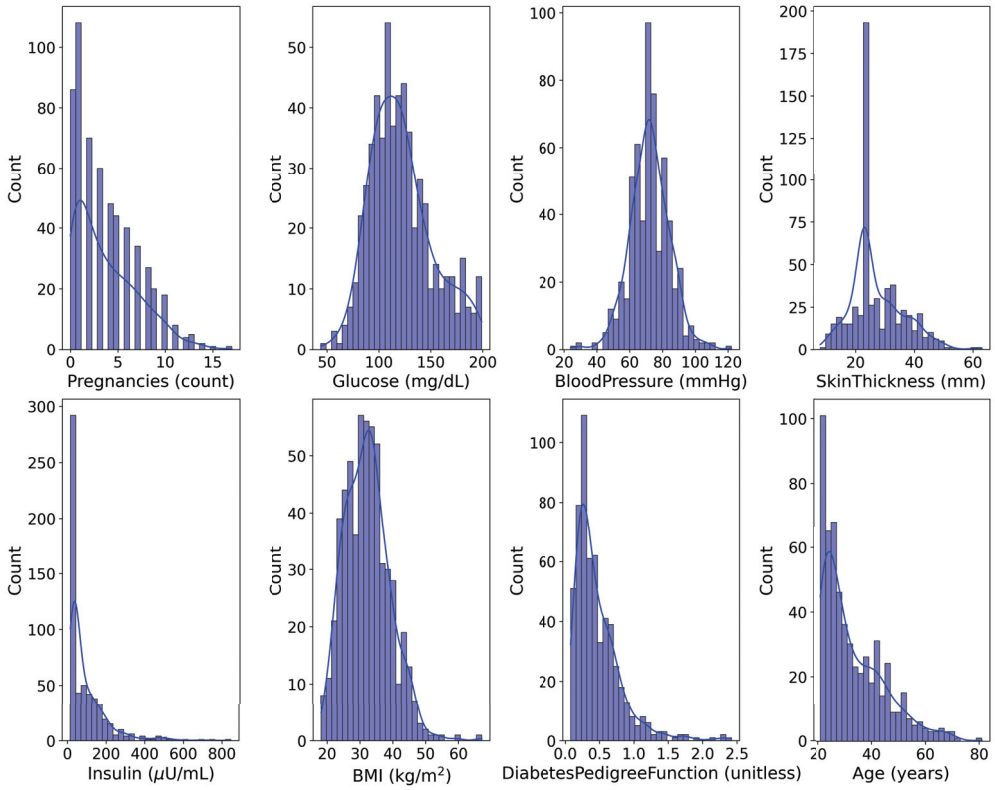
(**b**) EHRs dataset

Figure 1. Correlation plots for the two diabetes datasets. Each plot shows the pairwise correlations between the features in the dataset, where the color gradient represents the degree of correlation.

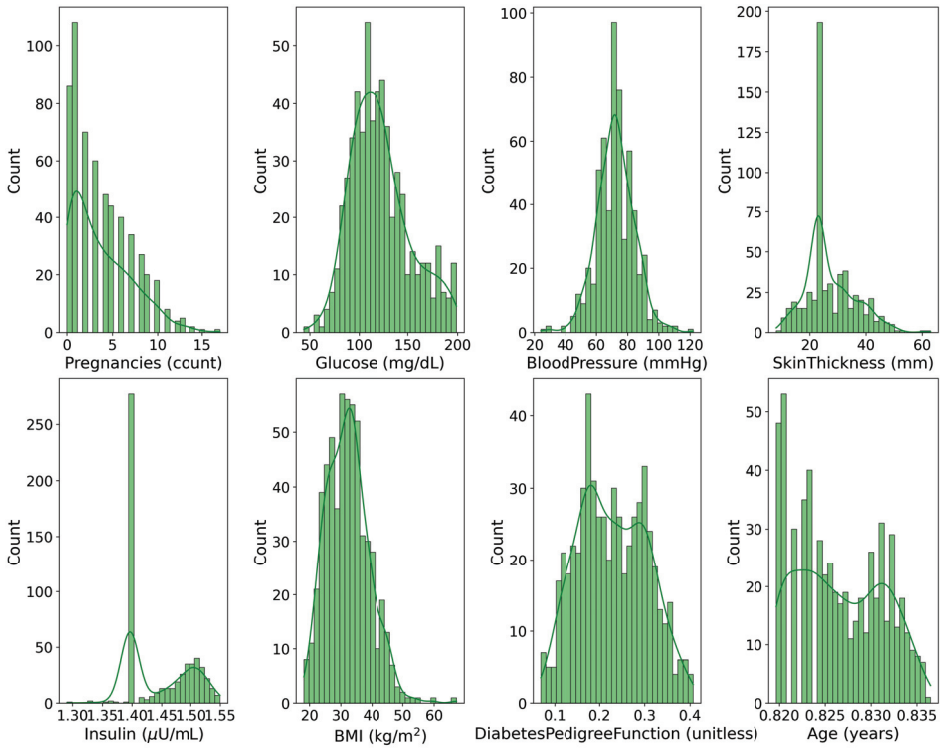
Although GMM is a probabilistic model that assumes data is generated from a mixture of multiple Gaussian distributions, it can still be applied to non-Gaussian data because it models clusters using multiple Gaussian components. However, when a dataset exhibits complex shapes (e.g., skewed or multimodal distributions), GMM might struggle to fit the data properly, leading to unstable clustering and less accurate cluster assignments that may not reflect meaningful underlying structure.

To address this issue, we apply the Box–Cox transformation, a power transformation that aims to stabilize variance and make the data more normally distributed. Figure 2 (green histograms) illustrates the transformed feature distributions, demonstrating a significant reduction in skewness. The DiabetesPedigreeFunction, Glucose, and BloodPressure distributions become more symmetric, making them more amenable to clustering. The transformation notably improves Age and Insulin, although the latter still exhibits a zero-inflated pattern. While some features retain minor deviations from normality, the overall transformation effectively enhances Gaussian-like behavior across most variables. The distributions of the non-categorical features from the PIMA dataset are shown both before and after the Box–Cox transformation. On the other hand, Figure 3 illustrates that the distributions of BMI are significantly improved in the EHRs dataset, becoming more symmetric and Gaussian-like, making it more suitable for GMM, whereas the transformation has a limited effect on features such as blood glucose level, HbA1c level, and age.

https://doi.org/10.3390/a19020122



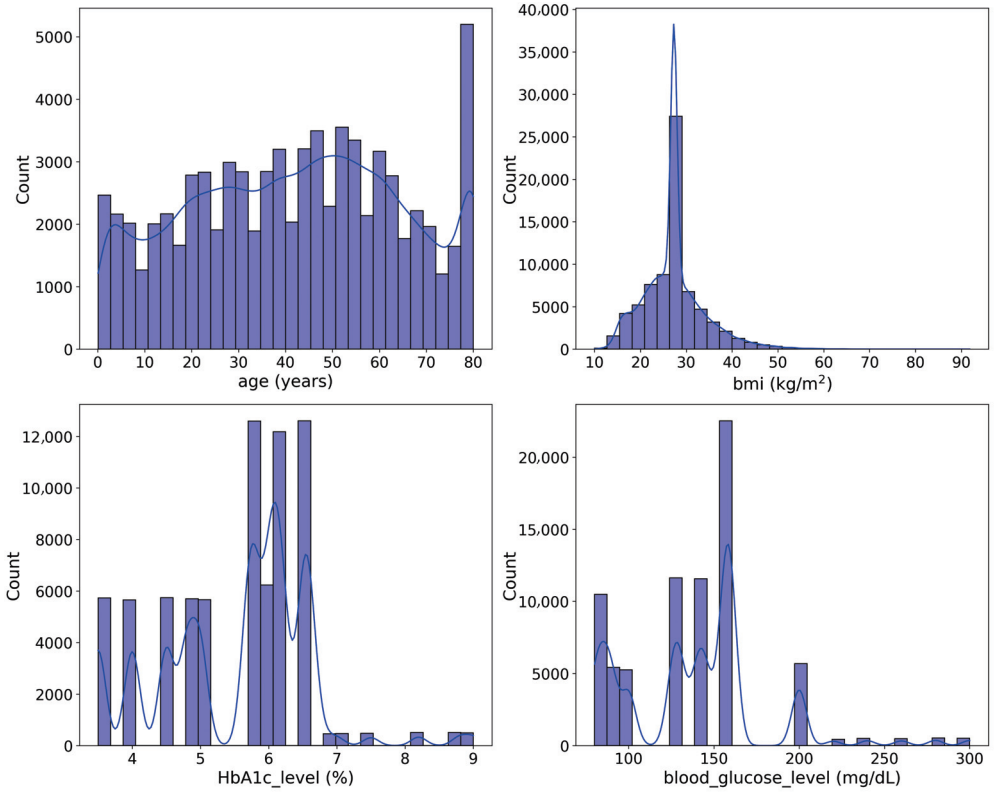
(**a**) Original distributions of features.



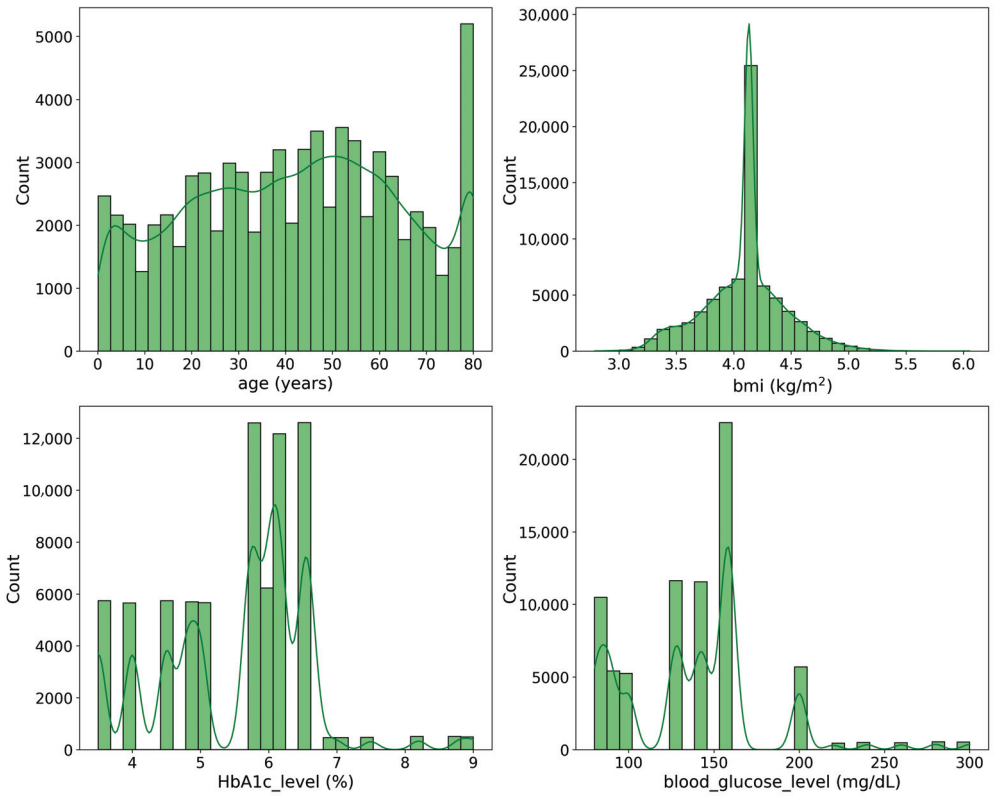
(**b**) Distributions after Box–Cox transformation.

Figure 2. The distribution of features in the PIMA dataset before and after applying a transformation to improve their suitability for clustering. The blue histograms present the original distributions of non-categorical features, while the green histograms present the distributions after applying the Box–Cox transformation.

https://doi.org/10.3390/a19020122



(**a**) Original distributions of features.



(**b**) Distributions after Box–Cox transformation.

Figure 3. The distribution of features in the EHRs dataset before and after applying a transformation to improve their suitability for clustering. The blue histograms present the original distributions of non-categorical features, while the green histograms present the distributions after applying the Box–Cox transformation.

https://doi.org/10.3390/a19020122

Feature Space Enrichment Using Kernel Density Estimation (KDE)

KDE is a non-parametric technique used to estimate the probability density function of a random variable based on observed data points without assuming any specific distribution [35]. It effectively captures the local density of the data points, highlighting regions of high and low data concentration [35]. In contrast to GMM, which is a parametric and assumes the data follow a specific distribution that has a number of parameters, such as mean and covariance [36], KDE constructs the density estimate by centering a kernel function at each data point and summing their contributions to form a smooth approximation of the underlying distribution [37]. In this study, we employed a Gaussian kernel, which is a symmetric kernel function centered at zero, under the assumption that the data originate from a smooth latent distribution.

A key challenge in using KDE is finding the appropriate bandwidth, which determines the smoothness of the density estimation. A narrow bandwidth captures more specific details but increases the risk of overfitting, whereas a broad bandwidth generates a smoother estimate that may oversimplify the data and obscure important structural information [35,37]. To address this, we adopted a cross-validated bandwidth selection strategy. Specifically, candidate values were drawn from the grid [0.1,0.2,0.5,1,2], and the optimal bandwidth for each feature was chosen by maximizing the training log-likelihood, ensuring stable and informative KDE-derived features.

After applying GMM to each feature, we further enriched the feature space using KDE to capture the local probability density of each data point. While GMM provides the probability that a data point belongs to a specific cluster (e.g., "low" or "high"-glucose clusters), KDE complements this by providing a non-parametric estimation of how typical or anomalous a sample is within the broader distribution of each feature. For example, a patient assigned to a "high-risk" GMM cluster for glucose levels might have a low KDE value, suggesting that their data value is in a sparsely populated region that likely indicates an outlier or rare clinical case.

Therefore, this hybrid approach that relies on both GMM and KDE to enrich the feature space can capture both subgroup membership and local uncertainty, which is particularly important in medical datasets.

2.2.2. Classification Frameworks

After enriching the feature space with GMM and KDE-based features, three classification frameworks were designed to evaluate the effectiveness of the proposed approach. The first framework trains the classification model on the concatenation of the original and enriched features to assess the direct contribution of feature augmentation. Building on this, two ensemble learning frameworks were introduced to further investigate the predictive value of the enriched feature space, as illustrated in Figure 4. The first ensemble employs a block-wise stacking strategy, in which base classifiers are trained on individual feature blocks (each original feature together with its corresponding GMM and KDE representations), and their outputs are combined using a meta-classifier. The second ensemble adopts a block-wise averaging scheme, where probability predictions from the same block-wise classifiers are aggregated through simple averaging to form the final decision. These three frameworks enable a systematic comparison between single-model learning and ensemble-based strategies, providing insights into the impact of feature enrichment on predictive performance.

https://doi.org/10.3390/a19020122

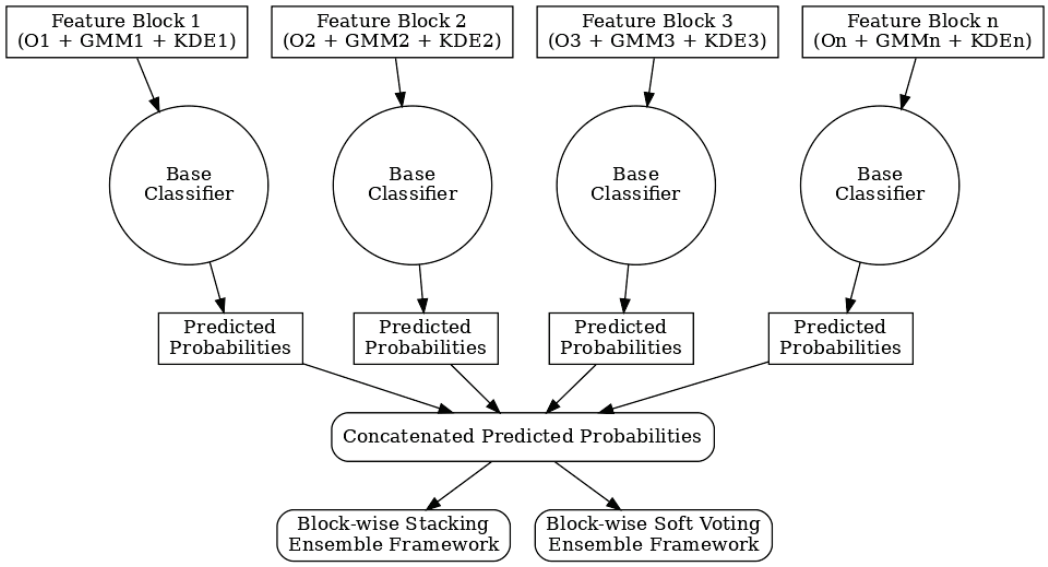


Figure 4. The block-wise ensemble learning framework for diabetes prediction. Each feature is first enriched with its corresponding GMM cluster probabilities and KDE log-density values, forming a feature block consisting of an original feature (O) together with its GMM and KDE representations. For each block, a base classifier is trained independently to generate predicted probabilities. These probability outputs are then concatenated to construct a meta-feature space. Two ensemble strategies are subsequently applied. In the block-wise stacking ensemble framework, the concatenated probability vectors serve as input to a meta-classifier for final prediction. In the block-wise soft voting ensemble framework, the average of the predicted probabilities across all feature blocks is used to determine the final classification.

2.3. Evaluation Metrics

To assess the performance of the classification models, particularly in the context of imbalanced datasets, we employ several evaluation metrics that capture both minority and majority class performance:

- **Accuracy** measures the proportion of correctly predicted cases to the total cases:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

  where TP (True Positive) represents cases in which the model correctly predicts diabetic instances, TN (True Negative) represents cases in which the model correctly predicts non-diabetic instances, FP (False Positive) represents cases in which the model incorrectly predicts diabetic instances, and FN (False Negative) represents cases in which the model incorrectly predicts non-diabetic instances.
  Although widely used, accuracy alone can be misleading in imbalanced datasets, as a model biased toward the majority class may achieve high accuracy while failing to detect minority (diabetic) cases.
- **Precision** measures the proportion of correctly predicted positive cases out of all cases predicted as positive:

$$\text{Precision} = \frac{TP}{TP + FP}$$

  It reflects the reliability of the model in minimizing false positives, which is crucial for avoiding unnecessary interventions.
- **Sensitivity (Recall or True Positive Rate)** indicates the proportion of actual positive cases that were correctly identified:

$$\text{Sensitivity} = \frac{TP}{TP + FN}$$

https://doi.org/10.3390/a19020122

- **Specificity (True Negative Rate)** measures the proportion of actual negative cases correctly identified:
$$\text{Specificity} = \frac{TN}{TN + FP}$$
- **F1-score** is the harmonic mean of precision and recall, providing a balance between the two:
$$\text{F1-score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$
It provides a balanced measure between precision and recall, particularly suitable for imbalanced datasets.
- **Area Under the ROC Curve (AUC)** evaluates the ability of the model to distinguish between the positive and negative classes across all classification thresholds. An AUC of 1.0 indicates perfect classification, while 0.5 represents random guessing.
- **Matthews Correlation Coefficient (MCC)** is a balanced measure that takes into account all four quadrants of the confusion matrix and is especially informative in imbalanced datasets:
$$\text{MCC} = \frac{(TP \cdot TN - FP \cdot FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$
MCC is especially informative for imbalanced datasets as it provides a balanced evaluation even when class sizes differ substantially.

In the context of diabetes prediction, the selected evaluation metrics provide complementary insights into clinical performance. Sensitivity is particularly important because false negatives may correspond to missed diabetes cases, delaying early intervention and increasing the risk of complications. Specificity helps limit false positives, thereby avoiding unnecessary diagnostic follow-up or treatment for non-diabetic individuals. AUC provides an overall measure of discriminative ability independent of decision thresholds, which is essential for clinical settings where risk stratification may vary. F1-score balances precision and recall under class imbalance, while MCC offers a robust, threshold-independent summary of model quality that accounts for all four outcomes in the confusion matrix. Together, these metrics enable a comprehensive and clinically relevant evaluation of model performance in both balanced and highly imbalanced diabetes prediction.

3. Experimental Results

This section presents the experimental setup and the results obtained using the enriched feature vectors generated by the proposed approach.

3.1. Experimental Setup

To evaluate the effectiveness of the proposed feature space enrichment framework, experiments were conducted on two publicly available diabetes datasets: the PIMA Indian Diabetes dataset and the Kaggle Diabetes Prediction dataset. Both datasets contain a combination of categorical and continuous variables, with a binary target variable indicating the presence or absence of diabetes.

Each dataset was randomly split into training (75%) and testing (25%) subsets using a stratified sampling strategy to preserve the original class distribution. The preprocessing pipeline included imputing physiologically implausible zero values (for selected PIMA features) using the median of the corresponding feature, applying the Box–Cox transformation to reduce skewness in non-categorical features, and standardizing all continuous variables to achieve zero mean and unit variance.

https://doi.org/10.3390/a19020122

For feature space enrichment, GMMs were applied independently to each transformed feature. The optimal number of mixture components was selected using the Bayesian Information Criterion (BIC), and the resulting soft cluster membership probabilities were incorporated as additional features. In parallel, KDE was applied to each feature, with the bandwidth automatically selected via cross-validation. The resulting log-density values were then appended to the feature set. Finally, the concatenation of the original and enriched features formed the extended feature space.

To address class imbalance in both datasets, a class-weighting strategy was employed. Specifically, the contribution of each class to the loss function was adjusted in proportion to its prevalence, such that misclassification of minority-class (diabetic) instances incurred a higher penalty. This approach enhances sensitivity to the minority class without introducing the drawbacks commonly associated with synthetic oversampling methods, such as data duplication or overfitting, while preserving the original data distribution.

Logistic Regression (LR) was used across all experimental setups to train the diabetes prediction models. It was selected for this study due to its simplicity and interpretability, and it demonstrated effectiveness in diabetes prediction, where it has been shown in many cases to outperform more complex models [28,38–47].

Hyperparameter tuning was conducted using grid search with stratified 5-fold cross-validation, repeated three times on the training data. Both L1 (Lasso) and L2 (Ridge) regularization penalties were evaluated, with the inverse regularization strength parameter (C) tuned over the values 0.01, 0.1, 1, and 10. The maximum number of iterations (max_iter) was set to 5000 to ensure model convergence.

3.2. Impact of Feature Enrichment on Prediction Performance

This section presents the experimental results of the proposed feature space enrichment framework evaluated on two diabetes datasets. The objective is to examine whether enriching the original feature space with GMM and KDE-based features can improve classification performance, particularly with respect to detecting minority-class (diabetic) cases.

For both datasets, Logistic Regression was used as the base classifier across all experimental setups. We evaluated three methodological configurations:

1. Baseline model: Trained on the original features only.
2. Feature-enriched model: Trained on the concatenation of original features with GMM-derived soft probabilities and KDE-based log-density estimates.
3. Ensemble models: Incorporating enriched feature blocks through two ensemble strategies, block-wise stacking with Logistic Regression as the meta-classifier and block-wise soft voting.

Each experimental configuration was evaluated using stratified 5-fold cross-validation repeated three times on the training set, followed by independent assessment on the held-out test set. Performance was reported using a comprehensive set of metrics appropriate for imbalanced classification tasks, including accuracy, precision, sensitivity (recall), specificity, F1-score, AUC, and the MCC.

To facilitate interpretation, the results are presented separately for each dataset.

3.2.1. PIMA Dataset Results

The performance of the proposed framework on the PIMA dataset was evaluated using both the baseline and enriched feature representations. As shown in Table 1, the baseline model, which trains an LR classifier using the original features, achieved an AUC of 0.85 on the validation set, which decreased to 0.79 on the independent test set. Similarly, the F1-score reached 0.68 on the validation set but declined to 0.62 on the test set. In terms of sensitivity, the model correctly identified 68% of diabetic cases. These results indicate

https://doi.org/10.3390/a19020122

that, although the baseline model exhibits reasonable predictive performance, the observed decrease from validation to test metrics highlights limitations in generalizability when relying solely on the original feature space.

When feature space enrichment using GMM and KDE was applied, the F1-score improved to 0.67 on the validation set and 0.68 on the test set, while the MCC increased from 0.38 to 0.49 on the test set. Although the improvement in AUC was modest, reaching 0.80 on the test set, the enriched representation consistently enhanced sensitivity, which is particularly important for identifying diabetic cases. Moreover, the consistent performance observed across the validation and test sets suggests that the enriched features contribute to improved model stability and generalizability.

The block-wise stacking ensemble achieved the strongest overall performance among all evaluated models. On the validation set, it attained an F1-score of 0.76 and an MCC of 0.52, representing the highest values across all approaches. On the test set, the stacking model maintained strong performance, achieving an F1-score of 0.68, an MCC of 0.47, and an AUC of 0.82, thereby outperforming both the baseline and the enriched LR models. This consistent performance indicates that the stacking strategy effectively integrates information from multiple feature blocks.

The block-wise soft voting ensemble also improved performance over the baseline, but its gains were less than stacking. On the test set, it reached an F1-score of 0.67 and MCC of 0.44, with an AUC of 0.81. While these results indicate moderate improvement, the soft voting framework did not leverage the feature-block structure as effectively as stacking.

Table 1. Performance comparison of baseline, feature-enriched, and ensemble models on the PIMA dataset.

Model	Set	Accuracy	AUC	Precision	Sensitivity	Specificity	F1-Score	MCC
Baseline (original features)	Validation	0.761559	0.84727	0.64028	0.721795	0.615385	0.676701	0.493382
	Test	0.697917	0.787204	0.566265	0.681159	0.707317	0.618421	0.376261
Feature-enriched model	Validation	0.750615	0.842844	0.620176	0.741880	0.692308	0.672399	0.482699
	Test	0.750000	0.799340	0.626506	0.753623	0.747967	0.684211	0.485819
Block-wise stacking ensemble	Validation	0.765625	0.848007	0.630081	0.778894	0.758621	0.758621	0.516730
	Test	0.739583	0.824673	0.609195	0.768116	0.723577	0.679487	0.473933
Block-wise soft voting ensemble	Validation	0.746528	0.834597	0.605578	0.763819	0.737401	0.675556	0.480677
	Test	0.713542	0.813228	0.572917	0.797101	0.666667	0.666667	0.445048

To evaluate whether the observed gains in sensitivity are achieved at the expense of increased false positive rates or reduced precision, we analyzed the trade-offs among evaluation metrics. Compared with the baseline model, the feature-enriched framework achieves higher sensitivity while maintaining comparable precision and specificity, as shown in Table 2. In particular, the increase in the true positive rate is not accompanied by a disproportionate rise in false positives, indicating a balanced precision–recall behavior. This trade-off is clinically favorable, as reducing false negatives (missed diabetic cases) is generally prioritized over a moderate increase in follow-up screening.

Table 2. Metric trade-off analysis on the PIMA dataset (test set).

Model	Precision	Sensitivity (Recall)	Specificity	FPR
Baseline (original features)	0.566265	0.681159	0.707317	0.292683
Feature-enriched model	0.626506	0.753623	0.747967	0.252033
Block-wise stacking ensemble	0.609195	0.768116	0.723577	0.276423
Block-wise soft voting ensemble	0.572917	0.797101	0.666667	0.333333

https://doi.org/10.3390/a19020122

Figure 5 presents the confusion matrices of the four models evaluated on the PIMA test set. The results show that the feature-enriched and ensemble configurations (stacking and soft voting) reduced the number of false negatives compared with the baseline model, indicating improved sensitivity to diabetic cases. In particular, the feature-enriched model and the stacking ensemble achieved a more balanced trade-off between correctly identifying diabetic and non-diabetic patients. These findings demonstrate the effectiveness of the proposed feature space enrichment in enhancing model generalization and reliability.

Analysis of the confusion matrices reveals a consistent reduction in false negatives with only a marginal change in false positives, indicating that the proposed enrichment framework improves the detection of diabetic cases without substantially increasing unnecessary clinical alarms.

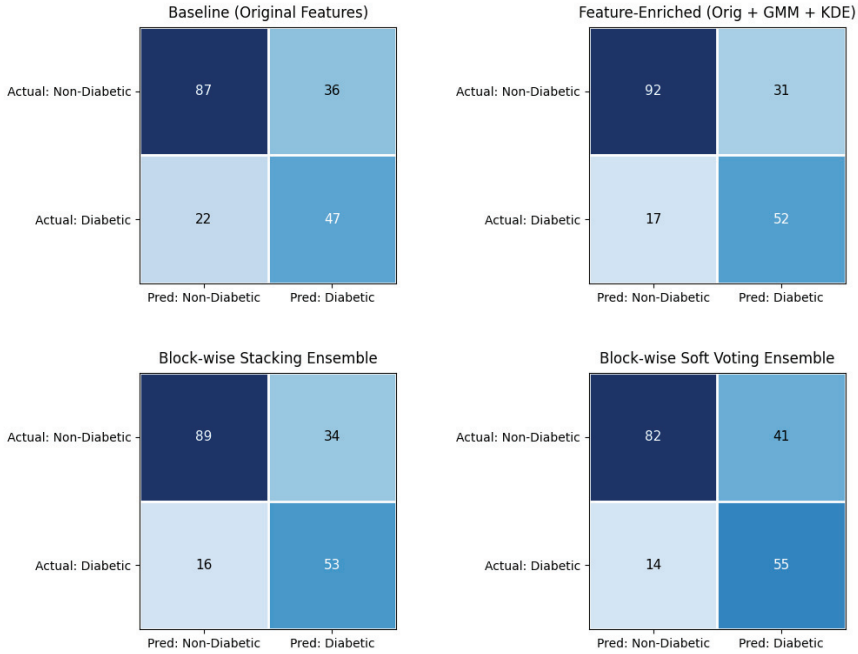


Figure 5. Confusion matrices of the LR models evaluated on the PIMA test set. Each matrix displays the distribution of correctly and incorrectly classified cases, with true diabetic and non-diabetic samples on the vertical axis (actual class) and predicted outcomes on the horizontal axis (predicted class). The top left and bottom right cells represent correctly identified non-diabetic (TN) and diabetic (TP) cases, respectively, while the off-diagonal cells correspond to misclassifications: FP and FN.

To assess robustness, we report the mean and standard deviation of performance metrics across stratified cross-validation folds, as shown in Table 3. The enriched model exhibits lower variability across folds while consistently improving sensitivity, indicating that the observed performance gains are stable and not driven by favorable data splits.

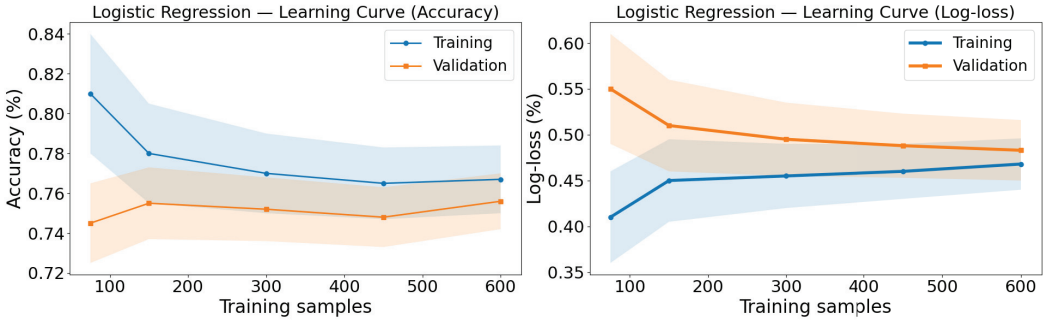
Table 3. Cross-validation performance variability (mean ± standard deviation) across test folds.

Model	Sensitivity (Mean ± std)	Precision (Mean ± std)	Specificity (Mean ± std)	MCC (Mean ± std)
Baseline	0.69 ± 0.05	0.57 ± 0.04	0.71 ± 0.03	0.38 ± 0.05
Feature-enriched	0.76 ± 0.03	0.63 ± 0.03	0.75 ± 0.03	0.48 ± 0.04

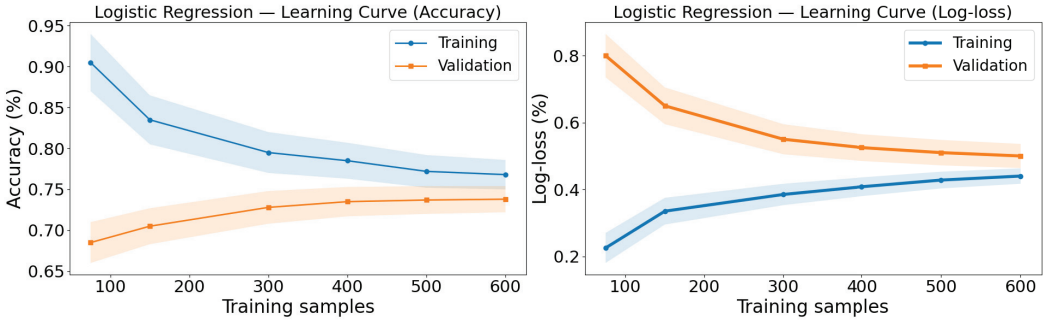
Figure 6 presents the learning curves for the LR models trained on the PIMA dataset under three configurations: baseline, feature-enriched, and block-wise stacking ensemble. The baseline model exhibits a noticeable gap between training and validation accuracy, indicating mild overfitting due to limited feature expressiveness. In contrast, the feature-enriched model demonstrates smoother convergence and reduced log-loss, reflecting

https://doi.org/10.3390/a19020122

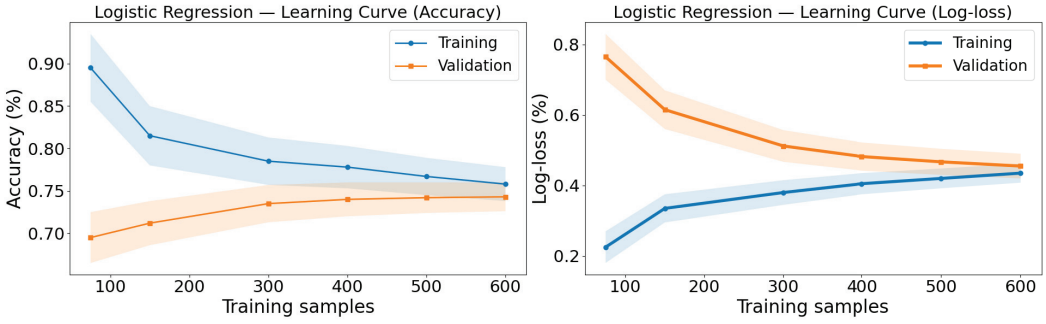
improved generalization. The stacking ensemble further stabilizes the learning process, achieving higher validation accuracy and lower loss across varying sample sizes. These learning curves confirm that the proposed enrichment strategy not only enhances predictive performance but also promotes more stable and efficient model training, thereby mitigating overfitting in small and imbalanced medical datasets.



(**a**) Baseline (original features) model.



(**b**) Feature-enriched model.



(**c**) Block-wise stacking ensemble model.

Figure 6. Learning curves for LR models trained on the PIMA dataset. Each panel shows the evolution of training and validation performance in terms of accuracy (**left**) and log-loss (**right**) as the number of training samples increases. (**a**) Baseline model using original features, (**b**) feature-enriched model incorporating GMM and KDE representations, and (**c**) block-wise stacking ensemble. The shaded regions represent the standard deviation over cross-validation folds. The curves demonstrate that feature enrichment and stacking improve convergence stability and reduce the generalization gap between training and validation performance.

Overall, as illustrated in Figure 7, the enriched features improve performance, stability, and generalization in the PIMA prediction task. In addition, the stacking ensemble achieves the highest AUC, making it the most effective configuration for this dataset.

https://doi.org/10.3390/a19020122

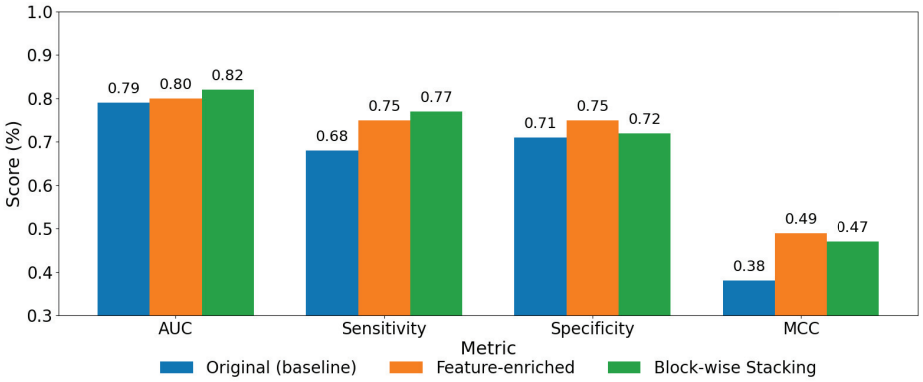


Figure 7. Comparison of test performance metrics for PIMA prediction across three approaches: baseline (original features), feature space enrichment (original + GMM + KDE), and block-wise stacking. The bar plots show values of AUC (%), Sensitivity (%), Specificity (%), and MCC (%). Results indicate that feature enrichment and stacking consistently improve the values relative to the baseline.

To evaluate the robustness of the probabilistic enrichment with respect to modeling choices, we conducted a sensitivity analysis by varying the number of Gaussian components (K) around the BIC-selected value for each feature. Table 4 reports the classification performance across all tested configurations. The results indicate that performance remains largely stable for different values of K, with variations in F1-score, AUC, and MCC generally confined to a narrow range (typically ≤ 0.02).

Notably, for features such as Pregnancies, SkinThickness, and DiabetesPedigreeFunction, performance remained nearly unchanged across all tested values of K, indicating that the mixture representations for these variables are inherently stable. In contrast, for Glucose, Insulin, and BloodPressure, slightly higher MCC and AUC values were observed at $K = 5$, $K = 3$, and $K = 4$, respectively. Moreover, BMI and Age benefited more substantially from higher values of K (4 and 10), reflecting greater heterogeneity in the underlying distributions of these clinical attributes.

The final selected numbers of components (last column in Table 4) coincide with the best or near-best performing configurations, confirming that BIC-based selection is not only theoretically well grounded but also empirically stable under perturbation.

Table 4. Sensitivity of enriched model performance to varying GMM components (K). The final selected K per feature (determined by BIC) is indicated in the last column.

Feature	K	F1-Score	AUC	MCC	Selected K (BIC)
Pregnancies	9	0.606	0.791	0.351	
Pregnancies	10	0.606	0.789	0.351	**9**
Pregnancies	11	0.606	0.790	0.351	
Glucose	3	0.599	0.791	0.335	
Glucose	4	0.620	0.797	0.370	**5**
Glucose	5	**0.625**	**0.796**	**0.376**	
BloodPressure	2	0.620	0.793	0.370	
BloodPressure	3	0.620	0.794	0.370	**4**
BloodPressure	4	**0.629**	0.789	**0.384**	
SkinThickness	1	0.620	0.793	0.370	
SkinThickness	2	**0.620**	**0.795**	**0.370**	**2**
SkinThickness	3	0.608	0.795	0.349	

Table 4. *Cont.*

Feature	K	F1-Score	AUC	MCC	Selected K (BIC)
Insulin	1	0.620	0.793	0.370	
Insulin	2	0.613	0.794	0.354	**3**
Insulin	3	**0.625**	**0.797**	**0.376**	
BMI	2	0.620	0.794	0.370	
BMI	3	0.620	0.797	0.370	**4**
BMI	4	**0.646**	**0.802**	**0.409**	
DiabetesPedigreeFunction	1	0.620	0.793	0.370	**1**
DiabetesPedigreeFunction	2	0.620	0.787	0.370	
Age	8	0.650	0.804	0.420	
Age	9	0.642	0.800	0.406	**10**
Age	10	**0.675**	**0.812**	**0.469**	

Note: Bold values indicate the best performance achieved for each feature across different numbers of GMM components. The selected *K* corresponds to the optimal number of components determined using the BIC.

In addition to the GMM component analysis, we evaluated the robustness of the KDE-based density features by varying the kernel bandwidth parameter within the range [0.5,2.0]. As summarized in Table 5, model performance remained nearly identical across all bandwidth settings, with no measurable change in F1-score and only negligible fluctuations in AUC and MCC (less than 0.002). These results indicate that the KDE-derived enriched features are highly stable and exhibit minimal sensitivity to the choice of bandwidth.

The selected bandwidth value of 1.0 (highlighted in the table) yields performance equivalent to or slightly better than the other tested settings. This stability indicates that KDE-based enrichment does not require precise hyperparameter tuning to achieve optimal performance and can generalize effectively without extensive parameter search, thereby enhancing reproducibility for future applications and external datasets.

Table 5. Impact of KDE bandwidth variation on enriched model performance.

Bandwidth	F1-Score	AUC	MCC
0.5	0.6203	0.7932	0.3704
1.0 (selected)	0.6203	0.7934	0.3704
1.5	0.6203	0.7938	0.3704
2.0	0.6203	0.7926	0.3704

The subgroup misclassification analysis presented in Table 6 shows that the proposed GMM–KDE enrichment improves prediction performance across most clinical intervals compared with the baseline model. The most notable gains are observed in borderline ranges, particularly for mid-level Glucose (+0.1087 and +0.1295), BMI (+0.1221 and +0.1043), and Age (+0.1292) subgroups, where class boundaries are inherently more ambiguous. These findings indicate that the probabilistic enrichment facilitates better discrimination among patients whose clinical profiles fall within overlapping distributions. In contrast, extreme values (e.g., very low or very high glucose levels and very high age) exhibit smaller improvements, suggesting that these cases remain more challenging and leave room for further methodological enhancements. Overall, the results confirm that the enriched feature space primarily strengthens discrimination in clinically ambiguous regions while maintaining stable performance across different subgroups.

https://doi.org/10.3390/a19020122

Table 6. Subgroup misclassification analysis comparing baseline and enriched model performance across clinical quartiles. The results show that GMM-KDE feature enrichment provides the largest improvement for borderline cases such as mid-range Glucose, BMI, and Age groups.

Feature	Group (Quartile)	Baseline F1	Enriched F1	Change
Glucose	Mid-Low	0.5000	0.6087	**+0.1087**
Glucose	Mid-High	0.6531	0.7826	**+0.1295**
Glucose	Low	0.7500	0.6667	-0.0833
Glucose	High	0.8457	0.8611	+0.0160
BMI	Mid-High	0.6800	0.7619	**+0.0819**
BMI	Mid-Low	0.6471	0.7692	**+0.1221**
BMI	Low	0.6957	0.8000	**+0.1043**
BMI	High	0.7459	0.8105	+0.0646
Age	High	0.8373	0.8533	+0.0160
Age	Mid-High	0.7500	0.8333	**+0.0833**
Age	Mid-Low	0.6400	0.7692	**+0.1292**
Age	Low	0.5833	0.6000	+0.0167

Note: Bold values indicate improved F1-scores of the enriched model compared to the baseline. The "Change" column reports (Enriched - Baseline), where positive values denote performance gains.

Given the limited sample size of the PIMA dataset and the use of probabilistic feature enrichment, we conducted an explicit robustness analysis using nested cross-validation to assess the risk of overfitting and optimistic bias. A stratified 5×5 nested cross-validation protocol was employed, in which the outer loop was used for unbiased performance estimation and the inner loop for model and hyperparameter selection. All preprocessing steps, including feature scaling, Box–Cox transformation, and GMM/KDE-based feature enrichment, were performed strictly within each outer training fold and subsequently applied to the corresponding validation and test partitions. This procedure ensured that no distributional information was shared across folds.

Table 7 reports the mean and standard deviation of performance metrics across the outer folds. The baseline model trained on the original features achieved an average AUC of 0.81 ± 0.03 and an MCC of 0.39 ± 0.05, while the enriched model achieved a higher AUC of 0.84 ± 0.02 and MCC of 0.48 ± 0.04. Importantly, the enriched representation also improved sensitivity (0.76 ± 0.05) compared to the baseline (0.69 ± 0.06), with comparable variability across folds.

Table 7. Nested cross-validation results on the PIMA dataset (mean $\pm$ std).

Model	AUC	MCC	Sensitivity	Specificity
Baseline (Original features)	0.81 ± 0.03	0.39 ± 0.05	0.69 ± 0.06	0.74 ± 0.04
Feature-enriched (GMM + KDE)	0.84 ± 0.02	0.48 ± 0.04	0.76 ± 0.05	0.75 ± 0.03

The relatively low standard deviations observed across all evaluation metrics indicate stable performance and demonstrate that the improvements achieved through probabilistic feature enrichment are consistent across different data splits, rather than being driven by a small number of favorable partitions. These results confirm that the proposed enrichment strategy enhances generalization on the PIMA dataset without inducing overfitting.

3.2.2. EHRs Dataset Results

The EHRs dataset is substantially larger than the PIMA dataset and exhibits pronounced class imbalance, with a significantly higher number of non-diabetic samples compared to diabetic cases. Despite this imbalance, Table 8 shows that the baseline LR model already demonstrates strong predictive capability, achieving AUC values of ap-

proximately 0.96 on both the validation and test sets. These results indicate excellent discrimination between diabetic and non-diabetic cases.

Introducing feature enrichment through GMM and KDE further improved performance relative to the baseline model. The results show that sensitivity increased from 0.88 to 0.91, enhancing the model's ability to correctly identify diabetic cases. This improvement was accompanied by increases in the F1-score (from 0.57 to 0.61) and MCC (from 0.56 to 0.61). These findings indicate that feature enrichment improves the balance between sensitivity and specificity while preserving overall accuracy and maintaining stable AUC performance.

Although the ensemble frameworks based on stacking and soft voting improved performance relative to the baseline model, the observed gains were modest. In particular, the stacking approach achieved a slight increase in AUC compared with the feature-enriched model, while no substantial improvements were observed across the remaining evaluation metrics.

Table 8. Performance comparison of baseline, feature-enriched, and ensemble models on the EHRs dataset.

Model	Set	Accuracy	AUC	Precision	Sensitivity	Specificity	F1-Score	MCC
Baseline (original features)	Validation	0.886764	0.961615	0.41951	0.877654	0.883103	0.567626	0.558029
	Test	0.887240	0.961073	0.424611	0.876279	0.888271	0.572036	0.561023
Feature-enriched model	Validation	0.905480	0.979433	0.47454	0.923256	0.903807	0.626875	0.621224
	Test	0.900600	0.973665	0.460635	0.911628	0.899562	0.612022	0.605172
Block-wise stacking ensemble	Validation	0.897693	0.988085	0.448477	0.906772	0.896854	0.600135	0.593931
	Test	0.897600	0.984330	0.452036	0.898605	0.897505	0.601494	0.592871
Block-wise soft voting ensemble	Validation	0.886320	0.963858	0.420106	0.900945	0.884967	0.573017	0.567550
	Test	0.887120	0.964192	0.425893	0.898141	0.886083	0.577798	0.570617

On the EHRs dataset, feature enrichment increases sensitivity while slightly reducing the false positive rate compared with the baseline model, as shown in Table 9, indicating improved detection of diabetic cases without an increase in false alarms.

Table 9. Metric trade-off analysis on the EHRs dataset (test set).

Model	Precision	Sensitivity (Recall)	Specificity	FPR
Baseline (original features)	0.424611	0.876279	0.888271	0.111729
Feature-enriched model	0.460635	0.911628	0.899562	0.100438
Block-wise stacking ensemble	0.452036	0.898605	0.897505	0.102495
Block-wise soft voting ensemble	0.425893	0.898141	0.886083	0.113917

Figure 8 displays the confusion matrices for the EHRs dataset. Across all implemented methodologies, the models exhibit strong discrimination between diabetic and non-diabetic samples, consistent with the high AUC values reported earlier. The feature-enriched and ensemble approaches (stacking and soft voting) further reduced the number of false negatives compared with the baseline model, correctly identifying a greater proportion of diabetic cases while maintaining low false-positive rates. These results confirm that the enriched feature representations enhance the model's ability to generalize and sustain balanced predictive performance, even in the presence of substantial class imbalance.

https://doi.org/10.3390/a19020122

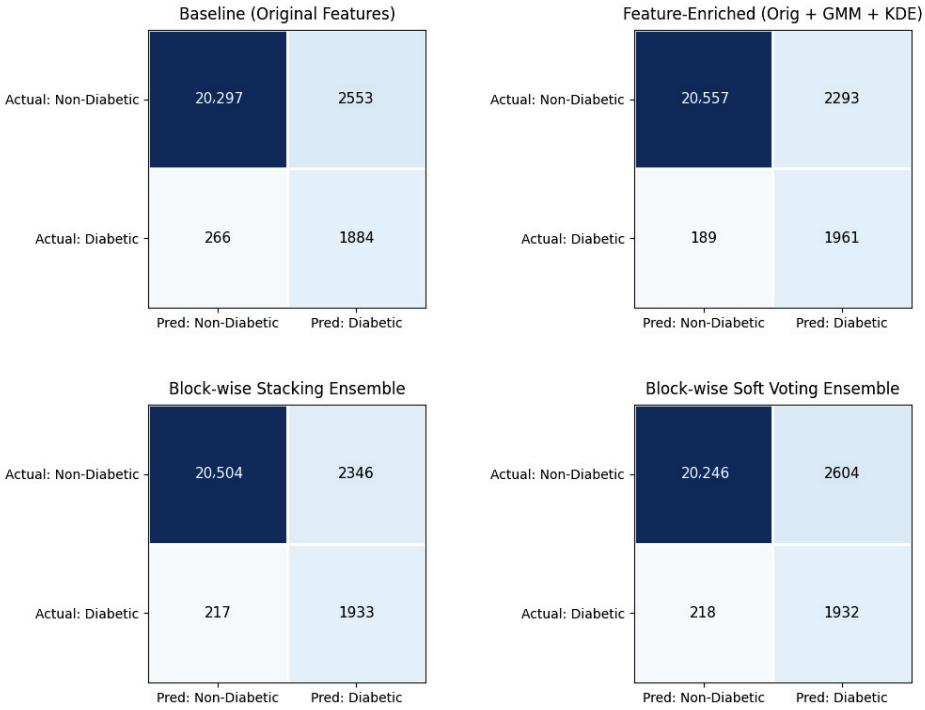
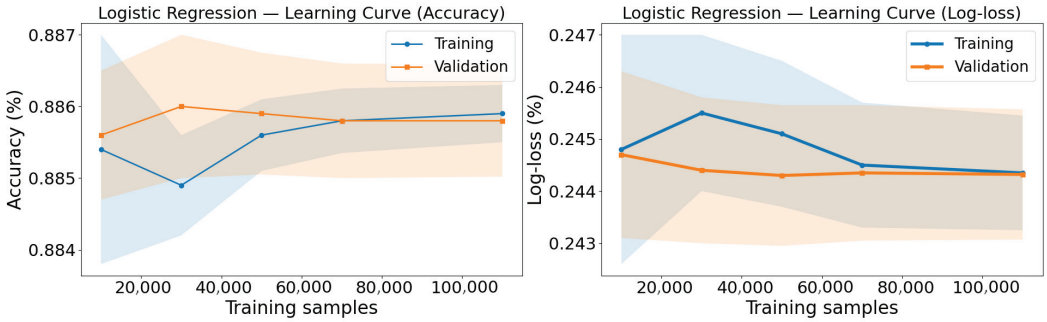


Figure 8. Confusion matrices of the LR models evaluated on the EHRs test set. Each matrix displays the distribution of correctly and incorrectly classified cases, with true diabetic and non-diabetic samples on the vertical axis (actual class) and predicted outcomes on the horizontal axis (predicted class). The top left and bottom right cells represent correctly identified non-diabetic (TN) and diabetic (TP) cases, respectively, while the off-diagonal cells correspond to misclassifications: FP and FN.

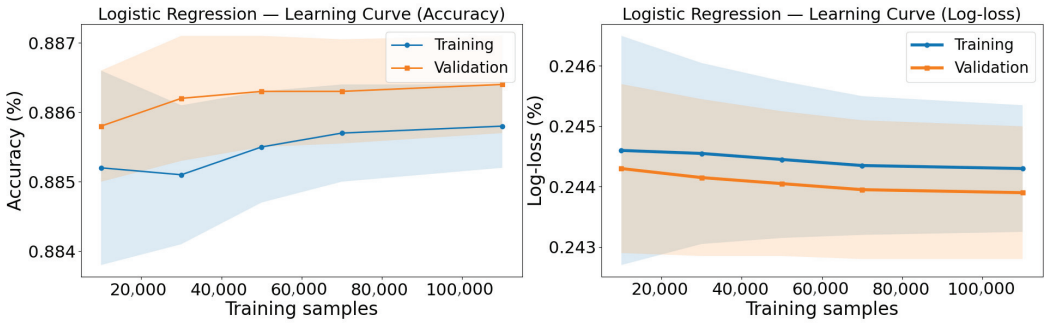
Figure 9 presents the learning curves of LR models for the EHRs dataset, comparing the baseline and feature-enriched configurations. Unlike the smaller PIMA dataset, both models exhibit rapid convergence and minimal divergence between training and validation curves, indicating well-generalized learning without evidence of overfitting. The stability of the log-loss across increasing sample sizes further confirms reliable optimization behavior and model consistency. The feature-enriched configuration demonstrates slightly higher accuracy and lower log-loss, suggesting that the GMM and KDE-derived features enable the model to capture finer probabilistic structure in the data. Overall, these learning curves highlight the robustness and scalability of the proposed enrichment framework when applied to large and heterogeneous medical datasets.

Overall, as illustrated in Figure 10, the feature-enriched model improves predictive performance, particularly by increasing sensitivity and enhancing the F1-score and MCC, while maintaining stable AUC and accuracy. In contrast, ensemble methods such as stacking and voting yield only marginal additional gains, achieving slight improvements in AUC but having minimal impact on the remaining evaluation metrics compared with the feature-enriched model.

https://doi.org/10.3390/a19020122



(**a**) Baseline (original features) model.



(**b**) Feature-enriched model.

Figure 9. Learning curves for LR models trained on the EHRs dataset. Each panel shows the evolution of training and validation performance in terms of accuracy (**left**) and log-loss (**right**) as the number of training samples increases. (**a**) Baseline model using original features, and (**b**) feature-enriched model incorporating GMM and KDE representations. The shaded regions represent the standard deviation over cross-validation folds. The curves demonstrate that feature enrichment improves convergence stability and reduces the generalization gap between training and validation performance.

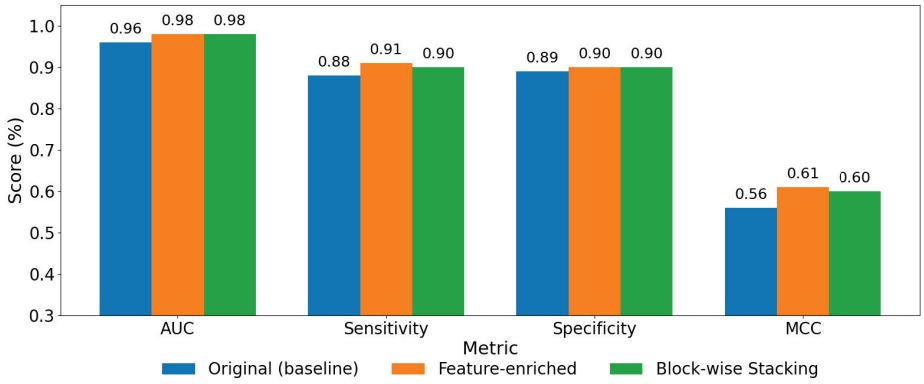


Figure 10. Comparison of test performance metrics for EHRs prediction across three approaches: baseline (original features), feature space enrichment (original + GMM + KDE), and block-wise stacking. The bar plots show values of AUC (%), Sensitivity (%), Specificity (%), and MCC (%). Results indicate that feature enrichment model provides the best results.

3.3. Interpretability and Feature Contribution Analysis

To enhance transparency and identify which enriched variables contribute most to diabetes prediction, a post hoc interpretability analysis was conducted using SHAP (SHapley Additive exPlanations) [48]. SHAP quantifies the marginal contribution of each feature to the model output and enables assessment of whether the performance gains achieved through enrichment arise from a small subset of highly informative probabilistic features or from a more distributed contribution across the enriched feature space.

https://doi.org/10.3390/a19020122

Figure 11 presents the mean absolute SHAP values for the Logistic Regression model trained on the enriched PIMA dataset. After feature scaling, the results indicate that *BMI* is the most influential predictor, followed closely by *Glucose*, which is consistent with established clinical knowledge. Importantly, several enriched variables appear prominently among the top contributors, including *Age_KDE*, *Age_GMM_Prob_5*, *BMI_GMM_Prob_1*, and *BMI_GMM_Prob_2*. This demonstrates that the enrichment process introduces meaningful latent representations rather than redundant transformations. In addition, multiple KDE-based features rank in the mid-to-lower importance range, indicating that density-based features contribute to prediction, albeit with lower relative impact compared to GMM-derived probabilities and primary clinical variables.

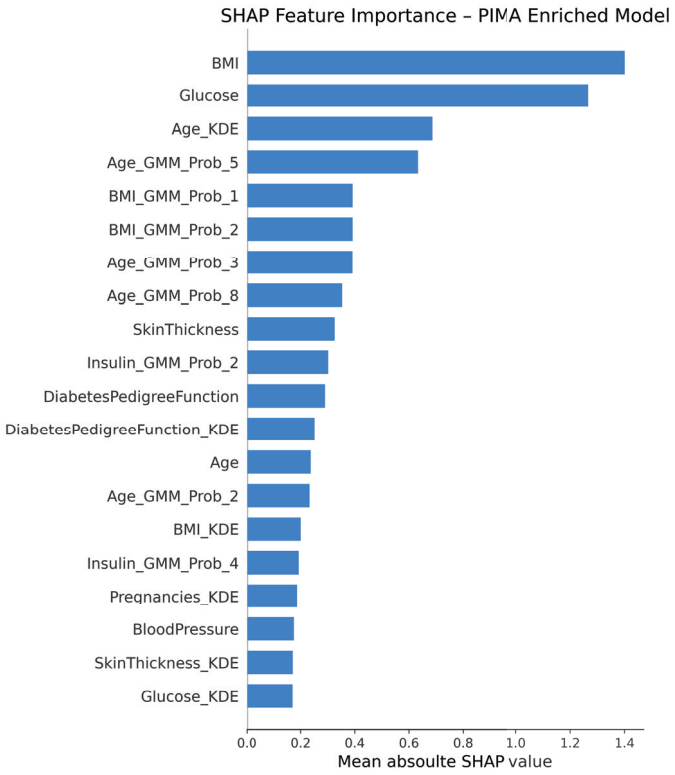


Figure 11. Mean absolute SHAP feature importance for the enriched PIMA model, showing contribution strength of original, GMM-derived, and KDE-derived features after scaling.

The SHAP beeswarm visualization in Figure 12 provides deeper insight into the directional effect of each feature. Each point represents a sample, with red indicating high feature values and blue representing low values. A clear positive trend is visible for both *BMI* and *Glucose*, where high values push predictions toward the diabetic class, matching real clinical behavior. Several enriched variables, particularly *Age_KDE* and GMM-based cluster memberships such as *Age_GMM_Prob_5* and *BMI_GMM_Prob_1*, also show structured and non-random contribution patterns, confirming that the enrichment framework captures latent metabolic subgroups and density-based deviations that influence classification. KDE features tend to cluster near zero with occasional strong positive impact for low–density regions, suggesting their role in detecting atypical or outlier-like risk profiles.

Overall, the SHAP analysis confirms that the enriched feature space contributes meaningfully to prediction rather than acting as a passive dimensional expansion. The integration of GMM and KDE-derived features alongside the original clinical variables results in a more expressive representation, leading to improved sensitivity and enhanced classification stability. Although some lower-ranked KDE and GMM features exhibit limited

https://doi.org/10.3390/a19020122

individual impact, they do not degrade performance, suggesting that targeted feature pruning or redundancy analysis could further optimize dimensionality in future work. This interpretability assessment corroborates the performance improvements reported in Section 3.2.1.



Figure 12. SHAP beeswarm summary for the enriched PIMA model. Color reflects feature magnitude (red = high; blue = low), while horizontal displacement indicates positive or negative effect on diabetes prediction.

4. Discussion

The experimental results obtained from both the PIMA and EHRs datasets demonstrate the effectiveness of the proposed feature space enrichment framework in improving diabetes prediction. The consistent performance gains observed across datasets of different sizes and class imbalance levels underscore the generalizability of the proposed approach.

On the smaller and more imbalanced PIMA dataset, the baseline LR model achieved only moderate performance, with noticeable reductions in generalization from the validation to the test sets. Incorporating GMM and KDE-derived features significantly improved both stability and sensitivity, confirming that feature space enhancement is particularly valuable when working with limited or weakly informative predictors. Furthermore, the block-wise stacking ensemble achieved the best overall performance, supporting prior findings that meta-learning approaches can more effectively integrate complementary feature information than simple averaging strategies [13,14].

In addition to quantitative performance evaluation, the SHAP-based interpretability analysis demonstrated that both original clinical variables and enriched probabilistic features contribute meaningfully to the final predictions. After feature scaling, *BMI* and *Glucose* emerged as the most influential predictors in the PIMA model, consistent with

established clinical knowledge. Importantly, several enriched features, including *Age_KDE*, *Age_GMM_Prob_5*, *BMI_GMM_Prob_1*, and *BMI_GMM_Prob_2*, exhibited substantial impact, confirming that the performance gains reported in Section 3.2.1 arise from the additional probabilistic structure introduced by GMM and KDE rather than from the original features alone. The SHAP beeswarm visualization further revealed interpretable patterns: higher values of *BMI* and *Glucose* consistently shifted predictions toward the diabetic class, while KDE-based features frequently highlighted low-density patient profiles associated with atypical or high-risk metabolic conditions. Collectively, these findings indicate that the enriched features provide complementary information beyond traditional clinical indicators, enhancing model transparency and reinforcing the clinical relevance of the proposed enrichment framework.

Because the PIMA dataset contains a relatively small number of samples, we carefully addressed the risk of overfitting and potential information leakage during probabilistic feature construction. All GMM and KDE models were fitted exclusively on the training folds within the cross-validation pipeline, and enriched features for validation and test instances were generated using models learned solely from their corresponding training subsets. This procedure prevents the density estimators from accessing information outside their respective folds and avoids leakage of distributional patterns. Furthermore, the use of stratified repeated 5-fold cross-validation, together with regularized classifiers and class-weighting, reduces variance and helps preserve model generalization.

In contrast, on the substantially larger EHRs dataset, where the baseline model already demonstrated strong discriminative performance, the proposed enrichment yielded incremental yet consistent improvements in sensitivity, F1-score, and MCC without compromising AUC or specificity. These findings indicate that the proposed framework not only generalizes effectively to large-scale datasets but also provides measurable benefits in identifying minority diabetic cases, even when overall predictive performance is already high.

Although the enriched features are mathematically derived, they offer intuitive interpretations in clinical settings. GMM-based probabilities reflect the degree to which a patient belongs to latent subgroups characterized by similar patterns in variables such as glucose, BMI, or HbA1c. These subgroups may correspond to clinically meaningful phenotypes, including patients exhibiting early metabolic dysregulation or those with more stable metabolic profiles. In parallel, KDE-derived density values indicate how typical or atypical a patient's measurements are relative to the broader population. Low-density values may signal rare or extreme combinations of risk factors, which are often associated with elevated clinical concern. Taken together, these probabilistic features help identify clusters of patients with shared characteristics while also highlighting individuals who reside in high-risk or unusual regions of the feature space, thereby enhancing clinical interpretability and supporting more informed decision-making.

The proposed framework introduces additional preprocessing through GMM and KDE modeling; however, the associated computational cost remains modest, as both models are fitted independently for each feature rather than on the full multivariate space. This per-feature design results in computational complexity that scales approximately linearly with the number of variables, enabling efficient feature enrichment even for large datasets. Moreover, feature enrichment is performed only once during model training, while deployment in a clinical environment requires evaluating mixture probabilities and KDE values for new patient records, followed by logistic regression inference. These operations are computationally lightweight and can be executed in real time, making the proposed framework practical for routine clinical use.

https://doi.org/10.3390/a19020122

The proposed GMM–KDE enrichment framework demonstrates consistent performance gains; however, this study focuses exclusively on probabilistic feature representations. Other feature space transformation strategies, such as Principal Component Analysis (PCA) [49], Independent Component Analysis (ICA) [50], autoencoder-based embeddings [51], or hybrid pipelines combining statistical and deep representations, may yield competitive or complementary improvements. Consequently, the observed performance enhancement should not be interpreted as unique to GMM and KDE-derived features, and systematic benchmarking against alternative enrichment techniques remains an important direction for future work.

A second limitation concerns the scope of the feature relevance assessment. While SHAP analysis was conducted and demonstrated that enriched probabilistic features contribute meaningfully to the prediction process, the study does not evaluate whether all generated GMM/KDE components are equally informative or whether some may be redundant or collinear with the original clinical predictors. Future work should therefore extend the relevance analysis toward redundancy-aware feature selection, for example, by using mRMR [52,53], ReliefF [54], or sparsity-inducing models such as LASSO or ElasticNet [55], to retain only the most discriminative representations and avoid unnecessary dimensionality inflation. In addition, projection-based relevance evaluation may help assess whether the enriched features meaningfully separate diabetic and non-diabetic cases.

5. Conclusions

This study introduces a probabilistic feature enrichment framework for early diabetes prediction by integrating GMM and KDE into the feature representation process. By augmenting the original clinical variables with probabilistic descriptors, the proposed framework captures latent distributional characteristics that may not be fully represented by conventional features, thereby improving the discriminative capability of standard classifiers without increasing model complexity. Comprehensive experiments conducted on two benchmark datasets—the smaller and more imbalanced PIMA dataset and the larger EHRs dataset—demonstrate the robustness of the proposed approach across heterogeneous data settings. On the PIMA dataset, feature enrichment led to substantial improvements in sensitivity and generalization performance, underscoring its effectiveness in scenarios with limited or weakly informative predictors. On the EHRs dataset, where the baseline model already exhibited strong discriminative performance, the proposed enrichment consistently improved sensitivity, F1-score, and MCC, confirming its ability to enhance predictive balance in highly imbalanced clinical data. Importantly, the use of class-weighted learning enabled improved detection of minority diabetic cases without the drawbacks associated with synthetic oversampling techniques.

Overall, the findings highlight the clinical significance of probabilistic feature enrichment, as improved sensitivity translates into fewer false negatives and a reduced risk of missed diabetes diagnoses—an essential factor for early intervention and the prevention of complications. Future work will extend this framework to more advanced classifiers, such as Support Vector Machines (SVMs) and XGBoost, and will further evaluate its applicability across broader healthcare prediction tasks.

Future work will focus on refining the enriched feature space through feature attribution and relevance–redundancy analysis to identify the most informative probabilistic components and reduce dimensionality without sacrificing discriminative performance. In addition, extending the framework to multi-class prediction settings, performing case-level and subgroup-specific error analyses, and exploring alternative or hybrid enrichment strategies, such as PCA, ICA, and autoencoder-based representations—represent promising directions for future research.

https://doi.org/10.3390/a19020122

Author Contributions: Conceptualization, H.A. and S.Q.; methodology, H.A. and Y.Q.; validation, Y.Q. and M.H.; formal analysis, H.A. and Y.Q.; investigation, H.A.; writing—original draft preparation, H.A.; writing—review and editing, S.Q. and M.H.; visualization, H.A. and M.H.; supervision, S.Q.; project administration, S.Q. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The dataset analyzed in this study is publicly available from Kaggle (https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database, accessed on 29 June 2025 and https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset, accessed on 29 June 2025). All code used for preprocessing, probabilistic feature enrichment (GMM and KDE), and classification experiments is available from the corresponding author upon reasonable request.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Sun, H.; Saeedi, P.; Karuranga, S.; Pinkepank, M.; Ogurtsova, K.; Duncan, B.B.; Stein, C.; Basit, A.; Chan, J.C.; Mbanya, J.C.; et al. IDF Diabetes Atlas: Global, regional and country-level diabetes prevalence estimates for 2021 and projections for 2045. *Diabetes Res. Clin. Pract.* **2022**, *183*, 109119. [CrossRef] [PubMed]
2. International Diabetes Federation. *IDF Diabetes Atlas*, 11th ed.; Technical Report; International Diabetes Federation: Brussels, Belgium, 2025.
3. World Health Organization. Diabetes Fact Sheet. 2024. Available online: https://www.who.int/news-room/fact-sheets/detail/diabetes (accessed on 29 June 2025).
4. Chatterjee, S.; Khunti, K.; Davies, M.J. *Type 2 Diabetes Mellitus*; StatPearls Publishing: Treasure Island, FL, USA, 2024. Available online: https://www.ncbi.nlm.nih.gov/books/NBK513253/ (accessed on 24 June 2025).
5. Campos, C.d.S. Chronic hyperglycemia and glucose toxicity: Pathology and clinical sequelae. *Postgrad. Med.* **2012**, *124*, 90–97. [CrossRef] [PubMed]
6. Tabák, A.G.; Herder, C.; Rathmann, W.; Brunner, E.J.; Kivimäki, M. Prediabetes: A high-risk state for diabetes development. *Lancet* **2012**, *379*, 2279–2290. [CrossRef] [PubMed]
7. Shailaja, K.; Seetharamulu, B.; Jabbar, M.A. Machine Learning in Healthcare: A Review. In Proceedings of the 2018 International Conference on Electronics, Communication and Aerospace Technology (ICECA), Coimbatore, India, 29–31 March 2018; pp. 910–914. [CrossRef]
8. Kavakiotis, I.; Tsave, O.; Salifoglou, A.; Maglaveras, N.; Vlahavas, I.; Chouvarda, I. Machine Learning and Data Mining Methods in Diabetes Research. *Comput. Struct. Biotechnol. J.* **2017**, *15*, 104–116. [CrossRef]
9. Toleva, B.; Atanasov, I.; Ivanov, I.; Hooper, V. An Effective Methodology for Diabetes Prediction in the Case of Class Imbalance. *Bioengineering* **2025**, *12*, 35. [CrossRef]
10. Sharma, A.; Guleria, K.; Goyal, N. Prediction of Diabetes Disease Using Machine Learning Model. In *Proceedings of the International Conference on Communication, Computing and Electronics Systems. Lecture Notes in Electrical Engineering*; Springer: Singapore, 2021; Volume 768, pp. 551–558. [CrossRef]
11. Ağraz, M.; Eğrioğlu, E.; Baş, E.; Chen, M.; Göksülük, D.; Furkan Burak, M. Diabetes development prediction using a hybrid model combining dendritic artificial neuron model and logistic regression. *Endocrinol. Res. Pract.* **2025**, *29*, 84–93. [CrossRef]
12. Ahmed, A.; Khan, J.; Arsalan, M.; Ahmed, K.; Shahat, A.A.; Alhalmi, A.; Naaz, S. Machine Learning Algorithm-Based Prediction of Diabetes Among Female Population Using PIMA Dataset. *Healthcare* **2024**, *13*, 37. [CrossRef]
13. Reza, M.; Islam, M.R.; Das, N.; Hoque, M.N.; Rahman, M.S.; Rahman, M.M. Improving diabetes disease patients classification using stacking ensemble method with PIMA and local healthcare data. *Heliyon* **2024**, *10*, e24536. [CrossRef]
14. Sivashankari, M.; Vengatesan, K. Diabetes prediction using machine learning algorithms. *Mater. Today Proc.* **2022**, *51*, 382–386. [CrossRef]
15. Tripathi, R.P.; Sharma, M.; Gupta, A.K.; Pandey, D.; Pandey, B.K.; Shahul, A.; George, A.S.H. Timely prediction of diabetes by means of machine learning practices. *Augment. Hum. Res.* **2023**, *8*, 1. [CrossRef]
16. Shi, J.; Yu, G.; Zhang, C.; Han, L.; Li, X.; Xu, W.; Zhang, L. Integrated Bagging–Random Forest learning model for diabetes prediction. *PeerJ Comput. Sci.* **2024**, *10*, e2302. [CrossRef] [PubMed]
17. Talukder, M.A.; Islam, M.M.; Uddin, M.A.; Kazi, M.K.I.; Khalid, M.; Akhter, A.; Chowdhury, S.Z.; Kim, J.; Moni, M.A. Toward reliable diabetes prediction: Innovations in data engineering and machine learning applications. *Digit. Health* **2024**, *10*, 20552076241271867. [CrossRef] [PubMed]

18. Uddin, M.A.; Islam, M.M.; Talukder, M.A.; Hossain, M.A.A.; Akhter, A.; Aryal, S.; Muntaha, M. Machine Learning Based Diabetes Detection Model for False Negative Reduction. *Biomed. Mater. Devices* **2024**, *2*, 427–443. [CrossRef]
19. Qawqzeh, Y.K. Neural Network-based Diabetic Type II High-Risk Prediction using Photoplethysmogram Waveform Analysis. *Int. J. Adv. Comput. Sci. Appl.* **2019**, *10*. [CrossRef]
20. Suwalska, A.; Polańska, J. GMM-Based Expanded Feature Space as a Way to Extract Useful Information for Rare Cell Subtypes Identification in Single-Cell Mass Cytometry. *Int. J. Mol. Sci.* **2023**, *24*, 14033. [CrossRef]
21. Ghaderi, P.; Nosouhi, M.; Jordanic, M.; Marateb, H.R.; Mañanas, M.A.; Farina, D. Kernel Density Estimation of Electromyographic Signals and Ensemble Learning for Highly Accurate Classification of a Large Set of Hand/Wrist Motions. *Front. Neurosci.* **2022**, *16*, 795943. [CrossRef]
22. Chushig-Muzo, D.; Soguero-Ruiz, C.; de Miguel-Bohoyo, P.; Mora-Jiménez, I. Interpreting clinical latent representations using autoencoders and probabilistic models. *Comput. Methods Programs Biomed.* **2021**, *207*, 106181. [CrossRef]
23. Smith, J.; Everhart, J.; Dickson, W.; Knowler, W.; Johannes, R. Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. In Proceedings of the Annual Symposium on Computer Application in Medical Care, Orlando, FL, USA, 7–11 November 1988; pp. 261–265.
24. Santhanam, T.; Padmavathi, M.S. Comparison of K-Means clustering and statistical outliers in reducing medical datasets. In *Proceedings of the 2015 International Conference on Computing and Network Communications (CoCoNet)*; IEEE: Piscataway, NJ, USA, 2015; pp. 1–6.
25. Padmavathi, M.S.; Sumathi, C.P. A New Method of Data Preparation for Classifying Diabetes Dataset. *Indian J. Sci. Technol.* **2019**, *12*, 1–9. [CrossRef]
26. Asha, V.; Lamani, M.R.; Padmaja, K.; Gondhale, T.A. Enhancing diabetes prediction accuracy with AdvanSVM: A machine learning approach using the PIMA dataset. *Seybold Rep. J.* **2024**, *19*, 51–72.
27. Idris, N.; Ismail, M.A.; Jaya, I.; Osman, A.; Abulfaraj, A.; Binzagr, F. Stacking with Recursive Feature Elimination-Isolation Forest for classification of diabetes mellitus. *PLoS ONE* **2024**, *19*, e0302595. [CrossRef]
28. Xia, Y. A Machine Learning-Based Prediction Study for Type 2 Diabetic Mellitus. *Sci. Technol. Eng. Chem. Environ. Prot.* **2024**, *1*. [CrossRef]
29. Deng, T.; Luo, W.; Huang, K. Identifying Key Factors that Influence Diabetes Prediction: A Meta Analysis of Two Datasets and Three Machine Learning Models. *Appl. Comput. Eng.* **2025**, *132*, 200–212. [CrossRef]
30. Maza, D.; Ojo, J.O.; Akinlade, G.O. A predictive machine learning framework for diabetes. *Turk. J. Eng.* **2024**, *8*, 583–592. [CrossRef]
31. Mustafa, M. Diabetes Prediction Dataset. 2022. Available online: https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset (accessed on 24 May 2025).
32. Reynolds, D.A. Gaussian mixture models. In *Encyclopedia of Biometrics*; Springer: Boston, MA, USA, 2009; pp. 659–663.
33. Zeng, H.; Cheung, Y.M. A new feature selection method for Gaussian mixture clustering. *Pattern Recognit.* **2009**, *42*, 243–250. [CrossRef]
34. Zhou, B.; Hansen, J.H. Unsupervised audio stream segmentation and clustering via the Bayesian information criterion. In Proceedings of the INTERSPEECH, Beijing, China, 16–20 October 2000; pp. 714–717.
35. Peter, D.H. Kernel estimation of a distribution function. *Commun. Stat.-Theory Methods* **1985**, *14*, 605–620. [CrossRef]
36. Gidel, S.; Blanc, C.; Chateau, T.; Checchin, P.; Trassoudaine, L. Comparison between GMM and KDE Data Fusion methods for particle filtering: Application to pedestrian detection from laser and video measurements. In Proceedings of the 13th International Conference on Information Fusion, Edinburgh, UK, 26–29 July 2010; pp. 1–7.
37. Weglarczyk, S. Kernel density estimation and its application. *ITM Web of Conf.* **2018**, *23*, 00037. [CrossRef]
38. Shakya, S.; Ceh-Varela, E.; Maes, L. Machine Learning Analysis of Factors Contributing to Diabetes Development. *Cloud Comput. Data Sci.* **2024**, *5*, 157–182. [CrossRef]
39. Rajput, M.; Khedgikar, S. Diabetes prediction and analysis using medical attributes: A Machine learning approach. *J. Xi'an Univ. Archit. Technol.* **2022**, *14*, 98–103. [CrossRef]
40. Mousa, G.; Abu-Hassan, H.; Al-Rimmawi, H. Prediction of Type 2 Diabetes using logistic regression techniques. *Turk. J. Comput. Math. Educ.* **2024**, *15*, 1–13.
41. Tigga, N.; Garg, S. Predicting Type 2 Diabetes Using Logistic Regression. In *Fourth International Conference on Microelectronics, Computing and Communication Systems*; Springer: Ranchi, India, 2019; pp. 491–500.
42. Joshi, R.; Dhakal, C. Predicting Type 2 Diabetes Using Logistic Regression and Machine Learning Approaches. *Int. J. Environ. Res. Public Health* **2021**, *18*, 7346. [CrossRef]
43. Rajendra, P.; Latifi, S. Prediction of diabetes using logistic regression and ensemble techniques. *Comput. Methods Programs Biomed. Update* **2021**, *1*, 100032. [CrossRef]
44. Pradhan, A. A Comprehensive Analysis of Diabetes Risk Prediction Using Logistic Regression. *Int. J. Comput. Trends Technol. (IJCTT)* **2024**, *72*, 192–219. [CrossRef]

45. Lynam, A.L.; Dennis, J.M.; Owen, K.R.; Oram, R.A.; Jones, A.G.; Shields, B.M.; Ferrat, L.A. Logistic regression has similar performance to optimised machine learning algorithms in a clinical setting: Application to the discrimination between type 1 and type 2 diabetes in young adults. *Diagn. Progn. Res.* **2020**, *4*, 6. [CrossRef] [PubMed]
46. Nguyen, B.; Zhang, Y. A Comparative Study of Diabetes Prediction Based on Lifestyle Factors Using Machine Learning. *arXiv* **2025**, arXiv:2503.04137. [CrossRef]
47. Noh, J.H.; Kim, J.H. Development of Causal Discovery-Based Predictive Models for Type 2 Diabetes Mellitus Using a Nationwide Health Screening Dataset. *Biomedicines* **2025**, *13*, 490. [CrossRef]
48. Lundberg, S.M.; Lee, S.I. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*; Curran Associates Inc.: Red Hook, NY, USA, 2017; pp. 4768–4777.
49. Pechenizkiy, M.; Tsymbal, A.; Puuronen, S. PCA-based feature transformation for classification: Issues in medical diagnostics. In Proceedings of the 17th IEEE Symposium on Computer-Based Medical Systems, Bethesda, MD, USA, 25 June 2004; Volume 17, pp. 535–540. [CrossRef]
50. Kwak, N.; Choi, C.H. Feature Extraction Based on ICA for Binary Classification Problems. *IEEE Trans. Knowl. Data Eng.* **2003**, *15*, 1374–1388. [CrossRef]
51. Li, H.; Feng, A.; Lin, B.; Su, H.; Liu, Z.; Duan, X.; Pu, H.; Wang, Y. A novel method for credit scoring based on feature transformation and ensemble model. *PeerJ Comput. Sci.* **2021**, *7*, e579. [CrossRef]
52. Mary, P.; Nagarajan, R. Feature Selection in Microarray using Proposed Hybrid Minimum Redundancy-Maximum Relevance (MRMR) and Modified Genetic Algorithm (MGA). *Int. J. Exp. Res. Rev.* **2024**, *39*, 82–91. [CrossRef]
53. Wang, G.; Lauri, F.; Hajjam, A. Feature Selection by mRMR Method for Heart Disease Diagnosis. *IEEE Access* **2022**, *10*, 100786–100796. [CrossRef]
54. Li, J.; Liu, H. Feature Selection and Classification Model Construction on Type 2 Diabetic Patients' Data. In *Proceedings of the 4th International Conference on Advances in Data Mining*; Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany, 2004; Volume 3275, pp. 349–364. [CrossRef]
55. Maeda-Gutiérrez, V.; Galván Tejada, C.; Galván Tejada, J.; Cruz, M.; Celaya Padilla, J.; Gamboa-Rosales, H.; García, A.; Luna-Garcia, H.; Villalba, K. Evaluating Feature Selection Methods for Accurate Diagnosis of Diabetic Kidney Disease. *Biomedicines* **2024**, *12*, 2858. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

algorithms

MDPI

Article

DenseNet-Based Classification of EEG Abnormalities Using Spectrograms

Lan Wei and Catherine Mooney *

FutureNeuro Research Ireland Centre, UCD School of Computer Science, University College Dublin, D04C1P1 Dublin, Ireland; lan.wei@ucd.ie
* Correspondence: catherine.mooney@ucd.ie

Abstract

Electroencephalogram (EEG) analysis is essential for diagnosing neurological disorders but typically requires expert interpretation and significant time. Purpose: This study aims to automate the classification of normal and abnormal EEG recordings to support clinical diagnosis and reduce manual workload. Automating the initial screening of EEGs can help clinicians quickly identify potential neurological abnormalities, enabling timely intervention and guiding further diagnostic and treatment strategies. Methodology: We utilized the Temple University Hospital EEG dataset to develop a DenseNet-based deep learning model. To enable a fair comparison of different EEG representations, we used three input types: signal images, spectrograms, and scalograms. To reduce dimensionality and simplify computation, we focused on two channels: T5 and O1. For interpretability, we applied Local Interpretable Model-agnostic Explanations (LIME) and Gradient-weighted Class Activation Mapping (Grad-CAM) to visualize the EEG regions influencing the model's predictions. Key Findings: Among the input types, spectrogram-based representations achieved the highest classification accuracy, indicating that time-frequency features are especially effective for this task. The model demonstrated strong performance overall, and the integration of LIME and Grad-CAM provided transparent explanations of its decisions, enhancing interpretability. This approach offers a practical and interpretable solution for automated EEG screening, contributing to more efficient clinical workflows and better understanding of complex neurological conditions.

Keywords: EEG; spectrogram; DenseNet121; explainable; automated classification

1. Introduction

Electroencephalogram (EEG) is a important tool that assists in the analysis of various neurological disorders, including epilepsy, Parkinson's disease, and Alzheimer's disease [1]. By recording electrical activity in the brain, EEG provides valuable insights into the brain's functioning and helps in the diagnosis and monitoring of these conditions. The classification of EEG signals into normal and abnormal patterns is a fundamental first step in this analysis [2], as it enables clinicians to identify potential abnormalities that may indicate underlying neurological issues. This classification not only facilitates timely intervention but also guides treatment decisions, ultimately improving patient outcomes and advancing our understanding of these complex diseases.

Previous research on EEG-based event classification methods has relied on estimating features from the time, frequency, and time–frequency domains, which are then used as inputs for machine learning algorithms [3,4]. This approach requires a solid background in

signal processing, and the performance of these models heavily depends on the quality of the estimated features.

In response, deep learning methods have been introduced to classify EEG events without the need for manual feature extraction [5–7]. However, these techniques require large EEG datasets for training to achieve high performance. Additionally, complex deep-learning architectures demand substantial training time and computational resources [8]. Transfer learning has emerged as a viable solution, allowing researchers to leverage pre-trained weights to address these challenges [9]. Consequently, more researchers [10–12] are adopting transfer learning methods to analyze EEG events.

Despite advancements in artificial intelligence techniques for EEG analysis that enhance performance, the interpretability of these methods has become increasingly challenging [13]. Explainability is particularly crucial in the medical domain, where understanding model decisions can significantly impact patient care [14]. Techniques such as SHapley Additive exPlanations (SHAP) [15], Local Interpretable Model-agnostic Explanations (LIME) [16] and Gradient-weighted Class Activation Mapping (Grad-CAM) [17] have been proposed to improve the interpretability of these models and enhance our understanding of their decision-making processes.

The Temple University Hospital Abnormal EEG Dataset (TUAB) [18] was created from archived records at Temple University Hospital and is regarded as the largest publicly available collection of clinical EEG recordings worldwide [19]. The TUAB dataset has been annotated as either normal or abnormal and has been widely utilized in various studies [20–25] to develop methods for classifying abnormal EEG signals.

Channels T5 and O1 are often chosen in TUAB EEG tasks [20,23], because they provide critical information about temporal and occipital brain activity, regions commonly involved in pathological conditions like epilepsy and encephalopathy [19]. Channel T5 located in the left temporal lobe, is sensitive to abnormalities such as spikes and sharp waves [26], while channel O1, in the left occipital region, detects patterns like occipital slowing or periodic discharges. These channels are less prone to artifacts from eye movements compared to frontal channels [27,28], making them ideal for identifying abnormalities with higher sensitivity and signal quality. Therefore, channels T5 and O1 were chosen as inputs in this study for classifying normal and abnormal EEGs in the TUAB dataset [20,23].

Signal images, spectrograms, and scalograms are commonly used for EEG event analysis, providing insights from various domains. To ensure a fair comparison among these EEG representations, we evaluated their performance under consistent model settings. To improve performance with limited data and to reduce computational costs, we used these representations as input to a DenseNet transfer learning-based strategy, with post-processing applied across multiple images. To enhance model interpretability, we used LIME and Grad-CAM techniques, which visualize the regions of the input data most influential to the predictions, enabling researchers to gain a clearer understanding of the model's decision-making process. Using only channels T5 and O1 of the TUAB EEG data our method obtained results comparable to multi-channel models for classifying normal and abnormal EEG signals. The main contributions of this study are as follows:

- Presents a transfer learning model for automated classification of EEG recordings as normal or abnormal, enabling high-throughput screening in clinical workflows. This facilitates early identification of potential neurological abnormalities, allowing clinicians to focus on abnormal EEGs, support timely intervention, and improve diagnostic efficiency and patient outcomes.
- Presents an EEG classification approach that achieves competitive performance using only two channels (T5 and O1), thereby reducing computational complexity and wiring requirements compared to full multi-channel models.

- Explores signal images, spectrograms, and scalograms as complementary representations of EEG signals, and compares their effectiveness in capturing features relevant to abnormal EEG detection across the time, frequency, and time–frequency domains.
- Applies a DenseNet-based transfer learning strategy to enhance model performance with limited EEG data and minimize the need for training deep networks.
- Implements a post-processing approach across multiple image representations to further improve classification performance.
- Incorporates explainable AI techniques (LIME and Grad-CAM) to visualize the most influential regions of input data, thereby improving the interpretability and transparency of the model's decision-making process.

While several studies have explored deep learning for EEG analysis, relatively few have combined DenseNet-based transfer learning with signal images, spectrograms, and scalograms for abnormality detection using only two channels. By integrating lightweight representations, an efficient model architecture, and explainable AI techniques, this work contributes a novel and practical approach to interpretable, low-complexity EEG classification, well suited for real-world clinical applications where computational and interpretability constraints are important.

2. Related Work

Previous methods for TUAB EEG normal abnormal classification have predominantly focused on feature extraction techniques [20] and deep learning models [23,29]. Some studies [30,31] trained and tested models on the same patient data, without evaluating performance on unseen datasets. Additionally, many classification approaches [20,25] used only the first 60 s of EEG recordings or excluded this period due to artifacts [31], potentially limiting their generalizability to full EEG recordings or other datasets. Furthermore, many deep learning-based methods lack interpretability, which can lead to distrust from researchers.

Despite the widespread use of the TUAB dataset, existing studies vary considerably in preprocessing pipelines, channel selection, model design, and the amount of EEG data utilized. Table 1 provides a structured comparison of representative studies, highlighting these methodological differences.

Table 1. Comparison of existing studies using the TUAB dataset for EEG classification.

	Model Type	Channels	EEG Duration	Accuracy
[31]	Deep neural networks	21	<15 min	86.16%
[2]	ChronoNet (RNN)	All	First 1 min	86.57
[30]	CLBP + SVM	24	First 1 min	93.84–98.19%
[20]	KNN/RF	T5–O1	First 1 min	58.2% (KNN), 68.3% (RF)
[23]	CNN	T5–O1	First 1 min	79.34%
[29]	WaveNet + LSTM	T5–O1	First 30 s	88.76%
[21]	CNN	All	Up to 11 min	76.90%
[32]	Deep Learning	All	Up to 20 min	RMSE: 0.47–1.75

As shown in Table 1, previous studies demonstrate substantial variability in their use of TUAB EEG data, particularly in terms of selected channels and recording duration. To address these inconsistencies, our study utilizes the full-length EEG recordings and evaluates performance on an independent test set to ensure robust and generalizable results.

3. Methodology

3.1. Dataset

The TUAB dataset is divided into two folders: training sessions and test sessions, containing a total of 2993 EEG recordings collected at sampling frequencies of 250 Hz, 256 Hz, and 512 Hz. For this study, we excluded all recordings sampled at 256 Hz (189 files) and 512 Hz (18 files), retaining only the 2786 recordings with a sampling frequency of 250 Hz. Integrating data with varying sampling rates would have required downsampling the 256 Hz and 512 Hz recordings to 250 Hz, a process that may introduce inconsistencies, particularly in time-frequency representations such as spectrograms and scalograms, which are highly sensitive to sampling rate differences. Such discrepancies can degrade feature quality and negatively impact model training and performance. To ensure consistency across the dataset and facilitate reliable feature extraction, only recordings originally sampled at 250 Hz were included in the analysis.

Among these, 2518 recordings are from the training session folder and 268 from the test session folder. The training data were further split into a training set (80%, n = 2015) and a validation set (20%, n = 503). All 268 recordings in the test session folder were used as an independent test set to evaluate the performance of the proposed method. There is no overlap between the patients in the training and test sets. To avoid any potential data leakage, the test set was not used at any stage of model development or hyperparameter tuning. Furthermore, all preprocessing steps were performed using parameters computed solely from the training data, not from the full dataset.

3.2. Data Preprocessing

To reduce artifacts and enhance signal quality, a Butterworth filter (an infinite impulse response filter) was applied to obtain the frequency band of interest (0.1–100 Hz; delta, theta, alpha, beta, and gamma wave) in the TUAB EEG recordings. Additionally, a notch filter was used to remove powerline interference at 60 Hz. For classification of normal and abnormal EEGs in the TUAB dataset, channels T5 and O1 were selected as input, providing a minimal yet informative representation while reducing data dimensionality.

The EEG recordings used in this study are labelled at the file level as either normal or abnormal, without annotations indicating the precise timing or spatial location of abnormal events within each recording. This presents a fundamental challenge for training a classification model, as a label of "abnormal" does not imply that the entire EEG trace is abnormal; many abnormal-labelled EEGs contain long stretches of normal brain activity. Training on full-length recordings without accounting for this could introduce significant label noise and degrade model performance.

To address this issue and better align the input data with the provided labels, we segmented each EEG recording into 5 min non-overlapping epochs. This epoch length reflects a strategic compromise: it is sufficiently long to capture a representative temporal context and allow for the manifestation of clinically relevant EEG patterns (e.g., seizure discharges, slowing, or asymmetries), yet short enough to minimize the inclusion of extensive normal segments in recordings labelled as abnormal. In practice, 5 min windows improve the likelihood that at least part of each segment in an "abnormal" file contains diagnostically useful activity, thereby reducing the risk of mislabeling during model training.

Moreover, shorter segments would increase the number of training samples but might miss important temporal dependencies, while much longer segments could introduce redundant or irrelevant information, increase computational demands, and hinder real-time applicability. Each 5 min epoch was then transformed into three complementary representations: signal images, spectrograms, and scalograms—to capture time-domain, frequency-domain, and time–frequency characteristics, respectively, for robust classification.

3.2.1. Signal Images

EEG signal images are extensively used in event classification tasks and have become an essential tool in fields such as clinical diagnostics, neuroscience, and brain–computer interface development [33]. In this study, 5 min EEG signal images served as input to the transfer learning model (See Figure 1 'Signal image'). These signal images display the EEG signals from channels T5 and O1, with the X-axis representing time (0 to 5 min) and the Y-axis showing the amplitude (-0.0003 μV to 0.0003 μV), which corresponds to the voltage fluctuations of the EEG signal.

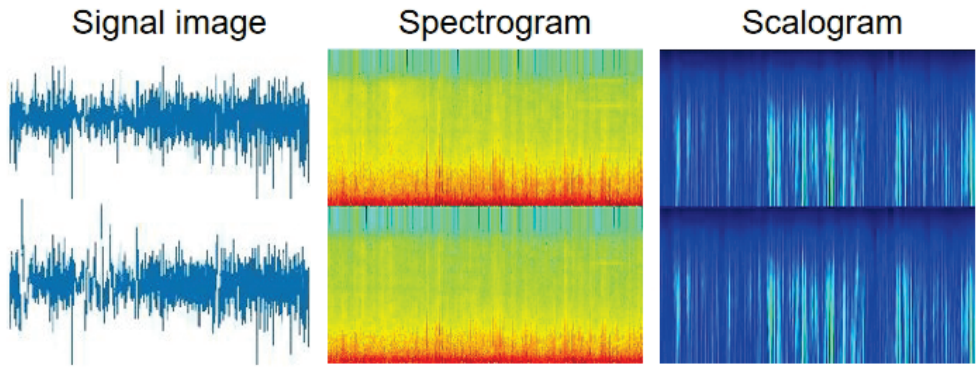


Figure 1. Examples of the signal image, spectrogram, and scalogram used as inputs in this study are presented here. The signal images display the EEG signals from channels T5 (top channel) and O1 (bottom channel), with the X-axis representing time (0 to 5 min) and the Y-axis indicating amplitude (-0.0003 μV to 0.0003 μV). In the spectrograms, the X-axis represents time (0 to 5 min), while the Y-axis depicts frequencies ranging from 0.1 to 100 Hz. For the scalograms, the X-axis corresponds to time (0 to 5 min), and the Y-axis represents wavelet scales that align with different frequency bands (0.1 to 100 Hz).

3.2.2. Spectrograms

Spectrograms are time-frequency representations that provide insights into frequency variations over time, making them particularly useful for detecting EEG changes during events such as epileptic seizures [12]. Figure 1 'Spectrogram' shows an example of the spectrogram used in this study. The X-axis in spectrograms represents time (0 to 5 min), while the Y-axis represents frequencies between 0.1 and 100 Hz. The color intensity indicates the power or magnitude of each frequency component at a given time, with hotter colors (red, yellow) showing higher power and cooler colors (blue, green) indicating lower power.

The spectrogram is generated by computing the Fast Fourier Transform (FFT) of short overlapping segments of the signal. The mathematical formula for the STFT is as follows:

1. **Short-Time Fourier Transform (STFT)**:

$$X(t,f) = \int_{-\infty}^{\infty} x(\tau)\, w(t-\tau) e^{-j2\pi f\tau}\, d\tau \tag{1}$$

where:

- $x(t)$ is the signal in the time domain.
- $w(t-\tau)$ is the window function (in this case, a Hamming window).
- f is the frequency.
- t is the time index for the center of the window.

2. **Hamming Window Function**:

$$w(t) = 0.54 - 0.46\cos\left(\frac{2\pi t}{N-1}\right) \tag{2}$$

where:

- N is the length of the window.
- t is the time index within the window.

3. **Spectrogram**:

The spectrogram is the squared magnitude of the STFT:

$$\text{Spectrogram}(t, f) = |X(t, f)|^2 \tag{3}$$

3.2.3. Scalograms

Scalograms, derived from the continuous wavelet transform (CWT), have gained attention for their ability to capture temporal variations in EEG signals, making them highly effective for distinguishing time-frequency features of various brain events [34]. In scalograms, the X-axis represents time (0 to 5 min), while the Y-axis reflects wavelet scales that correspond to different frequency bands (0.1 and 100 Hz). The color intensity, which represents the magnitude of wavelet coefficients, highlights the strength of specific frequency components at particular times, offering a detailed view of EEG signal variations.

The scalogram using the continuous wavelet transform (CWT), which is applied to each EEG epoch (after preprocessing and filtering).

1. **Continuous Wavelet Transform (CWT)**:

$$W_x(a, b) = \int_{-\infty}^{\infty} x(t)\, \psi^* \left(\frac{t - b}{a} \right) dt \tag{4}$$

where:

- $x(t)$ is the EEG signal.
- $\psi(t)$ is the mother wavelet.
- a is the scale parameter.
- b is the translation parameter.

2. **Morlet Wavelet** (Complex Morlet Wavelet):

$$\psi(t) = \pi^{-\frac{1}{4}}\, e^{i2\pi f_0 t}\, e^{-\frac{t^2}{2\sigma^2}} \tag{5}$$

where:

- f_0 is the central frequency of the wavelet.
- σ controls the width of the wavelet in time.

3. **Scale and Frequency Relation**: The scale a is related to the central frequency f_0 and σ as:

$$a = \frac{1}{2\pi f_0 \sigma} \tag{6}$$

The corresponding frequency for each scale is given by:

$$f(a) = \frac{f_0}{a} \tag{7}$$

4. **Scalogram**:

The scalogram is the absolute value of the CWT coefficients:

$$\text{Scalogram}(t, f) = |W_x(a, b)| \tag{8}$$

Figure 1 illustrates examples of the signal image, spectrogram, and scalogram used in this study.

3.3. Model Development

DenseNet121 is a convolutional neural network architecture that utilizes dense connections between layers, where each layer is connected to all subsequent layers in a feed-forward manner [35]. This structure helps mitigate the vanishing gradient problem, resulting in several benefits: it reduces the training complexity of deep learning models, enables the reuse of features, and decreases the number of parameters compared to other architectures [36]. DenseNet121 is widely used for image classification tasks, particularly in medical image analysis, as it can achieve high accuracy with fewer parameters, making it suitable for solving complex medical classification challenges [37].

In this study, signal images, spectrogram and scalogram serve as inputs separately. We utilize the DenseNet121 architecture pre-trained on ImageNet as a feature extractor for EEG-based classification tasks. We exclude the top fully connected layers of the DenseNet121 model, allowing fine-tuning of the model for normal and abnormal classification. Table 2 shows the input and outputs of our study.

Table 2. Input and output features used in the model.

Feature Type	Description	Data Format	Purpose
Signal Images	Raw EEG signal converted to time-domain images	2D images	Capture temporal EEG patterns
Spectrograms	Time-frequency representation of EEG signals	2D images	Capture frequency changes over time
Scalograms	Time-frequency representation using wavelets	2D images	Capture localized time-frequency features
Output	EEG classification label	Binary label (Normal/Abnormal)	Identify EEG abnormality status

To improve the model's adaptability to the task, we unfreeze the last 10 layers of DenseNet121, making them trainable while keeping the earlier layers frozen to retain pre-learned ImageNet features, followed by three fully connected layers with 128, 64, and 32 units, respectively, each using ReLU activation and L2 regularization (0.005) to prevent overfitting. The final output layer consists of a single neuron with a sigmoid activation function for normal and abnormal classification.

The model is compiled with the Adam optimizer with a learning rate of 0.01, and the loss function is binary cross-entropy, which is appropriate for the binary classification task. The performance is evaluated using accuracy, recall, and precision. To prevent overfitting and ensure optimal model performance, early stopping (patience = 20) is implemented, monitoring validation loss, while a model checkpoint is used to save the best-performing model based on validation accuracy. The model is trained for up to 100 epochs using the training and validation datasets. Table 3 shows the hyperparameters used for model training and validation.

Table 3. Hyperparameters used for model training and validation.

Hyperparameter	Value	Justification
Base Model	DenseNet121 (pre-trained on ImageNet)	Provides strong feature extraction capability from image-based EEG representations
Trainable Layers	Last 10 layers unfrozen	Allows task-specific fine-tuning while preserving general visual features from ImageNet
Top Layers	Three dense layers (128, 64, 32 units)	Adds task-specific nonlinearity and representation capacity
Activation Function	ReLU (for hidden layers), Sigmoid (for output)	ReLU prevents vanishing gradient; Sigmoid suitable for binary classification
Regularization	L2 = 0.005	Helps prevent overfitting by penalizing large weights
Loss Function	Binary Cross-Entropy	Appropriate for binary classification (normal vs. abnormal)
Optimizer	Adam	Adaptive learning rate, widely used and effective for deep networks
Learning Rate	0.01	Empirically chosen to balance convergence speed and stability
Early Stopping	Patience = 20 (monitoring validation loss)	Stops training when performance no longer improves to avoid overfitting
Model Checkpoint	Best model saved by validation accuracy	Ensures optimal model selection during training

3.4. Post-Processing

We combine the predictions from multiple image outputs by averaging them to obtain the final prediction. A full EEG recording typically lasts around 20 min (see Table 4), and we used a 5 min segment. Therefore, each EEG file produces approximately three to five images. The final prediction for each EEG file is made by calculating the average of the prediction probabilities across all images generated from that file. Figure 2 shows the flowchart of the proposed method.

Table 4. Number and duration of the TUAB EEGs used in the training, validation, and test set.

	Normal	Abnormal	Total
Train	1045	970	2015
Duration/file (mins)	20.50	20.37	20.44
Duration (mins)	21,425	19,760	41,185
Validation	261	242	503
Duration/file (mins)	20.06	20.91	20.47
Duration (mins)	5235	5060	10,295
Test	148	120	268
Duration/file (mins)	20.17	20.79	20.45
Duration (mins)	2985	2495	5480

Let n_i represent the number of images generated from EEG file i, where $n_i \in [3, 5]$.

For each image j from EEG file i, let p_{ij} denote the prediction probability.

The final prediction P_i for EEG file i is calculated as the average of the prediction probabilities across all images:

$$P_i = \frac{1}{n_i} \sum_{j=1}^{n_i} p_{ij} \tag{9}$$

The performance of the model is evaluated based on these final predictions.

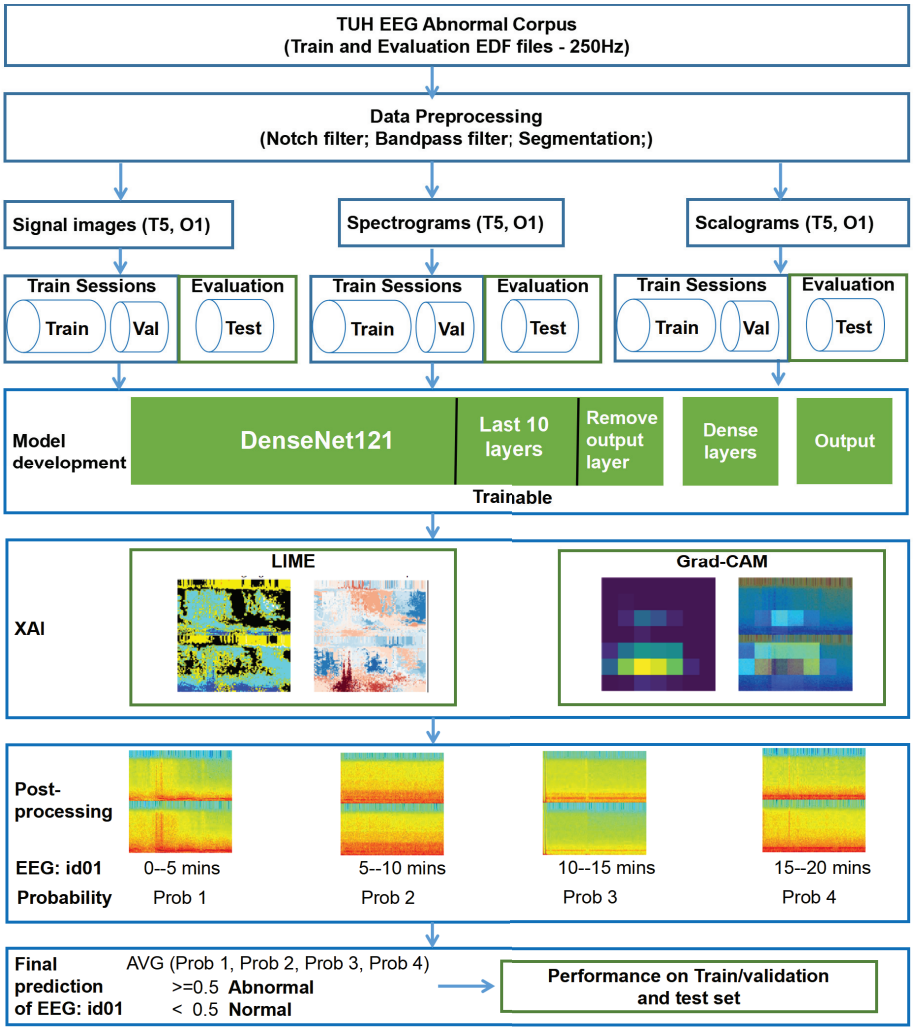


Figure 2. Overview of the TUAB EEG classification method: the TUAB EEG classification method involves the selection of EEG normal and abnormal files with a sampling frequency of 250 Hz for this study. The EEG signals are processed using a 60 Hz notch filter to remove powerline interference, followed by a bandpass filter (0.1–100 Hz) to obtain the frequency band of interest. Subsequently, the signals are segmented into 5 min intervals from channels T5 and O1. Each segment is transformed into signal images, spectrograms, and scalograms, which are then individually input into the DenseNet121 model. To analyze the model's output, we employed LIME and Grad-CAM techniques. The final classification for each EEG file as normal or abnormal is determined by averaging the prediction probabilities from all images generated for that file.

3.5. Explainable Methods

In this section, we present the visualisation and interpretability methods used in our study. Specifically, LIME and Grad-CAM, which are introduced in Sections 3.5.1 and 3.5.2, respectively, are employed to interpret the model's predictions.

3.5.1. Local Interpretable Model-Agnostic Explanations

In this study, the LIME technique was used to interpret the model's predictions [16]. LIME was specifically applied to generate explanations for the classification results by creating perturbed versions of the input image and assessing their impact on the model's predictions. The process involves preprocessing the input image, performing model infer-

ence, explaining the model's prediction through LIME, followed by the visualization of the explanation using heatmaps.

a. Image Preprocessing

Given an input EEG spectrogram I of shape (H, W, C), where H represents the height, W the width, and C the number of channels, the image is resized to match the model's input dimensions of $224 \times 224 \times 3$.

b. Model Prediction

The pre-trained DenseNet121 model, denoted as f_{DenseNet}, is used to perform the inference. Given the input image $\mathbf{I}_{\text{batch}}$, the model outputs a probability vector $\mathbf{p} = [p_1, p_2, \ldots, p_K]$, where K is the number of classes. In our binary classification task, $K = 2$, corresponding to class 0 (normal) and class 1 (abnormal).

The predicted class $\hat{y}$ is defined as the class with the highest predicted probability:

$$\hat{y} = \arg \max_{i \in \{1,\ldots,K\}} p_i \tag{10}$$

The corresponding predicted probability, or confidence score, is defined as:

$$a = p_{\hat{y}} = \max(p_i), \quad i = 1, \ldots, K \tag{11}$$

This confidence score a indicates how certain the model is in its prediction and is used later as a reference value when interpreting the impact of individual superpixels in the LIME explanation. Specifically, higher confidence allows more robust interpretation of which regions most strongly contributed to the final decision.

c. Model Explanation using LIME

To interpret the model's prediction, we apply Local Interpretable Model-agnostic Explanations (LIME), a technique that approximates the black-box model with a surrogate interpretable model. Let x represent the input image, and $\mathcal{E}(x)$ represent the explanation for x, which can be formally written as:

$$\mathcal{E}(x) = \arg \min_{\mathcal{M}} \mathcal{L}(f, \mathcal{M}, x) \tag{12}$$

where:

- $\mathcal{M}$ is a simple interpretable model (e.g., linear regression or decision tree),
- $\mathcal{L}$ is the loss function that measures the fidelity of $\mathcal{M}$ to the original model f,
- f is the black-box model (DenseNet121),
- x is the input image.

LIME generates explanations by perturbing the input image and training a local surrogate model on these perturbed samples. The explanation is provided in the form of a set of superpixels, which are small regions of the image that are interpreted to be meaningful features.

d. Superpixel Mask and Heatmap Generation

LIME assigns a weight w_i to each superpixel x_i to indicate its contribution to the model's decision. The local explanation for a given class c can be represented as:

$$\text{Explanation}(x) = \sum_{i \in M} w_i x_i \tag{13}$$

where M denotes the set of superpixels, and x_i is the feature corresponding to superpixel i. The explanation weight w_i quantifies the importance of each superpixel in the model's

decision. Superpixels are obtained through image segmentation techniques, SLIC (Simple Linear Iterative Clustering), which partitions the image into small, homogeneous regions based on pixel similarity. The resulting superpixels are treated as features, and their contributions to the model's prediction are assessed.

Using these weights, a heatmap is generated to visually represent the importance of each region in the image. The heatmap value at each superpixel i is given by:

$$\text{Heatmap}(i) = w_i \tag{14}$$

e. Visualization of Explanation and Heatmap

The final explanation is visualized using multiple plots:

- The original image and the preprocessed image are displayed alongside the LIME explanation (positive superpixels only), which highlights the regions of the image that contributed most to the prediction.
- The heatmap is generated using the weights w_i corresponding to each superpixel, with a color map applied to visualize the contributions.

The heatmap values are normalized to the range $[-\max(\text{heatmap}), \max(\text{heatmap})]$, and visualized using a color map, typically a diverging colormap like 'RdBu':

$$\text{heatmap} = \text{normalize}(w_i) \quad \text{for each superpixel } i \tag{15}$$

The heatmap is then overlaid on the image to provide a clear visualization of the model's decision-making process. From this analysis, we extracted a mask that highlights significant features and visualized it alongside the original image, its preprocessed version, and a color-mapped heatmap indicating the contributions of different segments to the prediction. In this heatmap, cooler colors such as blue represent regions that negatively affect the model's decision, whereas warmer colors like red indicate areas that positively affect it. The visualization effectively demonstrated the model's decision-making process, facilitating a better understanding of the influential factors behind its classification results.

3.5.2. Gradient-Weighted Class Activation Mapping

Grad-CAM was employed to visualize the important regions in images classified by the DenseNet121 model. It offers visual explanations for predictions made by convolutional neural networks [17]. In this study, input images were resized to $224 \times 224 \times 3$ and passed through the pre-trained DenseNet121 architecture on ImageNet. We identified the last convolutional layer and used it to compute the gradients of the predicted class with respect to its feature maps. This gradient information was utilized to generate a heatmap highlighting areas contributing significantly to the model's predictions. The heatmap was colorized using a jet colormap and overlaid onto the original image to improve visual interpretation. Yellow regions indicate areas with the highest impact, while cooler colors, such as blue, represent lower importance. This approach provides a clearer understanding of the model's decision-making process and enhances its interpretability.

a. Image Preprocessing

Let I represent the input image with dimensions $H \times W \times C$, where H is the height, W is the width, and C is the number of color channels (usually three for RGB images). The image is resized to the target size $S \times S \times C$, where S is typically 224, the input size required by the model.

b. Grad-CAM Heatmap Generation

Given the preprocessed image $\mathbf{I}_{\text{batch}}$, Grad-CAM is applied to generate the heatmap that highlights the regions of the image most influential for the model's prediction. Grad-CAM works by computing the gradients of the predicted class with respect to the activations of the last convolutional layer.

Let $f(\mathbf{I})$ denote the model's prediction for the input image $\mathbf{I}$, and $\hat{y}$ be the predicted class. The output of the model is the predicted probability vector:

$$\mathbf{p} = f(\mathbf{I}_{\text{batch}}) \tag{16}$$

Let A represent the output feature map of the last convolutional layer, and $\mathbf{A} \in \mathbb{R}^{H \times W \times D}$, where D is the number of channels in the feature map. We compute the gradients of the class score C_k with respect to the feature map A, and we express this as:

$$\frac{\partial C_k}{\partial A} = \text{gradients of class score w.r.t feature map} \tag{17}$$

Using a gradient tape, we record the gradients with respect to the feature maps, and compute the mean of the gradients over all spatial locations to obtain a vector $\mathbf{g_k}$:

$$\mathbf{g_k} = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} \frac{\partial C_k}{\partial A_{i,j}} \tag{18}$$

We then compute the class activation heatmap by performing a weighted sum of the feature map channels using the gradients $\mathbf{g_k}$ as weights:

$$\mathbf{H}_{\text{gradcam}} = \sum_{k=1}^{D} g_k \cdot A_k \tag{19}$$

The resulting heatmap $\mathbf{H}_{\text{gradcam}}$ is of size $H \times W$, representing the class activation map that highlights the important regions for the model's prediction.

Finally, to normalize the heatmap between 0 and 1, we use the following transformation:

$$\mathbf{H}_{\text{gradcam}}{}^{\text{normalized}} = \frac{\mathbf{H}_{\text{gradcam}} - \min(\mathbf{H}_{\text{gradcam}})}{\max(\mathbf{H}_{\text{gradcam}}) - \min(\mathbf{H}_{\text{gradcam}})} \tag{20}$$

c. Superimposing the Heatmap onto the Image

To visualize the result, we superimpose the generated heatmap $\mathbf{H}_{\text{gradcam}}{}^{\text{normalized}}$ onto the original image I_{original}. The heatmap is colorized using a colormap, typically the jet colormap:

$$\text{colormap}(H_{\text{gradcam}}) = \text{jet}(H_{\text{gradcam}}) \tag{21}$$

The colorized heatmap is then resized to match the original image dimensions and combined with the original image I_{original} with a blending factor α:

$$I_{\text{superimposed}} = \alpha \cdot \text{colormap}(H_{\text{gradcam}}) + (1 - \alpha) \cdot I_{\text{original}} \tag{22}$$

This produces the final image with the heatmap overlaid, which can be visualized. Grad-CAM provides an interpretable visualization of the model's decision-making by highlighting the regions that most important to the model's output. The class activation map is obtained by backpropagating gradients from the predicted class through the last convolutional layer, followed by a weighted sum over the feature maps. This method improves our understanding of the parts of the input image that contribute most to the model's decision.

3.6. Performance Evaluation

The sensitivity, specificity, precision, accuracy, F1 score and balanced accuracy were used in estimating the performance of the DenseNet121-based EEG normal abnormal classification method.

$$Sensitivity = \frac{TP}{TP + FN} \tag{23}$$

$$Specificity = \frac{TN}{TN + FP} \tag{24}$$

$$Precision = \frac{TP}{TP + FP} \tag{25}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{26}$$

$$F1 = 2 \times \frac{Sensitivity \times Precision}{Sensitivity + Precision} \tag{27}$$

$$BalancedAccuracy = \frac{Sensitivity + Specificity}{2} \tag{28}$$

where:

- True positives (*TP*): the number of abnormal EEGs predicted as abnormal EEGs;
- False positives (*FP*): the number of normal EEGs predicted as abnormal EEGs;
- True negatives (*TN*): the number of normal EEGs predicted as normal EEGs;
- False negatives (*FN*): the number of abnormal EEGs predicted as normal EEGs.

4. Results

4.1. Performance Evaluation of DenseNet121 for EEG Classification

In this study, to mitigate overfitting and ensure optimal results, early stopping is applied based on validation loss. Additionally, a model checkpoint is employed to save the best-performing model according to validation accuracy. Training is conducted for up to 100 epochs using both the training and validation datasets. Figures 3–5 illustrate the training and validation loss and accuracy curves for the signal image, spectrogram, and scalogram inputs.

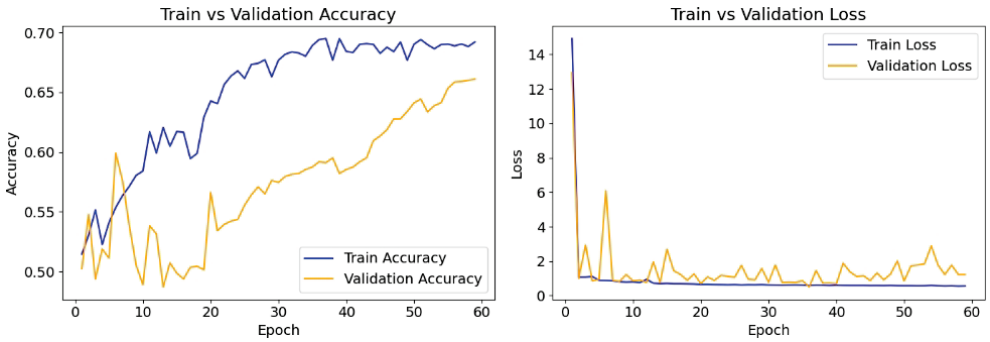


Figure 3. Training and validation loss and accuracy curves for signal image inputs.

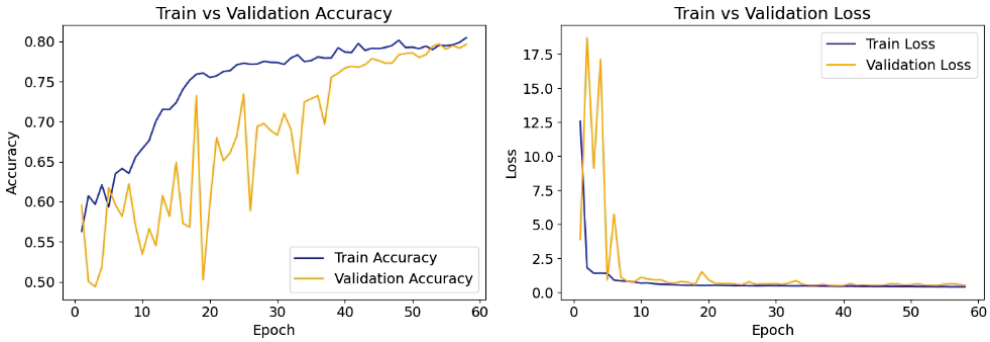


Figure 4. Training and validation loss and accuracy curves for spectrogram inputs.

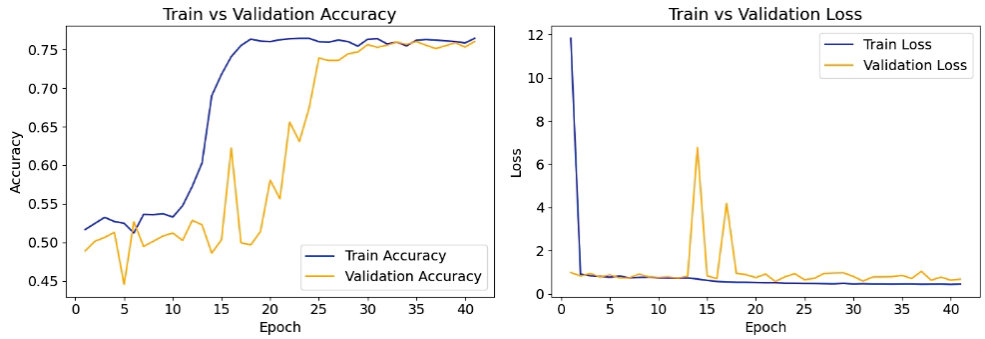


Figure 5. Training and validation loss and accuracy curves for scalogram inputs.

Table 5 shows the performance of the DenseNet121-based classification method for identifying normal and abnormal EEG signals using three different input types: signal images, spectrograms, and scalograms. Among these, spectrograms achieved the highest accuracy on the test set at 76.55%, compared to 69.53% for signal images and 74.56% for scalograms.

Table 5. Performance evaluation of the DenseNet121-based classification method for identifying normal and abnormal EEG signals was conducted using signal images, spectrograms, and scalograms as separate inputs. This evaluation was performed on the training, validation and test set.

		Sensitivity	Specificity	Precision	Accuracy	F1 Score	Balanced Accuracy
Signal Image	Train	54.78%	82.52%	74.30%	69.21%	63.06%	68.65%
	Validation	52.57%	79.18%	70.93%	66.10%	60.39%	65.87%
	Test	55.11%	81.58%	71.43%	69.53%	62.22%	68.35%
Spectrogram	Train	76.06%	84.55%	81.95%	80.48%	78.90%	80.31%
	Validation	74.31%	84.81%	82.55%	79.65%	78.21%	79.56%
	Test	70.94%	81.24%	75.97%	76.55%	73.37%	76.09%
Scalogram	Train	73.58%	79.14%	76.49%	76.47%	75.01%	76.36%
	Validation	71.44%	80.52%	77.99%	76.06%	74.57%	75.98%
	Test	70.34%	78.05%	72.82%	74.54%	71.56%	74.20%

4.2. Performance of Post-Processing Method

Table 6 shows the performance of the DenseNet121-based classification method (using signal images, spectrograms and scalograms as input) after post-processing, evaluated on the training, validation, and test sets. The results show a noticeable improvement in sensitivity, specificity, precision, accuracy and F1 score, following the application of the post-processing method (Section 3.4) on signal images, spectrograms and scalograms. Figure 6 shows the confusion matrix for the test set, where spectrograms are used as the input.

Table 6. Performance of the DenseNet121-based classification method after post-processing, evaluated on the training, validation, and test sets, using signal images, spectrograms, and scalograms as inputs.

		Sensitivity	Specificity	Precision	Accuracy	F1 Score
Signal Image	Train	56.65%	83.72%	76.39%	70.67%	65.05%
	Validation	53.62%	79.11%	71.07%	66.64%	61.12%
	Test	55.83%	80.41%	69.79%	69.40%	62.04%
Spectrogram	Train	77.95%	87.79%	85.58%	83.05%	81.59%
	Validation	75.17%	85.38%	83.11%	80.39%	78.94%
	Test	75.00%	83.78%	78.95%	79.85%	76.92%
Scalogram	Train	83.07%	89.17%	87.71%	86.23%	85.33%
	Validation	74.90%	81.20%	79.22%	78.12%	77.00%
	Test	73.33%	82.43%	77.19%	78.36%	75.21%

Sens: Sensitivity; Spec: Specificity; Prec: Precision; Acc: Accuracy

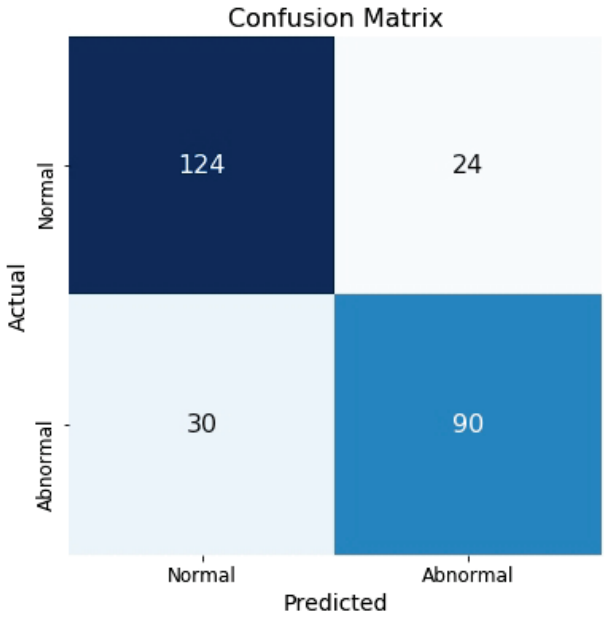


Figure 6. Confusion matrix for the test set.

4.3. Comparison with Previous Work

Table 7 compares the performance of the DenseNet121-based EEG normal–abnormal classification method with previous studies. Our method demonstrated higher sensitivity (75.0%) and accuracy (79.9%) compared to previous work on TUH abnormal EEG classification.

Table 7. Performance comparison of DenseNet121-based EEG normal–abnormal classification methods with previous works.

	Method	Sens (%)	Spec (%)	Acc (%)	BAC (%)
[20]	KNN	-	-	58.2	-
[20]	Random forest	-	-	68.3	-
[21]	1D-CNN	-	-	76.9	-
[21]	2D-CNN	-	-	70.39	-
[21]	TCNN-RNN	-	-	71.48	-
[22]	LSTM	-	-	79.18	-
[23]	Deep CNN	71.4	86.0	79.3	-
[24]	CNN	-	-	-	71.9
[25]	BD-Shallow	72–87	69–96	-	-
Ours	DenseNet	75.0	83.8	79.9	79.4

Sens: Sensitivity; Spec: Specificity; Acc: Accuracy; BAC: Balanced Accuracy; TCNN-RNN: Time-distributed convolutional recurrent neural network; BD-Shallow: Braindecode Shallow ConvNet;

4.4. Explainable Methods

Figures 7–10 presents examples of true positive, true negative, false positive, and false negative events. LIME highlights and heatmaps illustrate the explanation weights assigned to the DenseNet121 method. Grad-CAM highlights the image regions influencing the model's decision, with warmer colors indicating areas of higher contribution and cooler colors representing regions of lesser importance.

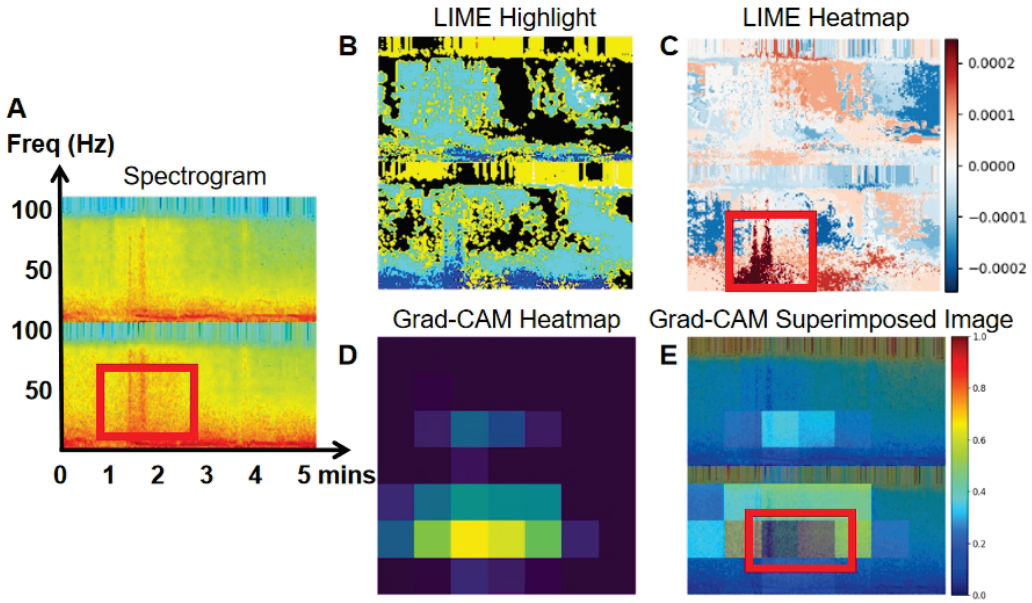


Figure 7. Example of a true positive event: (**A**) spectrogram with (**B**,**C**) LIME and (**D**,**E**) Grad-CAM visualizations (T5: top channel; O1: bottom channel), with a prediction probability of 0.705.

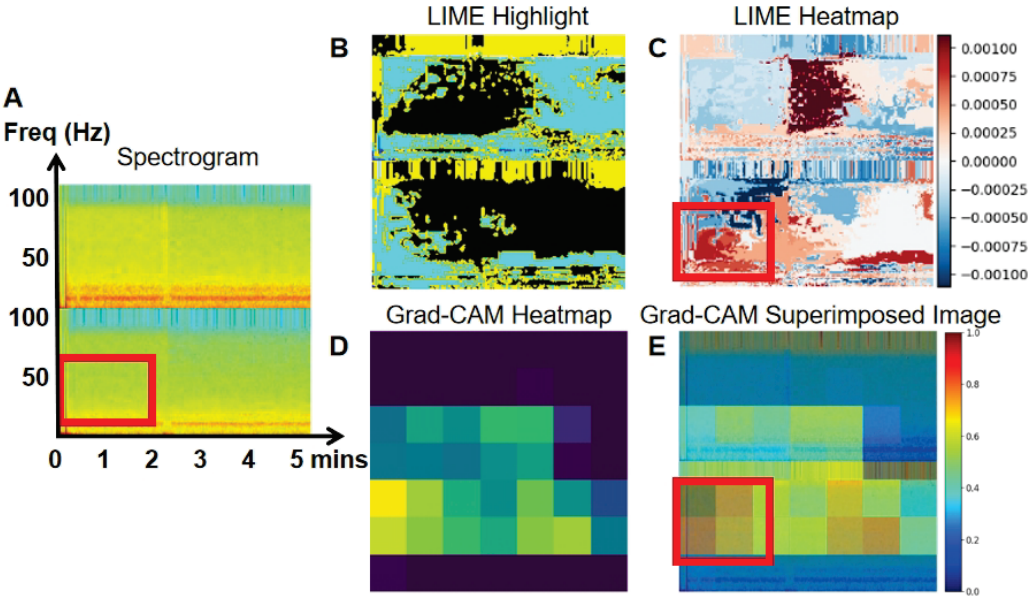


Figure 8. Example of a true negative event: (**A**) spectrogram with (**B**,**C**) LIME and (**D**,**E**) Grad-CAM visualizations (T5: top channel; O1: bottom channel), with a prediction probability of 0.121. The red box highlights the normal EEG recordings.

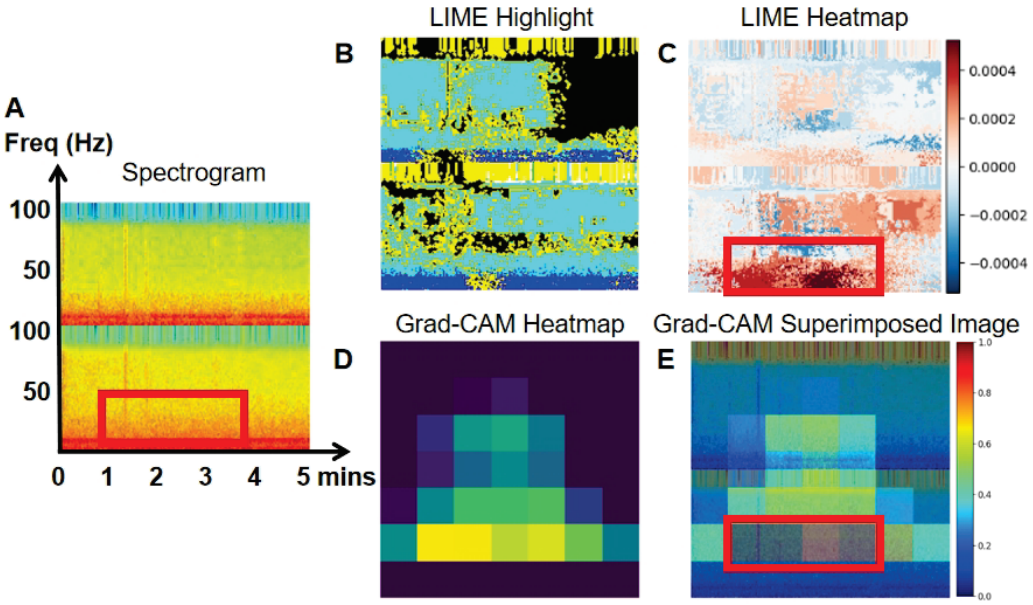


Figure 9. Example of a false positive event: (**A**) spectrogram with (**B**, **C**) LIME and (**D**, **E**) Grad-CAM visualizations (T5: top channel; O1: bottom channel), with a prediction probability of 0.757. The red box highlights the abnormal EEG recordings.

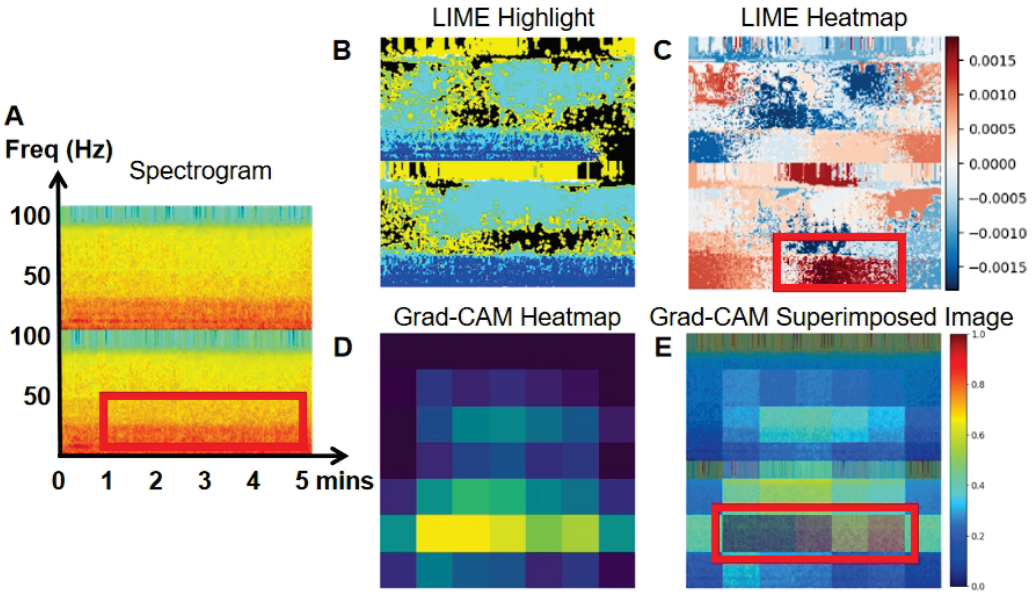


Figure 10. Example of a false negative event: (**A**) spectrogram with (**B**,**C**) LIME and (**D**,**E**) Grad-CAM visualizations (T5: top channel; O1: bottom channel), with a prediction probability of 0.166. The red box highlights the abnormal EEG recordings.

5. Discussion

In this study, we introduce a DenseNet-based method to classify normal and abnormal EEGs using the TUAB datasets. Unlike previous studies, which often relied on multiple EEG channels [25,31,32,38], our approach focuses on channels T5 and O1. This helps mitigate the challenges associated with using multiple EEG channels. First, it simplifies the direct identification of subject-dependent reactive bands, avoiding the need for automated identification processes [39]. Second, it reduces the dimensionality of the feature vector, which can otherwise negatively affect classifier performance.

Several previous studies have used the same TUAB dataset as this study to classify EEG recordings as normal or abnormal. However, their approaches vary widely in terms of which portions of the data they use. Gemein et al. [31] used 21 channels and excluded

the first 60 s of each recording to reduce artifacts, analyzing up to 20 min of EEG and achieving 86.16% accuracy with 5-fold cross-validation. In contrast, Roy et al. [2] trained a deep recurrent network (ChronoNet) using only the first minute of EEG, reporting 86.57% accuracy. Similarly, Tuncer et al. [30] focused on the first minute across 24 channels, applying chaotic local binary pattern analysis and obtaining accuracies between 93.84% and 98.19% using SVMs.

Other studies also concentrated on brief EEG segments on channel T5–O1 of TUAB EEGs: Lopez et al. [20] used the first 60 s from the T5–O1 channel, with error rates of 41.8% (KNN) and 31.7% (random forest); Yildirim et al. [23] applied a CNN to the same channel (T5–O1) and duration, reaching 79.34% accuracy. Albaqami et al. [29] used only the first 30 s with a WaveNet-LSTM model, reporting 88.76% accuracy. Roy et al. [21] included up to 11 min of EEG in a CNN model, achieving 76.90% test accuracy, while Kiessner et al. [32] removed the first minute and used up to 20 min, reporting RMSE values between 0.47 and 1.75.

These studies have adopted inconsistent strategies regarding the use of the first minute of EEG recordings in the TUAB dataset. Some excluded it due to concerns about artifacts, while others used only the first minute, suggesting it is representative of the entire recording. However, these conflicting conclusions create ambiguity in the classification process. Abnormal events may or may not occur within the initial minutes, so relying exclusively on or removing the first minute risks producing inconsistent and potentially biased results. Moreover, the rationale for these choices is often unclear or insufficiently justified. Such inconsistencies undermine the comparability and robustness of model performance across studies. To address this issue and ensure a more comprehensive and consistent representation of the data, we chose to use the entire EEG recording when developing our model. In addition, these prior studies were not evaluated on an independent test set, leaving their generalizability to unseen data uncertain. Nonetheless, as most prior studies employed cross-validation, we adopted the same approach to ensure fair and consistent performance comparisons. The detailed results of our 5-fold cross-validation are presented in Table A1.

In this work, we develop a DenseNet-based classification method using signal images, spectrograms, and scalograms as inputs, which are widely employed in EEG analysis [12,33,40,41]. For the signal images, we constrained the signal amplitude to a fixed range (-0.0003 μV to 0.0003 μV) to suppress extreme outliers and stabilize the visual representation. For spectrograms, we limited the frequency range to 0.1–100 Hz to focus on clinically relevant EEG activity and exclude high-frequency noise. Similarly, for scalograms, we selected wavelet scales corresponding to the same 0.1–100 Hz frequency band, which helps reduce the impact of irrelevant or noisy components outside this range.

Moreover, we used the entire TUAB EEG recordings in the development of our method. Each EEG recording used in this study lasts approximately 20 min (see Table 4). We split the EEG into 5 min segments, generating around three to five images per file. Not all 5 min segments in the EEGs labelled as abnormal will contain abnormal events. Similarly, some segments from normal EEGs may contain artifacts. To address this, we averaged predictions from multiple image outputs to obtain the final prediction for each file, thereby reducing the misclassification rate. Table 6 shows that this approach improves overall accuracy by averaging prediction probabilities across all images from each EEG file. Figure 6 presents the confusion matrix for the test set, where spectrograms are used as the input.

To enhance the interpretability of our model, we employed LIME and Grad-CAM techniques. Figures 7–10 present the spectrogram (A), LIME highlights (B), and the heatmap (C) generated by LIME. The spectrogram shows the frequency content of the EEG signal over time, while LIME highlights specific sections of the spectrogram that significantly affect the model's predictions, aiding in the understanding of classification decisions. The heatmap, generated by LIME, uses color gradients to indicate the importance of different spectrogram

regions in the model's decision-making. Warmer colors (red) represent a positive impact, while cooler colors (blue) indicate a negative impact.

Grad-CAM heatmaps (D) are also shown in Figures 7–10. Warmer colors, like red and yellow, highlight regions most influential in the model's decision, while cooler colors, like blue, represent less significant areas. In the superimposed images, the red and yellow regions align with EEG areas of higher power, frequency, and amplitude, providing deeper insight into the model's decision-making process in image classification.

Figures 7–10 illustrates true positive, true negative, false positive, and false negative events, along with their corresponding LIME (B and C) and Grad-CAM visualizations (D and E), highlighting the specific areas of the spectrogram that the model focused on during classification. In Figure 7, an example of a true positive event is shown. The LIME heatmap (Figure 7C) and Grad-CAM overlay (Figure 7E) clearly highlight the high-power signal with an elevated frequency on channel O1. This signal is marked in blue within the red block on the LIME heatmap, and as the yellow epoch within the red block on the Grad-CAM overlay. The original spectrogram in red indicates higher power at this point. These visualizations demonstrate that this high-power signal plays a crucial role in the model's prediction of abnormal EEG. The increased power corresponds to the classification of abnormal EEG, as higher power is considered a positive indicator of abnormalities. In contrast, Figure 8 presents a true negative event, where the lower-frequency power signal on channel O1 (within the red block in the Spectrogram, Figure 8A) is crucial for classifying the spectrogram as normal EEG. This comparison highlights that abnormal EEG often shows higher-frequency characteristics compared to normal EEG.

In Figure 9, a false positive event is shown, where elevated power in channels T5 and O1 led to the misclassification of a normal EEG as abnormal. This is highlighted in blue on the LIME heatmap (Figure 9C) and yellow on the Grad-CAM overlay (within the red block, Figure 9E). The misclassification can likely be attributed to artifacts in the spectrogram that misled the model. In contrast, Figure 10, a false negative event, depicts a case where high power at lower frequencies (indicated by the red portion in the spectrogram, Figure 10A) enabled the model to accurately classify the event as normal. However, it is important to note that not all abnormal spectrograms contain detectable abnormal events, resulting in potential misclassifications, particularly false negatives. To address these issues, visualizations produced by LIME (Figure 10C) and Grad-CAM (Figure 10E) can assist researchers in manually correcting such misclassifications. Additionally, we propose a post-processing method in this study aimed at reducing false classification events.

A limitation of the current work is the potential impact of artifacts in the EEG recordings on model performance. While the TUH EEG dataset provides a valuable resource, it may still contain various types of noise and artifacts, such as eye blinks, muscle activity, or electrical interference, which could compromise the quality of the signals. These artifacts, if not properly mitigated, could lead to misclassifications or reduce the overall performance of the model. Although our method includes a post-processing step to average predictions across multiple image outputs, the presence of artifacts in certain segments could still result in false positives or false negatives. Future work should explore more robust artifact removal or detection techniques, such as adaptive filtering or artifact subspace projection, to further enhance the model's accuracy and reliability in clinical applications.

Another limitation of the current study is that our method was developed and tested only on the TUH EEG dataset. While the results are promising, the model's performance and generalizability across other EEG datasets or real-world clinical settings remain untested. Variations in data acquisition protocols or patient populations could affect the model's effectiveness. To enhance robustness and generalizability, future work will involve validating the method on diverse EEG datasets and in real-world clinical environments.

This will help assess how well the method generalizes to different populations and recording conditions. Additionally, in future work, we will explore the impact of different EEG channel combinations on model performance, and further analyze the relative importance of each channel in abnormal EEG detection. Moreover, future work would focus on testing the proposed method using more specific datasets to determine its ability to distinguish disease symptoms or differentiate between distinct neurological diseases. For example, datasets related to epilepsy type classification, seizure detection, or seizure prediction could be employed to assess the method's effectiveness in identifying disease-specific patterns. This will be essential for validating the model's generalizability and clinical applicability. Furthermore, it is important to note that certain EEG events may be visible in the raw signal but not apparent in the spectrogram or scalogram, potentially affecting model performance [12]. To improve effectiveness and clinical applicability, we plan to integrate raw EEG signals with spectrograms or scalograms in future work, either through input fusion or decision-level ensemble methods.

In addition, to ensure the successful integration of this model into clinical practice, several challenges must be addressed. Real-time deployment in clinical settings requires effective handling of artifacts, rapid processing times, and minimizing false positives and false negatives to avoid clinical misinterpretation. The model's interpretability is also critical for clinical adoption. Techniques such as LIME and Grad-CAM provide transparency into the model's decision-making process, which is essential for fostering clinician trust. However, explainability methods may themselves produce inconsistent or overly complex outputs, which could confuse rather than assist clinicians. Therefore, co-designing interpretability outputs with domain experts is crucial. Further validation and refinement of these interpretability tools, in collaboration with clinicians, will be necessary to ensure the model can be integrated into clinical workflows and effectively support decision-making. Additionally, real-world deployment in industrial or embedded systems will require model optimization for resource-constrained hardware, along with rigorous testing to ensure reliability, safety, and compliance with healthcare regulations.

6. Conclusions

In this study, we presented a DenseNet-based method for the classification of entire EEG files as normal or abnormal, focusing on simplifying the process by using fewer channels (T5 and O1). The model demonstrated high accuracy, especially when using spectrogram inputs, and the post-processing method further improved performance by averaging predictions across multiple images. Furthermore, LIME and Grad-CAM techniques were applied to improve the interpretability of the model by visualizing which parts of the input data contributed most to the predictions. These visualizations help explain the model's decision-making process and allow for a better understanding and potential refinement of the classification method.

Author Contributions: Conceptualization, L.W.; data curation, L.W.; formal analysis, L.W.; funding acquisition, C.M.; investigation, L.W.; methodology, L.W.; project administration, C.M.; resources, L.W.; supervision, C.M.; validation, L.W.; visualization, L.W.; writing—original draft, L.W.; writing—review and editing, L.W. and C.M. All authors have read and agreed to the published version of the manuscript.

Funding: This publication has emanated from research conducted with the financial support of Taighde Éireann—Research Ireland, under Grant number 21/RC/10294_P2 at FutureNeuro Research Ireland Centre for Translational Brain Science.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available at https://isip.piconepress.com/projects/tuh_eeg/, reference number [18].

Acknowledgments: OpenAI's ChatGPT was employed to enhance grammar, clarity, and language expression. The authors retain full responsibility for the content and conclusions presented in this work.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A

Table A1 presents the performance metrics of the DenseNet121-based classification method using spectrograms, evaluated through five-fold cross-validation.

Table A1. Performance of the DenseNet121-based classification method (spectrogram) using five-fold cross-validation (mean ± standard deviation).

Metric	Value ($\pm$Std)
Sensitivity	70.18% $\pm$ 0.84%
Specificity	86.33% $\pm$ 0.59%
Precision	82.70% $\pm$ 0.75%
Accuracy	78.54% $\pm$ 0.71%
F1 Score	75.92% $\pm$ 0.59%

References

1. Jui, S.J.J.; Deo, R.C.; Barua, P.D.; Devi, A.; Soar, J.; Acharya, U.R. Application of entropy for automated detection of neurological disorders with electroencephalogram signals: A review of the last decade (2012–2022). *IEEE Access* **2023**, *11* , 71905–71924. [CrossRef]
2. Roy, S.; Kiral-Kornek, I.; Harrer, S. ChronoNet: A deep recurrent neural network for abnormal EEG identification. In Proceedings of the Artificial Intelligence in Medicine: 17th Conference on Artificial Intelligence in Medicine, AIME 2019, Poznan, Poland, 26–29 June 2019; Proceedings 17; Springer: Berlin/Heidelberg, Germany, 2019; pp. 47–56.
3. Amin, H.U.; Malik, A.S.; Ahmad, R.F.; Badruddin, N.; Kamel, N.; Hussain, M.; Chooi, W.T. Feature extraction and classification for EEG signals using wavelet transform and machine learning techniques. *Australas. Phys. Eng. Sci. Med.* **2015**, *38*, 139–149. [CrossRef]
4. Wei, L.; Boutouil, H.; Gerbatin, R.R.; Mamad, O.; Heiland, M.; Reschke, C.R.; Del Gallo, F.; Fabene, P.F.; Henshall, D.C.; Lowery, M.; et al. Detection of spontaneous seizures in EEGs in multiple experimental mouse models of epilepsy. *J. Neural Eng.* **2021**, *18*, 056060. [CrossRef]
5. Gao, Y.; Lee, H.J.; Mehmood, R.M. Deep learninig of EEG signals for emotion recognition. In Proceedings of the 2015 IEEE International Conference on Multimedia & Expo Workshops (ICMEW), Turin, Italy, 29 June–3 July 2015; pp. 1–5.
6. Schirrmeister, R.T.; Springenberg, J.T.; Fiederer, L.D.J.; Glasstetter, M.; Eggensperger, K.; Tangermann, M.; Hutter, F.; Burgard, W.; Ball, T. Deep learning with convolutional neural networks for EEG decoding and visualization. *Hum. Brain Mapp.* **2017**, *38*, 5391–5420. [CrossRef]
7. Wei, L.; Ventura, S.; Ryan, M.A.; Mathieson, S.; Boylan, G.B.; Lowery, M.; Mooney, C. Deep-spindle: An automated sleep spindle detection system for analysis of infant sleep spindles. *Comput. Biol. Med.* **2022**, *150*, 106096. [CrossRef] [PubMed]
8. Taye, M.M. Understanding of machine learning with deep learning: Architectures, workflow, applications and future directions. *Computers* **2023**, *12*, 91. [CrossRef]
9. Mukhlif, A.A.; Al-Khateeb, B.; Mohammed, M.A. An extensive review of state-of-the-art transfer learning techniques used in medical imaging: Open issues and challenges. *J. Intell. Syst.* **2022**, *31*, 1085–1111. [CrossRef]
10. Lin, Y.P.; Jung, T.P. Improving EEG-based emotion classification using conditional transfer learning. *Front. Hum. Neurosci.* **2017**, *11*, 334. [CrossRef] [PubMed]
11. Zhang, K.; Robinson, N.; Lee, S.W.; Guan, C. Adaptive transfer learning for EEG motor imagery classification with deep convolutional neural network. *Neural Netw.* **2021**, *136*, 1–10. [CrossRef]
12. Wei, L.; Mchugh, J.C.; Mooney, C. Transfer learning for the identification of paediatric EEGs with interictal epileptiform abnormalities. *IEEE Access* **2024**, *12*, 86073–86082. [CrossRef]
13. Kiani, M.; Andreu-Perez, J.; Hagras, H.; Rigato, S.; Filippetti, M.L. Towards understanding human functional brain development with explainable artificial intelligence: Challenges and perspectives. *IEEE Comput. Intell. Mag.* **2022**, *17*, 16–33. [CrossRef]
14. Antoniadi, A.M.; Du, Y.; Guendouz, Y.; Wei, L.; Mazo, C.; Becker, B.A.; Mooney, C. Current challenges and future opportunities for XAI in machine learning-based clinical decision support systems: A systematic review. *Appl. Sci.* **2021**, *11*, 5088. [CrossRef]

15. Mangalathu, S.; Hwang, S.H.; Jeon, J.S. Failure mode and effects analysis of RC members based on machine-learning-based SHapley Additive exPlanations (SHAP) approach. *Eng. Struct.* **2020**, *219*, 110927. [CrossRef]
16. Kumarakulasinghe, N.B.; Blomberg, T.; Liu, J.; Leao, A.S.; Papapetrou, P. Evaluating local interpretable model-agnostic explanations on clinical machine learning classification models. In Proceedings of the 2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS), Rochester, MN, USA, 28–30 July 2020; pp. 7–12.
17. Vinogradova, K.; Dibrov, A.; Myers, G. Towards interpretable semantic segmentation via gradient-weighted class activation mapping (student abstract). In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 13943–13944.
18. Lopez, S. Automated Identification of Abnormal EEGs. Master's Thesis, Temple University, Philadelphia, PA, USA, 2017. Available online: http://www.isip.piconepress.com/publications/ms_theses/2017/abnormal (accessed on 1 May 2025).
19. Obeid, I.; Picone, J. The temple university hospital EEG data corpus. *Front. Neurosci.* **2016**, *10*, 196. [CrossRef]
20. Lopez, S.; Suarez, G.; Jungreis, D.; Obeid, I.; Picone, J. Automated identification of abnormal adult EEGs. In Proceedings of the 2015 IEEE Signal Processing in Medicine and Biology Symposium (SPMB), Philadelphia, PA, USA, 12 December 2015; pp. 1–5.
21. Roy, S.; Kiral-Kornek, I.; Harrer, S. Deep learning enabled automatic abnormal EEG identification. In Proceedings of the 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Honolulu, HI, USA, 17–21 July 2018; pp. 2756–2759.
22. Cisotto, G.; Zanga, A.; Chlebus, J.; Zoppis, I.; Manzoni, S.; Markowska-Kaczmar, U. Comparison of attention-based deep learning models for eeg classification. *arXiv* **2020**, arXiv:2012.01074. [CrossRef]
23. Yıldırım, Ö.; Baloglu, U.B.; Acharya, U.R. A deep convolutional neural network model for automated identification of abnormal EEG signals. *Neural Comput. Appl.* **2020**, *32*, 15857–15868. [CrossRef]
24. Peh, W.Y.; Thomas, J.; Bagheri, E.; Chaudhari, R.; Karia, S.; Rathakrishnan, R.; Saini, V.; Shah, N.; Srivastava, R.; Tan, Y.L.; et al. Five-institution study of automated classification of pathological slowing from adult scalp electroencephalograms. *arXiv* **2020**, arXiv:2009.13554.
25. Kiessner, A.K.; Schirrmeister, R.T.; Gemein, L.A.; Boedecker, J.; Ball, T. An extended clinical EEG dataset with 15,300 automatically labelled recordings for pathology decoding. *NeuroImage Clin.* **2023**, *39*, 103482. [CrossRef] [PubMed]
26. Margerison, J.; Corsellis, J. Epilepsy and the temporal lobes. *Brain* **1966**, *89*, 499–530. [CrossRef]
27. Cavanagh, J.F.; Frank, M.J. Frontal theta as a mechanism for cognitive control. *Trends Cogn. Sci.* **2014**, *18*, 414–421. [CrossRef]
28. Barbur, J.; Moro, S.; Harlow, J.; Lam, B.; Liu, M. Comparison of pupil responses to luminance and colour in severe optic neuritis. *Clin. Neurophysiol.* **2004**, *115*, 2650–2658. [CrossRef] [PubMed]
29. Albaqami, H.; Hassan, G.M.; Datta, A. Automatic detection of abnormal eeg signals using wavenet and lstm. *Sensors* **2023**, *23*, 5960. [CrossRef]
30. Tuncer, T.; Dogan, S.; Acharya, U.R. Automated EEG signal classification using chaotic local binary pattern. *Expert Syst. Appl.* **2021**, *182*, 115175. [CrossRef]
31. Gemein, L.A.; Schirrmeister, R.T.; Chrabąszcz, P.; Wilson, D.; Boedecker, J.; Schulze-Bonhage, A.; Hutter, F.; Ball, T. Machine-learning-based diagnostics of EEG pathology. *NeuroImage* **2020**, *220*, 117021. [CrossRef] [PubMed]
32. Kiessner, A.K.; Schirrmeister, R.T.; Boedecker, J.; Ball, T. Reaching the ceiling? Empirical scaling behaviour for deep EEG pathology Classification. *Comput. Biol. Med.* **2024**, *178*, 108681. [CrossRef]
33. Wei, L.; Mchugh, J.C.; Mooney, C. Interictal Epileptiform Discharge Classification for the Prediction of Epilepsy Type in Children. In Proceedings of the 2023 10th International Conference on Biomedical and Bioinformatics Engineering, Kyoto, Japan, 9–12 November 2023; pp. 180–184.
34. Wei, Y.; Zhu, Y.; Zhou, Y.; Yu, X.; Luo, Y. Automatic sleep staging based on contextual scalograms and attention convolution neural network using single-channel EEG. *IEEE J. Biomed. Health Inform.* **2023**, *28*, 801–811. [CrossRef]
35. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July, 2017; pp. 4700–4708.
36. Albelwi, S.A. Deep architecture based on DenseNet-121 model for weather image recognition. *Int. J. Adv. Comput. Sci. Appl.* **2022**, *13*, 559–565. [CrossRef]
37. Rochmawanti, O.; Utaminingrum, F. Chest X-ray image to classify lung diseases in different resolution size using DenseNet-121 architectures. In Proceedings of the 6th International Conference on Sustainable Information Engineering and Technology, Malang, Indonesia, 13–14 September 2021; pp. 327–331.
38. Montoya, X.; Díaz, F.; Félix, J.; Paucar, J.; Ferrer, J.; Fonseca, P. Seizure detection by analyzing the number of channels selected by cross-correlation using TUH EEG seizure corpus. In Proceedings of the 18th International Symposium on Medical Information Processing and Analysis, Valparaiso, Chile, 9–11 November 2022; Volume 12567, pp. 397–406.
39. Wang, Y.; Veluvolu, K.C. Evolutionary algorithm based feature optimization for multi-channel EEG classification. *Front. Neurosci.* **2017**, *11*, 28. [CrossRef]

40. Türk, Ö.; Özerdem, M.S. Epilepsy detection by using scalogram based convolutional neural network from EEG signals. *Brain Sci.* **2019**, *9*, 115. [CrossRef] [PubMed]
41. Aslan, Z.; Akin, M. A deep learning approach in automated detection of schizophrenia using scalogram images of EEG signals. *Phys. Eng. Sci. Med.* **2022**, *45*, 83–96. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

algorithms

MDPI

Article

Affordable Audio Hardware and Artificial Intelligence Can Transform the Dementia Care Pipeline

Ilyas Potamitis

Department of Music Technology and Acoustics, Hellenic Mediterranean University, 71410 Heraklion, Greece; potamitis@hmu.gr

Abstract

Population aging is increasing dementia care demand. We present an audio-driven monitoring pipeline that operates either on mobile phones, microcontroller nodes, or smart television sets. The system combines audio signal processing with AI tools for structured interpretation. Preprocessing includes voice activity detection, speaker diarization, automatic speech recognition for dialogs, and speech-emotion recognition. An audio classifier detects home-care–relevant events (cough, cane taps, thuds, knocks, and speech). A large language model integrates transcripts, acoustic features, and a consented household knowledge base to produce a daily caregiver report covering orientation/disorientation (person, place, and time), delusion themes, agitation events, health proxies, and safety flags (e.g., exit seeking and falling). The pipeline targets real-time monitoring in homes and facilities, and it is an adjunct to caregiving, not a diagnostic device. Evaluation focuses on human-in-the-loop review, various audio/speech modalities, and the ability of AI to integrate information and reason. Intended users are low-income households in remote settings where in-person caregiving cannot be secured, enabling remote monitoring support for older adults with dementia.

Keywords: AI; elderly care; dementia; LLM; ESP32

1. Introduction

Population aging is accelerating worldwide, shifting the balance between those needing care and the working-age population available to provide it. Globally, dementia imposes substantial health-system and societal costs—estimated at USD 1.3 trillion in 2019 and projected to rise sharply in coming decades—while informal carergivers already deliver roughly half of total care hours [1–4]. In Europe, older adults ($\geq$65 years) accounted for 21.6% of the EU population on 1 January 2024, and projections indicate a sustained increase in the old-age dependency ratio through mid-century [2]. United Nations projections similarly forecast a rapid expansion of the $\geq$65 population over the next three decades [3]. Against this backdrop, enabling safe, scalable home-based care for people living with dementia is a pressing priority, both to preserve independence and to mitigate escalating institutional and family costs [1,4]. In many cases, in-home caregiving is provided by workers with limited or no formal training at all; services can be expensive, and reliable, high-quality support is not readily available on demand. In this work, we explore the integration of audio-based signal processing with new possibilities offered by artificial intelligence (AI) tools within the care pipeline for elderly patients with dementia. Our aim is to leverage recent advances in AI, and large language models (LLMs) in particular, to deliver an affordable service designed primarily for people of low-income and/or remote

locations where full-time human care is not practically feasible. Our approach is also applicable to the important and pressing case of childless elderly people they live alone and develop mobility and mental health conditions.

Ambient audio sensing is an attractive modality for home care: microphones are inexpensive, unobtrusive, and easily installed. They can capture both speech and non-speech events (e.g., coughs, thuds, alarms, cane taps, and expressions of pain or distress) without the visual privacy trade-offs of cameras and without the adherence burden of wearables. Prior research has demonstrated that speech carries information relevant to cognitive status. Early work using automatic speech analysis differentiated healthy controls, mild cognitive impairment (MCI), and Alzheimer's disease (AD), establishing the feasibility of acoustic biomarkers [5]. Subsequent reviews and empirical studies have evaluated acoustic (paralinguistic) features—prosody, timing, and pauses—alone and alongside linguistic variables for AD screening and monitoring, including remote collection paradigms [6–9]. Recent studies have further explored non-semantic, acoustic-only features and the feasibility and test–retest reliability of multi-day, remote assessments of speech acoustics, including associations with amyloid status and deep-learning approaches to voice recordings [10–15]. Collectively, this body of work supports speech-based digital biomarkers as a noninvasive window into cognitive health, but translation to day-to-day home-care workflows remains limited [8].

Beyond speech, audio has also been investigated for characterizing behavioral and psychological symptoms of dementia, and care environments. Persistent or inappropriate vocalizations are a common and burdensome symptom in advanced dementia and integrative reviews synthesize their phenomenology and implications for care [16]. Soundscape interventions in nursing homes show that targeted monitoring and staff feedback can improve acoustic environments and staff evaluations, suggesting that environmental audio is a modifiable determinant of wellbeing [17]. Concurrently, the growing literature examines audio-based detection of safety-critical events such as falls [18,19], spanning classical machine-learning methods and transformer-based approaches and complemented by reviews of fall-detection methods and wearable sensing for older adults [20,21]. Related work has demonstrated real-time acoustic detection of critical incidents on edge devices, underscoring feasibility for low-latency deployment [22]. However, most prior studies focus on single tasks (e.g., diagnostic screening or environmental monitoring), use longer or highly structured recordings, or are conducted in clinical or institutional settings rather than ordinary homes [23–28]. Integration of heterogeneous audio evidence into actionable, caregiver-facing summaries is rarely addressed.

Smart-home and telecare research for older adults has advanced through successive projects that illustrate different sensing and inference paradigms. An adaptive framework for activity recognition in domestic environments incorporated user feedback to refine models, thereby embedding personalization into resident-centered automation [29,30]. Practical issues of scaling were examined in multi-country deployments of assistive equipment for people with dementia, where installation procedures and technical support were identified as critical for sustained operation [31]. Inference methods progressed with plan-recognition models capable of interpreting incomplete observations to detect disorientation and atypical behaviors in Alzheimer's patients [32]. Multimodal capture of daily living activities was enabled through wearable audio-video devices, with indexing strategies to allow efficient clinical review of extended recordings [33]. For nocturnal monitoring, unobtrusive load-cell systems placed under beds were developed to quantify movement and posture changes, supporting analysis of sleep patterns and risk states [34]. Activity recognition in naturalistic home settings using simple state-change sensors demonstrated that ubiquitous instrumentation could yield meaningful behavior classification without

requiring cameras [35]. Within clinical contexts, models were proposed for quantifying patient activity in hospital suites, producing indicators such as mobility levels and displacement distributions that could be tracked continuously [36]. Earlier, lifestyle-monitoring systems established the feasibility of using ambient data streams to identify deviations from typical behavior and support independent living [37]. Finally, integrated e-healthcare prototypes in experimental "techno houses" are suggested in [38] combined physiological and environmental sensors to enable continuous monitoring across multiple aspects of daily life, including personal hygiene and rest [39,40].

This work targets that translational gap by proposing and evaluating a home-based pipeline that couples audio signal processing with an Audio Spectrogram Transformer (AST) [41] trained on the Audioset database [42] for non-speech event detection and a LLM (ChatGPT-5.1) constrained to emit auditable, structured reports. Speech and non-speech events salient for home care—coughs (including chronic cough burden), cane-tap sequences (as a mobility proxy via cadence variability), thuds, appliance beeps, and alarms—are time-stamped by AST and exported as a JSON record. The LLM receives as text input the context of the application, dialog transcripts, acoustic features, and event metadata together with a small, consented household knowledge base and returns reports covering orientation/disorientation (person/place/time), delusion themes and their persistence (e.g., fixed beliefs about "another home"), agitation likelihood, comprehension difficulty, instruction follow-through, and safety flags (e.g., possible fall). Unlike diagnostic-only approaches [43–50], the outputs are designed to guide daily care (e.g., de-escalation prompts, and check-ins) and to populate longitudinal dashboards for trend analysis.

Our contributions are threefold. First, we integrate speech and non-speech audio within a home pipeline, moving beyond single-task speech-only screening to multitask monitoring that reflects real-world caregiving priorities (safety, agitation, and communication efficacy). We use recently developed AI tools that were not available ten years ago (i.e., transformers, large databases, and LLMs) and introduce them to assisted care services. Second, we develop a real-time, affordable hardware solution, and we open source its code. Third, we use an LLM to integrate explicitly separate acoustic observations to reach a higher level of interpretation that cannot be achieved using single audio transcription—for example, detecting falls and delusional themes. In aggregate, the combination of audio processing, transformer classifiers, and LLM receiving input from real patients reframes ambient home audio from a raw signal into validated, longitudinal indicators of safety and wellbeing, complementing—not replacing—clinical assessment while addressing pressing needs in dementia care. A depiction of our approach can be seen in Figure 1.

This work integrates pretrained components and applies them in a zero-shot manner, with few-shot enrollment only for diarization. There is no training or finetuning on private data. The design reflects ethical and practical constraints that make large-scale end-to-end training on speech from older adults with dementia infeasible and difficult to scale (limited labeled data, privacy and consent requirements, participant burden, and annotation cost). Accordingly, we state explicitly the following: (i) all neural and transformer-based models remain frozen (i.e., AST, ASR, VAD, LLMs, and emotion recognition); (ii) emotion recognition specifically is zero-shot; and (iii) diarization uses few-shot enrollment without training from scratch or fine-tuning per se.

This study should be interpreted as a dense, longitudinal case study of two individuals rather than a generalizable clinical trial. Our conclusions are restricted to engineering feasibility and within-subject trend tracking. In Section 5.1 Future Prospects, we discuss multi-site studies with larger, diverse cohorts that will be required before any claims about diagnostic performance or broad deployment are justified.

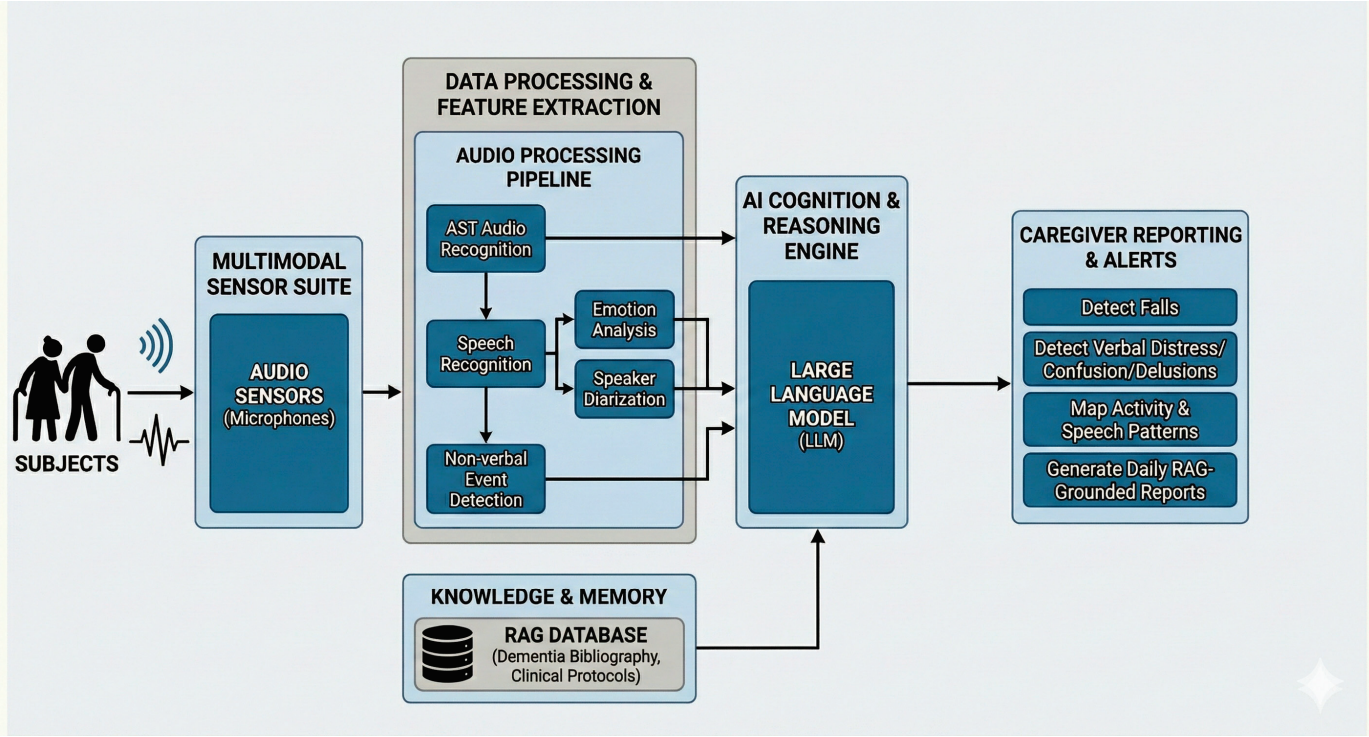


Figure 1. Architecture of the audio-based dementia care system. The system utilizes an Audio Spectrogram Transformer (AST) to classify audio events. Verbal instances undergo parallel processing for emotion recognition, speaker diarization, and Speech Recognition (ASR). Both the processed verbal data and classified non-verbal events are fed into a Large Language Model (LLM). The LLM integrates this multi-modal information, applies reasoning to detect critical events (e.g., falls, delusions, pain), and generates structured reports and alerts for timely caregiver notification.

2. Materials and Methods

The study participants are a couple living in their own home—a woman aged eighty-seven and a man aged ninety (as of 2025)—both formally diagnosed with stage-2 dementia. The main symptoms are as follows: increasing difficulty recalling recent events; forgetting names of familiar people; word-finding problems; short sentences; confusion about time and place; and repetitive questioning. The female subject has developed delusions and increased risk for falls. Both subjects have additional mobility problems and use canes and other mobility aids.

The subjects are attended by two caregivers in two 8 h shifts per day and occasionally receive additional care from practitioners. The establishment has in-house cameras with audio in all rooms, attended by professional security personnel during nights. Due to low mobility, the couple moves only to one floor and only to the bedroom, the bathroom, and the living room.

This study is a proof-of-concept demonstrating the feasibility of an audio processed by AI pipeline for home dementia care. Our goal is methodological and systems-oriented—to show that ambient audio can be captured on very low-cost hardware, transformed into text labels corresponding to the recognized audio events out of 527 audio classes, and summarized by an LLM (in this work ChatGPT 5.1 in thinking mode) that receives the labels and the context of the application into auditable, caregiver-facing outputs. We do not claim population-level diagnostic performance. Instead, we establish engineering viability, safety guardrails, and a structured scheme for longitudinal monitoring.

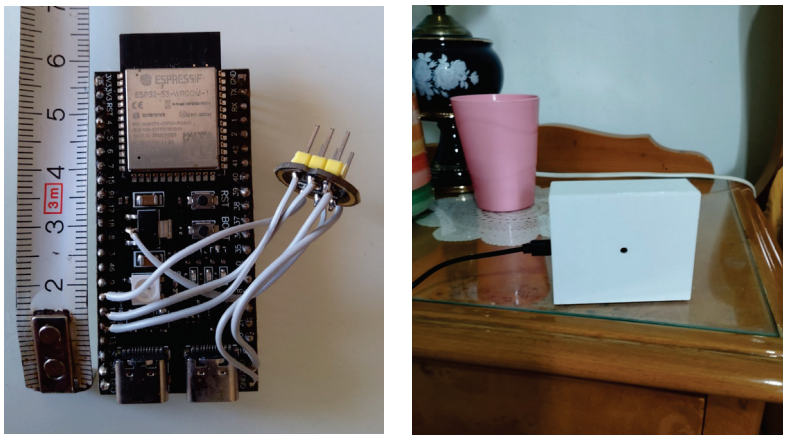
Although the cohort is small, the dataset is longitudinal and dense (thousands of 15 s snippets across months), enabling robust within-subject demonstrations of trend tracking

(e.g., cough burden, exit-seeking themes, and cane-tap variability). For personalization-centric care, repeated measures from the same individuals are more informative than sparse cross-sectional samples.

We acknowledge that two participants do not support population-level inference. However, we believe that reported evidence in this work justifies the position that LLM and suitable AI-bots can provide significant value to this problem by first monitoring and, in the near future, engaging the task with actions. Accordingly, we perform the following: (i) report per-subject results and descriptive statistics rather than pooled significance tests; (ii) release code and disidentified audio to enable replication (see Appendix A); and (iii) outline a prospective, multi-site study with power considerations for clinical endpoints (e.g., sensitivity to agitation/exit-seeking and caregiver workload impact). This sequencing—proof-of-concept → pilot → powered trial—is standard for translational systems entering safety-critical care.

2.1. Hardware

The approach presented is based on the ESP32 and MEMS type microphone that are very low cost. We implemented a distributed architecture that supports cloud, on-premises, and edge deployments. A network of ESP32-S3-DevKitC-1-N8R8 nodes (Espressif Systems, Shanghai, China) with SPH0645LM4H digital microphones (Syntiant Corp., Irvine, CA, USA) transmits MP3 compressed 15 s audio snippets over Wi-Fi to a secure remote server for classification (see Figure 2).



(**a**) (**b**)

Figure 2. Hardware and deployment of the audio monitoring node. (**a**) ESP32-S3 low-power audio node. The node integrates an ESP32-S3 microcontroller with a microphone attachment. The device operates in a continuous, audio-triggered state. Upon activation, it captures a 15 s audio segment at a 16 kHz sampling rate, compresses the data into the MP3 format, and streams the packet over Wi-Fi to a local server for subsequent processing. (**b**) Bedside Deployment. A typical deployment location showing the device positioned on the bedside table. This placement ensures optimal proximity for continuously monitoring both sleep-related events (e.g., coughing and breathing), speech, paralinguistic features and general audio (cane hits, thump/thud/hits, and glass breaking).

The SPH0645LM4H provides low-power, bottom-port I^2S output, eliminating the need for an external codec. Its flat 0.1–8 kHz response aligns with our 16 kHz sampling rate for AST-based analysis. The ESP32-S3, a dual-core Xtensa LX7 MCU with Wi-Fi/BLE, vector instructions, ample GPIO, and I^2S, was selected to handle audio capture and MQTT. The total cost of the device is at the order of EUR 20 (as per 14 August 2025). A more detailed description of the hardware can be found in [22]. However, the same service can be provided by a spare smartphone at zero extra cost that streams audio snippets to Wi-Fi, captured by a voice activity detector (VAD).

2.2. Embedded Software in ESP32

Using Espressif IoT Development Framework version 5.3 (ESP-IDF 5.3, Espressif Systems, Shanghai, China), the embedded nodes in C language capture audio from the microphone via the Inter-IC Sound (I^2S) digital audio interface in non-overlapping 1024-sample windows at 16 kHz. Then, they perform basic level checks using the Root Mean Square (RMS) value, encode the audio into MP3 format with an embedded MP3 Encoder (LAME) library (The LAME Project, San Diego, US), and publish the compressed data using the Message Queuing Telemetry Transport (MQTT) protocol secured with Transport Layer Security (TLS). Processing is divided between the two cores of the ESP32-S3 microcontroller (Core 0 handles capture and processing, while Core 1 handles encoding and transmission), with a ring buffer used for inter-task communication. Configuration is stored in Non-Volatile Storage (NVS) and supports remote updates via MQTT. Pseudo-Static Random Access Memory (PSRAM) is used for efficient buffering. Networking is provided over Wi-Fi by default to keep communications' cost low.

MQTT was chosen for its reliability under constrained bandwidth and variable link quality. Streams are handled by a multithreaded MQTT client with Quality of Service (QoS) tuning, buffering, error detection, and fast reconnection protocols. Latency is reduced through optimized packetization, I^2S Direct Memory Access (DMA) buffering, and adaptive send intervals. A User Datagram Protocol (UDP) fallback is enabled if MQTT delays increase. Local brokers operating over 2.4 GHz Wi-Fi minimize Wide Area Network (WAN) hops, while server-side asynchronous processing and Network Time Protocol (NTP) synchronization maintain throughput and temporal alignment across nodes.

At the server, MP3 clips are decoded using Fast Forward Moving Picture Experts Group (ffmpeg), converted into spectrograms via the Short-Time Fourier Transform (STFT), and rendered as time–frequency images for the Audio Spectrogram Transformer (AST). The AST, which follows a Vision Transformer (ViT) architecture, operates directly on 10 s spectrogram patches for audio event classification. Spectrograms for spoken snippets and environmental events (e.g., discussions, cough, and cane taps) are processed to produce labeled, timestamped outputs suitable for downstream integration.

2.3. The Elders with Dementia Dataset

The ESP devices work 24/7, and depending on the activity inside home, several hundred recordings can be produced per day. The recording sessions started on 4 August 2025 and continue up to now, reaching several thousands of recordings. Recordings contain mainly speech, non-verbal human sounds, and impact sounds of various objects (e.g., drawers, closets, doors, switches, mobility aids). The database they produced is the subject of this study.

2.4. The Audio Event Recognizer

The Audio Spectrogram Transformer (AST) [41] formulates audio tagging as patch-based transformer classification on log-Mel spectrograms and is commonly pre-trained or fine-tuned on AudioSet. Concretely, a waveform is converted to 128-dimensional log-Mel filter banks using a 25 ms window with 10 ms hop. The 2-D time–frequency map is then split into overlapping 16×16 patches that are linearly projected to tokens, augmented with positional embeddings, and processed by a ViT-style encoder (12 layers, 12 heads, 768 dim embeddings). The model is initialized from ImageNet ViT weights via patch and positional-embedding adaptation and optimized with binary cross-entropy using standard audio augmentations (mix-up, SpecAugment) and weight-averaged checkpoints. Evaluated on AudioSet—the 2.1 M-clip, 5.8 k h, 527-class collection of human-labeled 10 s YouTube excerpts—AST reports mean average precision (mAP) 0.4590 for a single

weight-averaged model and up to 0.4850 with model assembling on the full training split, indicating that transformer attention over spectrogram patches is competitive with or superior to CNN/attention hybrids for weakly labeled sound event recognition. In this configuration, AST leverages AudioSet's breadth to learn general sound representations that transfer to downstream tagging tasks without architectural changes.

AudioSet [42] is a large-scale, ontology-driven dataset of human-labeled sound events released by Google, comprising over 2 million 10 s audio clips drawn from YouTube videos. Its ontology includes 527 distinct classes spanning speech, environmental sounds, animal vocalizations, and human activities, providing a broad coverage for training and evaluating audio event recognition models. The AudioSet ontology includes the sounds that daily activity care produces, namely, impact, collisions, and knocks labels that are falls/accidents proxies. Specifically, the Knock, Tap, Bang, Thump/Thud, Slam, Smash, Crash, Clatter, and Shatter (glass) are included in the ontology. Regarding sounds of human vocal distress, pain, and agitation that can be associated with elderly activity and accidents, the following labels are included in the ontology and are semantically related: Screaming, Yell, Shout, Wail/moan, Whimper, Groan, Gasp, and Sigh. We are also interested in monitoring chronic conditions related to respiratory and health conditions and the following are included in the available set of labels: Cough, Sneeze, Breathing, Snoring, Throat Clearing, and Wheeze which appear under respiratory sounds in the ontology as well. Regarding mobility and movement (activity/restlessness cues), the labels are as follows: Footsteps, Bouncing, Wood, Stomp, Stamp, Surface Contact (e.g., scrape and scratch), Creak (e.g., floorboards/chairs), and Door (open/close). Safety/alerting signals (hazard or attention) are included in the ontology. Smoke detector, Smoke Alarm/Fire Alarm, Alarm Clock, Siren, Buzzer, Beep, Bleep, Doorbell, and Telephone Ringing. Conversation and caregiver interaction (dialog state cues) are handled by the Speech, Conversation, Narration, Monolog, Whispering, and loudness-related (Shout/Yell) categories.

The wide representation of audio categories in Audioset ontology practically covers all cases appearing in the context of our work (see Appendix A.1).

2.5. Voice Activity Detection

A VAD marks the part of the recording that is related to speech. There is a vast number of algorithms and implementations available. We have chosen a new approach from the TEN-VAD that has demonstrated better results compared with other state-of-the-art algorithms (see Appendix A.2). This algorithm is a small hybrid digital signal processing (DSP) followed by a recurrent neural network VAD. It is lightweight, supports streaming, and uses classic signal processing features and a small LSTM-style ONNX model for sequence modeling. It computes a 16 kHz STFT with a Hann window, applies a 40 band Mel filterbank, logs/normalizes features, and appends a pitch feature. These per-frame features are stacked over a short temporal context. The model emits a frame-level speech probability and a binary VAD flag after thresholding.

2.6. Ordinary Automatic Speech Recognition Models Do Not Currently Work Properly for Elders with Dementia

Whisper large-v3 is an open, transformer-based encoder–decoder ASR model that delivers strong accuracy across many languages with competitive inference speed. Automatic Speech Recognition (ASR) with Whisper 3-large (see Appendix A.3) underperformed on conversational Greek from older adults with dementia because multiple, interacting factors degrade the signal and violate the model's assumptions. We need to note that transcription accuracy varies among elder individuals and the various states of dementia. While baseline testing on healthy elderly subjects (aged 60–80+) showed high transcription fidelity that match the rates reported for Whisper v3 for the Greek language (see

https://github.com/openai/whisper, accessed on 1 November 2025), the ASR accuracy for the specific couple in this study is extremely low, with the Word Error Rate (WER) consistently exceeding 90% and the Sentence Error Rate (SER) approaching saturation 99% accuracy loss. We subsequently analyze the reasons for this failure. First, speaker physiology changes with age—presbyphonia, breathy or tremulous phonation, reduced articulatory precision from dentures or dry mouth, and frequent comorbid dysarthria—alter formant structure and sibilants and reduce the effective signal-to-noise ratio, making phonetic cues less distinct. Second, cognitive–linguistic patterns characteristic of dementia (see also Figure 3)—long hesitations, false starts, repetitions, elliptical utterances, and frequent repair sequences—disrupt the fluent, clause-level structures that end-to-end ASR models implicitly expect from their training data, increasing deletions and punctuation errors. Third, the home acoustic environment adds nonstationary interference: televisions and radios, and hard-surface reverberation all produce crosstalk and spectral interference that cannot be discriminated by front-end feature extraction. Fourth, Greek-specific variability—regional accents, intonation patterns, and rich inflectional morphology with clitics and enclitics—is underrepresented in common web-scale corpora, so systems tuned on broadcast or YouTube speech from younger speakers generalize poorly to elderly, informal Greek dialog. Fifth, pipeline constraints compound errors: tight voice-activity detection and diarization cuts fragment utterances and 10–15 s windows limit language-model context for disfluency recovery. In combination, these factors yield low confidence scores, insertion/deletion spikes, and semantic drift in transcripts, which in turn mislead downstream LLM components tasked with behavioral inference. Model adaptation of transformers to include the speech of elders is feasible but nontrivial: since systems like Whisper are trained on paired audio–text annotations, effective domain adaptation would require a sizable, consented corpus of elderly Greek conversational speech with accurate transcripts spanning dialects, acoustic conditions, and symptom severities—resources that are rarely available and costly to curate (see [51] for an adaptation of ASR models to elders using a 12 h adaptation corpus) and [52–54] for specifically adapting to dysarthric speech.

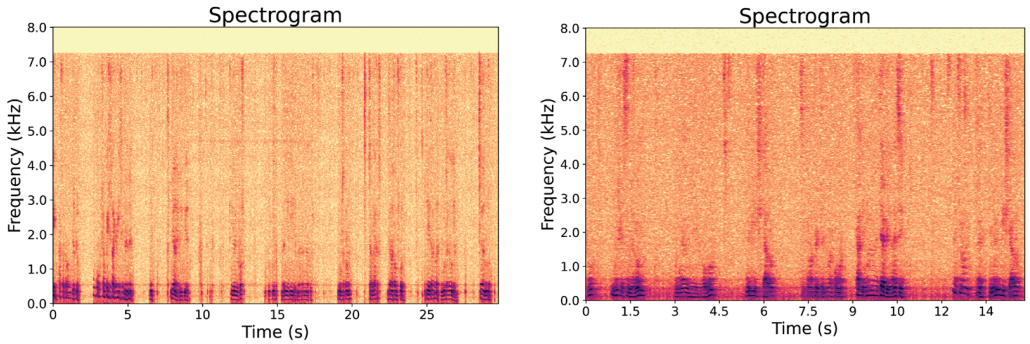


Figure 3. Characteristic spectrograms of speech from elders with dementia. *y*-axis the frequencies of speech and *x*-axis time in seconds. (**Left**) An example of female elder speech (Event_id: 186692). (**Right**) Male elder (Event_id: 172738). Both speakers were 1–1.5 m away from the microphone. Notice the fading of frequencies above 600 Hz and the prolonged pauses between vocalizations.

As shown in Figure 3, log-spectrograms are time-frequency representations of audio signals widely used in speech processing and recognition. A spectrogram is produced by first calculating the Short-Time Fourier Transform (STFT) of a signal, which yields the signal's energy distribution across various frequencies over short, overlapping time segments. The resulting magnitude plot is then processed by applying logarithmic compression to the values. This step is critical and conforms to the way the human ear perceives loudness. Log-spectrograms effectively compress the high-energy components (such as strong speech vowels) while simultaneously amplifying the subtle, information-rich low-energy

components (such as fainter fricatives or consonants). This conversion transforms the 1D audio signal into a 2D visual representation that acts as a "picture" input for machine learning models. This visibility and 2D format make log-spectrograms an ideal front-end for models like Transformers, leading to robust performance in tasks such as Automatic Speech Recognition (ASR), speaker diarization, and emotion identification.

2.7. Speech Diarization and Speaker Recognition

Speaker diarization splits a clip into speaker-homogeneous time segments and can flag overlaps. It does not know the identities of people talking. Speaker recognition is a different task and tags "who is this speaker." It matches a segment to an enrolled voiceprint (verification/identification), assuming the segment is single speaker. Determining who spoke when—is essential in monitoring elderly people with dementia because it attributes each utterance to the correct person (e.g., patient, caregiver, visitor). There are many approaches and implementations dealing with this task, but in our case, few-shot learning seems the appropriate solution. Few-shot learning relies on few samples to build a model for each speaker (~20 recordings of 15 s each), and the in-home application is expected to deal with a small number of speakers. In our setting, speaker tagging is convenient since the AST selects out speech that is subsequently tagged by a label of few speakers (six in our case: two participants, two caregivers, and two relatives) or any of their combinations. By segmenting audio into speaker-specific streams before ASR, diarization reduces attribution errors (such as mistaking a caregiver prompt for a patient belief) and yields cleaner, per-speaker transcripts for the LLM. This improves LLM's reasoning about orientation, agitation, and instruction follow-through, enables role-aware prompts and alerts, and supports longitudinal tracking of patient speech while preserving context about overlapping talk and background voices. Speech diarization has been quite successful in our experiments because, as usually the case is, there were a limited number of people in a house of elders. This allows modeling of the voices based on few examples and reliable attribution of speech queues. In this work we take the pyannote approach that implements diarization as a three-stage neural pipeline—local speaker segmentation, speaker embedding, and global clustering—that operates directly on mono 16 kHz audio and returns a time-stamped "who-spoke-when" annotation (see Appendix A.4). In version 2.1, the default pipeline applies a neural speaker segmentation model on short sliding windows (≈5 s with 0.5 s hop), producing frame-level (≈16 ms) posteriors for the activity of up to Kmax speakers. A related "powerset" segmentation variant explicitly models single- and two/three-speaker overlap classes to handle simultaneous talkers. Local speaker traces are then converted to speaker embeddings and aggregated globally via hierarchical agglomerative clustering (AHC) to obtain a consistent speaker inventory for the whole recording. The number of speakers can be estimated automatically or constrained by the user when known. The released embedding model is an x-vector TDNN architecture augmented with SincNet learnable filterbanks (trainable since convolutions replacing fixed Mel filters), yielding compact speaker representations suited to clustering. The toolkit also exposes overlapped-speech detection that can be combined with the pipeline to better account for concurrent speech. Collectively, this design frames diarization as end-to-end neural segmentation on overlapping windows, representation learning for speakers, and global AHC on embeddings—rather than K-means on spectra or hand-engineered MFCCs—to robustly infer speaker turns.

2.8. Emotion Recognition

In elderly people with dementia, depending on the age, health problems, and dementia stage, the emotional expression often differs significantly from that of younger populations and may be subtle, context-dependent, and shaped by the progression of cognitive decline. Neutral or baseline states, which can blend into apathy, are common and characterized by long silences, flat prosody, and reduced motivation or expressivity. Moments of happiness or contentment do occur, often triggered by familiar people, music, or reminiscence, and may be marked by brief laughter or a brighter tone, even singing. Sadness and depressive states manifest as softer speech, a slower pace, sighing, or tearfulness and are often accompanied by social withdrawal. Anger, irritability, and agitation are also possible, presenting as raised or shaky pitch, louder vocal bursts, or interruptions, particularly during care tasks, refusals, or episodes of confusion. Fear and anxiety are typically expressed through a tense or pressed voice, vocal tremor, and repetitive questioning and are especially prevalent during sundown. Apathy, a low-arousal negative state distinct from simple neutrality, can be pervasive, with minimal speech initiation and muted effect.

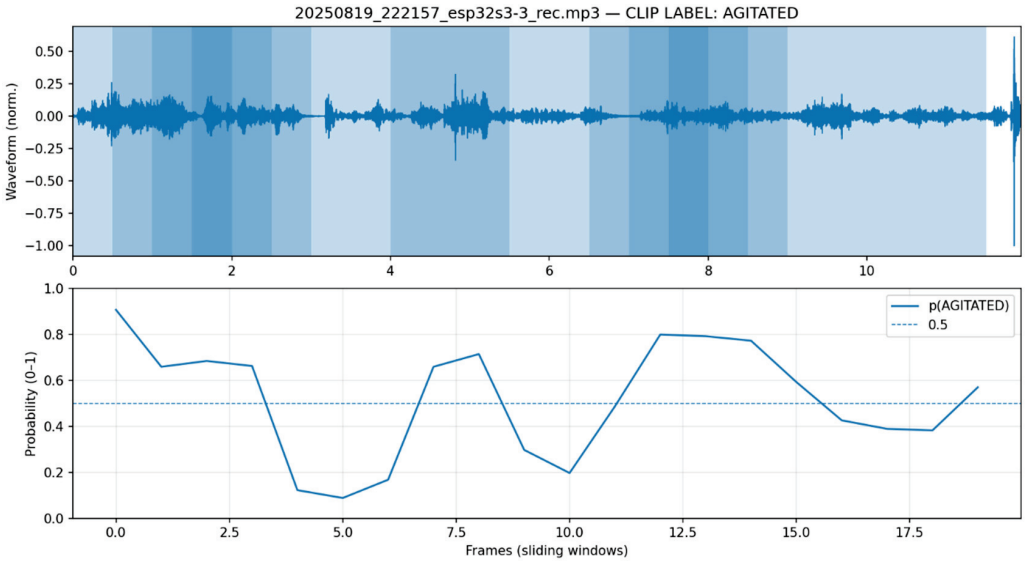
In this study, emotion recognition algorithms are based on signal processing audio only and extract low-level acoustic descriptors (pitch, loudness, speech rate, and rhythm, etc.) and then use statistical models to classify into emotions (e.g., neutral, happy, sad, angry, and afraid). Since they rely on acoustic–prosodic features (i.e., paralinguistic cues) rather than speech content, they can be largely language-independent—meaning they can work on any language, but accuracy may be slightly lower if the prosody differs from the languages it was trained on (typically English). Since the algorithms are not ASR-based, unclear or slurred speech will not cause the same errors as in transcription. Note that speech melody and expression vary between languages and cultures, which can affect accuracy. For example, Greek has naturally higher pitch variation than English, which might be misread as "excited" by some models. Many speech-based emotion recognition datasets use young-to-middle-aged actors. Age-related changes in the voice (pitch lowering in women, reduced clarity, slower speech) can cause misclassification. Moreover, in noisy, reverberant environments accuracy drops.

In this work we use the HuBERT-Large backbone fine-tuned for the SUPERB Emotion Speech Recognition (ESR) task (see Appendix A.5). The base encoder is hubert-large-ll60k, a self-supervised model trained on approximately 60,000 h of 16 kHz speech (LibriLight). During pretraining, HuBERT learns robust speech representations by masking spans of audio and predicting discrete pseudo-labels obtained from clustering acoustic features. For ESR, a lightweight classification head is attached to the frozen or partially fine-tuned encoder. The resulting checkpoint accepts raw 16 kHz mono waveforms and outputs posterior probabilities over four emotions. We did not fine-tune our data.

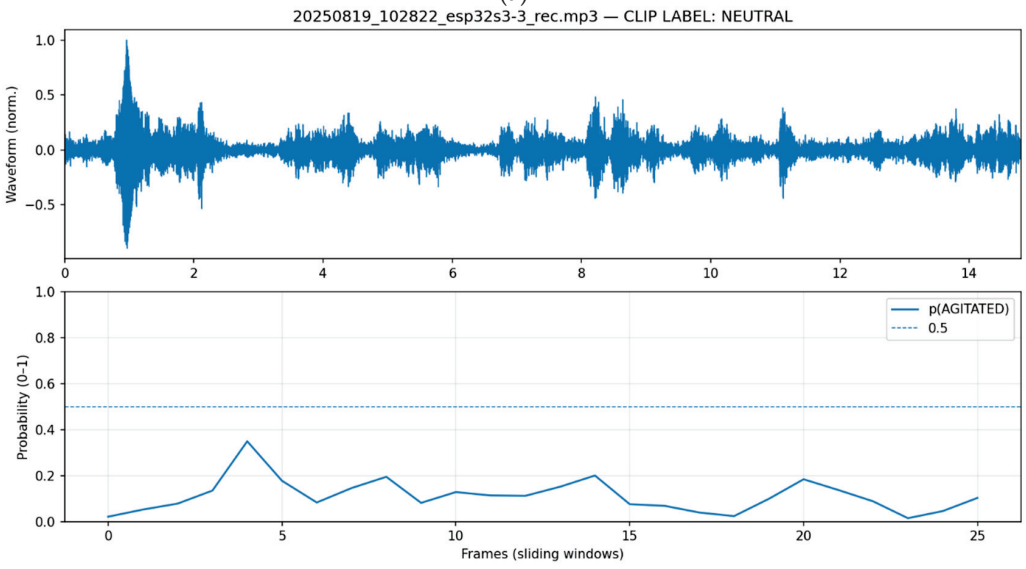
In the SUPERB benchmark, the ESR task is framed as a four-class utterance-level classification with the categories angry, happy, sad, and neutral. The widely used setup draws on the Interactive Emotional Dyadic Motion Capture (IEMOCAP) database and follows the common practice of focusing on these four balanced classes (other minority labels are omitted). The superb/hubert-large-superb-er checkpoint is fine-tuned to predict the label set—angry, happy, sad, and neutral.

Old adults with dementia do not demonstrate the expressivity of the young. Therefore, we merge the initial labels to improve robustness in this task. To meet the binary requirement of the application, we map categories into two states: AGITATED includes angry and happy, while NEUTRAL includes neutral and sad. Regarding the decision procedure, to handle long, real-world recordings, we apply sliding-window inference. Each input file is segmented into fixed windows (default 2.0 s) with overlap (0.5 s hop). For every window, the model produces a probability distribution over the four emotions. For each

window we compute p(agitated) as the sum of the angry and happy probabilities, and we mark the window as AGITATED when p(agitated) ≥ 0.5. A clip-level decision then uses a simple fraction rule: if at least 20% of windows are AGITATED, the entire clip is labeled AGITATED (see Figure 4a), otherwise, it is NEUTRAL (see Figure 4b). Window length, hop size, and the fraction threshold can be tuned to trade off sensitivity (catching brief shouts) versus specificity (avoiding false alarms on quiet, neutral speech). The 0.5 frame-level probability threshold and the "$\geq$20% agitated windows" clip criterion were chosen to favor sensitivity based on pilot inspection, not tuned on a held-out set.



(**a**)



(**b**)

Figure 4. Figures (**a**,**b**) depict two recordings that are analyzed to assess 'agitation' vs. 'neutral' emotional state. The top subfigure is the oscillogram and the bottom subfigure an agitation metric. (**a**) ESR predicts an emotional state for each speech frame because agitation sometimes manifests itself in bursts. (**b**) A calm conversation.

2.9. Privacy, Data Handling, and Ethics

The ESP32/S3 nodes are triggering autonomously and produce short snippets (fixed 15 s windows). They are designed not to include an SD shield, and their memory suffices for a single recording, therefore audio is buffered only for encoding and transmission and is not written to persistent storage. Continuous raw audio is not streamed.

The server either in the home or facility scenarios can be held on premises so nothing is transmitted outside, including audio visual content. On the server, raw MP3 is decoded to a spectrogram and discarded after feature extraction. We retain only timestamped, de-identified event metadata (AST labels, RMS, diarization tags, agitation flags) and, for the small manually transcribed subset, text without names or identifiers.

Data transport is encrypted, and all MQTT traffic is protected by TLS. Moreover, the experiments in this paper used an on-premises broker and server on the same local network. Participants are referred to only by anonymous IDs in all stored data. Access to the server and logs is restricted to the research team. Caregivers can see audiovisual data but can acquire only reports. We clarify that event logs used in this study are retained for 12 months for research analysis and then will be deleted or irreversibly anonymized. There will be no permanent archive of identifiable audio that can be uploaded to third parties.

3. Results

In this section we gather practical results and an evaluation of services.

3.1. Detecting Cough and Other Non-Verbal Audio Events

Cough events are counted according to the AST classification label (see Figure 5). In this dashboard we illustrate that, for 15 s snippets, the preprocessing module detects the presence of cough, enumerates cough bursts, and estimates cough intensity via the root-mean-square (RMS) amplitude of the waveform. Even in the context of a chronic cough, longitudinal variation in frequency and acoustic character is informative [55]: sudden increases may reflect intercurrent irritation or infection, and changes in sound quality (e.g., wetter/drier timbre or superimposed wheeze) may indicate shifts in respiratory status, and temporal alignment with dialog quantifies communication impact (interruptions and post-cough cessation of speech) as well as correlations with time of day or specific activities. The AST recognizer has 527 audio sources, and the frontend in Figure 5 allows filters to be applied to audio classes that queue and group the audio events of interest.

To operationalize these signals, we compute personalized baselines (e.g., coughs per hour) and multi-day deviation metrics. In Figure 6-top, we see the hours that the cough takes place over a configurable time span. Figure 6-bottom allows us to see the rate of the cough events per hour. The same way, mobility proxies are also captured by choosing the correct filters in the stream of captured-and-tagged audio events. In the AudioSet ontology, these sounds belong to the classes ‘bouncing’, ‘wood’, and ‘tap’. Cane-tap cadence or the impact on the floor by other mobility aids variability provides a lightweight indicator of stability. Aggregating snippet-level outputs yields daily and weekly indicators (e.g., routine adherence) and supports conservative alert policies that prioritize precision for non-urgent notifications while enabling immediate escalation for high-risk events. The same framework detects and quantifies other clinically relevant sounds, including snoring and sleep-related breathing.

2025/08/05 → 2025/08/14

Eve...	Timestamp	Device	A	Prob Cla...	Audio Clas...	Prob Class 2	Audio Cl...	Prob Class 3	RMS (dB)
171389	2025-08-14 15:08:32	esp32s3-3	Cough	0.65	Silence	0.19	Throat clear...	0.06	39.18
171388	2025-08-14 14:22:12	esp32s3-3	Cough	0.78	Throat clearing	0.09	Sneeze	0.08	37.84
171378	2025-08-14 10:34:52	esp32s3-3	Cough	0.6	Throat clearing	0.36	Sneeze	0.01	44.3
171371	2025-08-13 01:26:47	esp32s3-3	Cough	0.65	Throat clearing	0.31	Sneeze	0.01	35.12
171370	2025-08-13 01:25:24	esp32s3-3	Cough	0.57	Throat clearing	0.35	Sneeze	0.01	39.08
171369	2025-08-13 01:25:07	esp32s3-3	Cough	0.54	Throat clearing	0.44	Sneeze	0	40.67
171368	2025-08-13 00:18:09	esp32s3-3	Cough	0.69	Throat clearing	0.29	Sneeze	0	40.59
171270	2025-08-12 00:15:21	esp32s3-3	Cough	0.64	Throat clearing	0.34	Sneeze	0.01	38.49
171269	2025-08-12 00:14:41	esp32s3-3	Cough	0.77	Throat clearing	0.19	Sneeze	0.01	36.65
171265	2025-08-11 23:12:25	esp32s3-3	Cough	0.72	Throat clearing	0.23	Sneeze	0.02	43.35
171263	2025-08-11 23:04:14	esp32s3-3	Cough	0.68	Throat clearing	0.28	Sneeze	0.01	48.16
171262	2025-08-11 23:03:55	esp32s3-3	Cough	0.63	Throat clearing	0.35	Speech	0	47.42
171259	2025-08-11 23:00:41	esp32s3-3	Cough	0.67	Throat clearing	0.3	Speech	0	42.28
171254	2025-08-11 22:47:21	esp32s3-3	Cough	0.47	Throat clearing	0.35	Sneeze	0.05	46.19
171179	2025-08-11 20:31:00	esp32s3-3	Cough	0.57	Throat clearing	0.36	Sneeze	0.01	40.26
171172	2025-08-11 19:26:32	esp32s3-3	Cough	0.27	Speech	0.25	Throat clear...	0.09	58.65
171156	2025-08-11 00:00:31	esp32s3-3	Cough	0.71	Throat clearing	0.2	Sneeze	0.04	36.9
171083	2025-08-08 22:25:55	esp32s3-3	Cough	0.68	Throat clearing	0.25	Sneeze	0.03	42.28
171064	2025-08-08 22:04:09	esp32s3-3	Cough	0.85	Throat clearing	0.09	Sneeze	0.03	46.85
171059	2025-08-08 22:01:36	esp32s3-3	Cough	0.54	Sneeze	0.13	Throat clear...	0.11	44.03
170709	2025-08-08 13:26:01	esp32s3-3	Cough	0.03	Throat clearing	0.02	Sneeze	0.01	38.69
170697	2025-08-08 13:18:52	esp32s3-3	Cough	0.53	Throat clearing	0.27	Sneeze	0.13	34.96
170694	2025-08-08 13:12:03	esp32s3-3	Cough	0.36	Throat clearing	0.28	Sneeze	0.25	38.99
170693	2025-08-08 13:11:43	esp32s3-3	Cough	0.5	Throat clearing	0.3	Sneeze	0.13	37.62
170549	2025-08-08 12:21:18	esp32s3-3	Cough	0.07	Throat clearing	0.03	Sneeze	0.01	39.37
166843	2025-08-06 13:19:04	esp32s3-3	Cough	0.59	Throat clearing	0.3	Sneeze	0.06	34.96
166798	2025-08-06 13:04:35	esp32s3-3	Cough	0.53	Throat clearing	0.28	Sneeze	0.09	33.8

Figure 5. Audio events are classified using the AST, and they are gathered in a dashboard that has the role of a situation assessment board. The dashboard depicts from left to right columns: the ID of the audio event, its timestamp, the device node that originated from, the three higher-ranked class labels of the event with their corresponding probabilities, and the RMS (dB) of the event. Once events are cataloged, one can ask the system to report the audio incidents of 'cough', 'sneeze', and 'throat cleaning' of an elderly couple in an in-house setting and perform various statistics and time-series analysis spanning days to weeks.

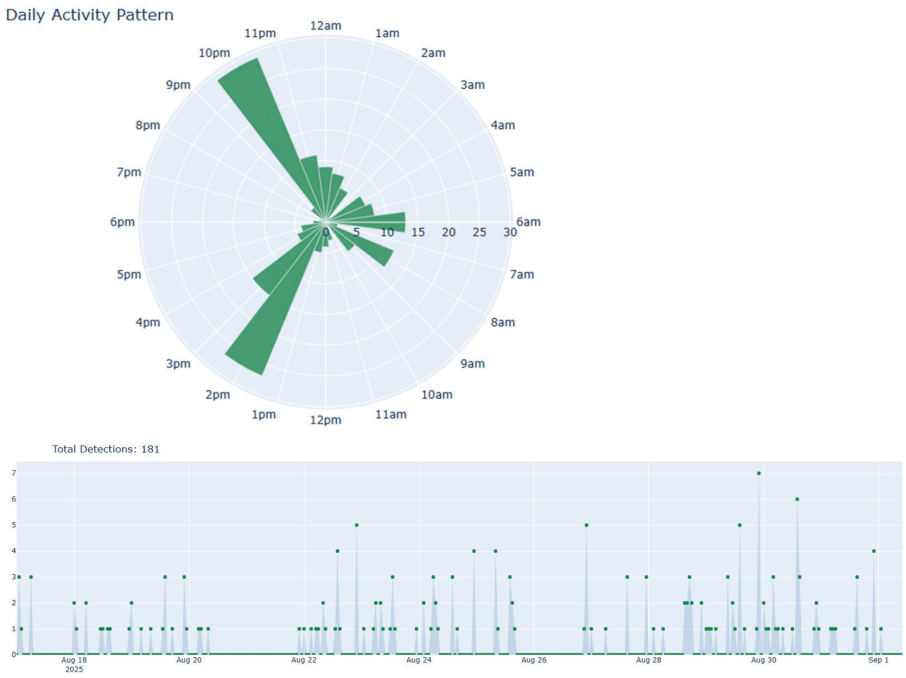


Figure 6. (**Top**) Pie chart of hourly activity of coughing from 17 August 2025–1 September 2025 and, (**Bottom**) time series visualization that assesses rate and trend in cough incidents and relates it to pre- and after treatment. x-axis denotes time and y-axis number of incidents for the target label. Note that the audio category is configurable, and we can visualize any of the 527 audio categories of the AST ontology.

The Audio Spectrogram Transformer (AST) demonstrated exceptional reliability in differentiating cough events from domestic background noise. To evaluate the model's performance in a real-world setting, we implemented a reliable sampling protocol based on continuous "in-the-wild" recordings from the deployed ESP32 nodes.

Sampling Methodology: Ground truth was established through manual annotation of the audio snippets. We curated a balanced validation set comprising 100 confirmed positive instances of respiratory events (merging 'cough', 'sneeze', and 'throat clearing' due to their similar acoustic signatures and clinical relevance) and 400 confirmed negative instances. The negative set included a diverse array of "hard negatives" typical of a home environment, such as speech, television audio, door slams, and silence, to rigorously test the model's false–positive rejection rate.

To accommodate the polyphonic nature of real-world home environments, where target sounds often overlap with background noise, we adopted a "Top-3" classification criterion. An audio segment was classified as a True Positive (TP) if any of the target labels—'cough', 'sneeze', or 'throat clearing'—appeared within the AST's top three predicted classes. Conversely, a false positive (FP) was recorded if any of these labels appeared within the AST's top three predicted classes for an audio segment containing only hard negative events. This threshold was selected to ensure robust detection even in multi-source scenarios where the target event might be identified as a secondary prediction rather than the primary output.

Under these criteria, the AST classifier achieved perfect separation between classes across the 500 sample validation set, effectively filtering out all non-respiratory acoustic interference. The system demonstrated perfect accuracy with the following performance metrics: accuracy, 1.0; precision, 1.0; recall (sensitivity), 1.0; and F1-score, 1.0.

These results confirm the AST's capability to serve as a highly reliable pre-filtering trigger, ensuring that the computationally intensive downstream LLM analysis is only initiated by verified health-related acoustic events.

The VAD is used to pinpoint the event in the recording so that a precise measurement of the RMS intensity and number of incidents is logged (see Figure 7). Note that the audio category is configurable, and we can visualize any of the 527 audio categories of the AST ontology.

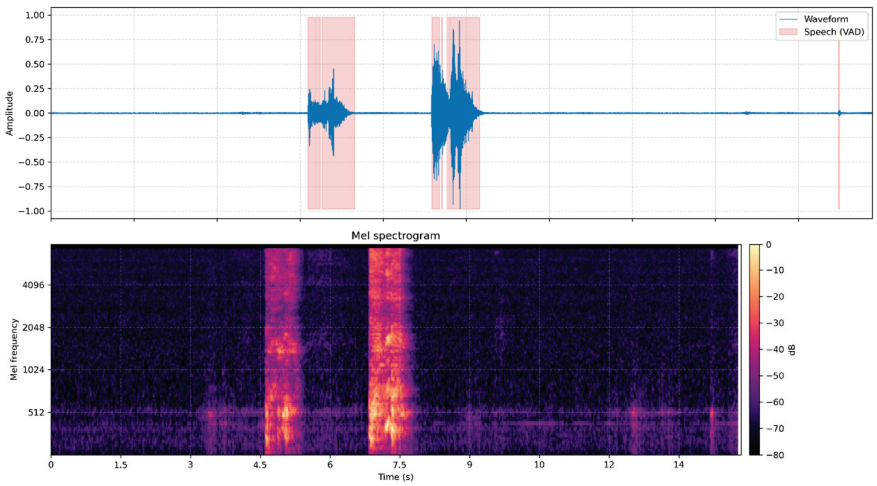


Figure 7. Voice activity detection on a cough recording. (**Top**) The vocal segments are identified and shaded, and the RMS value of the event and the number of events can be measured precisely. (**Bottom**) Spectrogram of a typical cough event.

Once events are recognized and logged, visualizations can support caregiver interpretation. Figure 8 illustrates one such summary: the couple briefly woke around 05:00 (speech detected) and then slept again, waking for the day after ~09:00. Speech dominates

the day's activity (≈68% of 177 detections). Notably, there are extended conversations after the 22:30 bedtime.

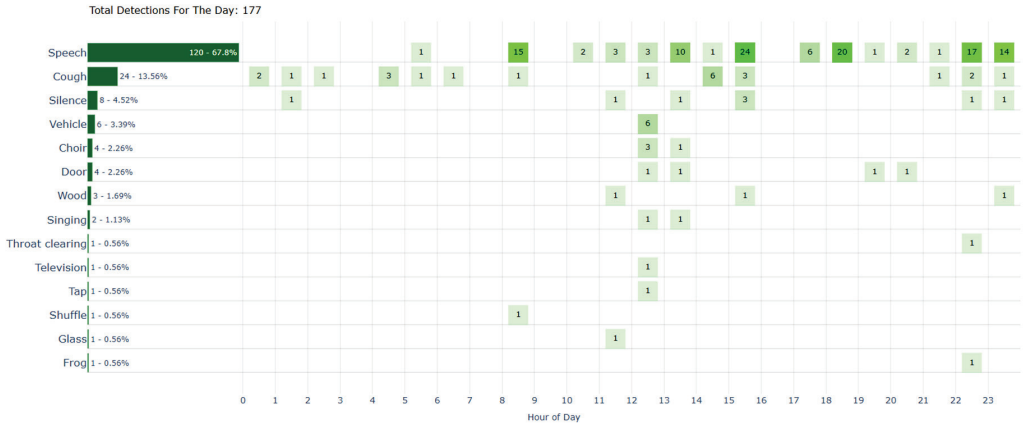


Figure 8. Automated audio event classification and hourly distribution of 177 audio events of an elderly couple with dementia in home. The *y*-axis shows the detected classes and green boxes their relative rates, while the *x*-axis localizes their hourly occurrence in a single day.

In Figure 8, aside from intermittent coughs during the night, sleep appears relatively uninterrupted. Cough events constitute ~14% of detections and occur intermittently across the day—few in the early morning, a larger cluster in the early–midafternoon, and several in the late evening. Cane-related impacts ("bounce," "tap," and "wood") are rare and isolated, suggesting minimal cane striking. No sounds indicative of a fall, pain, or distress were detected on this day.

3.2. Diarization

Diarization has been notably successful, even in a few-shot learning scheme. We extracted 20 recordings, 15 s each, for every person in the home environment to enroll speakers in the diarization application. The people are A: Relative 1, B: Male patient, C: Female patient, D: Caregiver 1, E: Caregiver 2, F: Relative 2. For finetuning, we included recordings including conversations such as B + C, A + B + C, etc. (20 recordings per case). The test set of 100 recordings was taken within one month from the training set. The results are very promising and have been achieved with few recordings per speaker, indicating that more thorough training can achieve near-perfect results.

In Tables 1–3, we present various accuracy metrics for this multilabel, multiclass problem. In multilabel evaluation, subset accuracy is the proportion of instances for which the predicted label set exactly matches the true set (all labels correct, with no missing or extra labels). Hamming loss measures the average fraction of misclassified label decisions over all instance–label pairs and penalizes both missing positives, and spurious positives and is lower when performance is better. The Jaccard index (intersection-over-union) compares predicted and true. Jaccard (micro) first pools true positives, false positives, and false negatives over all labels and instances, then applies the Jaccard formula. Jaccard (macro) computes Jaccard per label and averages labels equally, regardless of frequency. The F1 score is the harmonic mean of precision and recall: F1 (samples) computes per-instance F1 and averages across instances. F1 (micro) pool counts globally before computing precision, recall, and F1 (favoring frequent labels). F1 (macro) averages per-label F1 scores (treating each label equally and highlighting performance on rare labels). In Table 1, we gather the results with emphasis on the multilabel case.

Table 1. Multilabel, multiclass metrics.

Metric	Value
Subset accuracy	0.66
Hamming loss	0.07
Jaccard (samples)	0.80
Jaccard (micro)	0.73
Jaccard (macro)	0.80
F1 (samples)	0.86
F1 (micro)	0.84
F1 (macro)	0.88

Table 2. Single class metrics.

Speaker	Precision	Recall	F1-Score	Support
A	1.00	0.92	0.96	12
B	0.98	0.77	0.86	57
C	0.97	0.60	0.74	53
D	1.00	0.89	0.94	9
E	0.78	0.95	0.86	19
F	0.89	1.00	0.94	8

In classification tasks, precision quantifies the proportion of predicted positives that are truly positive, i.e., the ability of the model to avoid false positives. Recall (also called sensitivity) measures the proportion of true positives that are correctly identified among all actual positives, reflecting the ability to avoid false negatives. The F1-score is the harmonic mean of precision and recall, balancing the trade-off between them and providing a single metric that is especially informative when class distributions are imbalanced. In Table 2, we gather the results on a per-speaker basis.

Table 3. Micro and macro-averaging metrics.

Average Type	Precision	Recall	F1-Score	Support
Micro avg	0.94	0.77	0.84	158
Macro avg	0.94	0.85	0.88	158
Weighted avg	0.95	0.77	0.84	158
Samples avg	0.95	0.82	0.86	158

In multi-class or multi-label evaluation, microaverage computes a metric (e.g., precision, recall, F1, Jaccard) after aggregating true positives, false positives, and false negatives across all classes and instances, emphasizing performance on frequent classes. Macro average first computes the metric separately for each class and then takes the unweighted mean, giving equal weight to rare and common classes. Weighted average is like macro but weighs each class's metric by its support (number of true instances), mitigating the influence of very rare or very common classes on the overall score. Samples average is specific to multi-label settings: it computes the chosen metric per instance by comparing its predicted label set to its true set, then averages these per-instance values over all samples, reflecting how well complete label sets are recovered for each example. In Table 4, we gather the results on micro and macro averaging.

Table 4. Metrics of the emotion recognition task. There are two classes: 'agitated' and 'neutral'.

Class	Precision	Recall	F1
Neutral	0.91	0.91	0.91
Agitated	0.91	0.91	0.91
Accuracy			0.91
Macro avg	0.91	0.91	0.91
Weighted avg	0.91	0.91	0.91

3.3. Emotion Recognition Results

Caregivers and the first author have pulled together a corpus of 22 recordings in agitated and 22 in neutral state. They are based on observed behaviors cross-checked by the cameras and voice characteristics. The 22 Agitated and 22 Neutral clips were selected by querying the longitudinal data for episodes and confirming the label by listening. These class labels are not on formal scales such as Cohen–Mansfield Agitation Inventory (CMAI) or Neuropsychiatric Inventory (NPI).

The agitated state is not common for this couple while in neutral there are thousands of recordings. The algorithm has been tuned so that it picks up agitation even in small phrases (abrupt and short-time raise of voice). In Table 4, we gather the metrics and in Figure 9 we show the confusion matrix of this task. The metrics are the same as in the diarization case. The reported 91% accuracy should be interpreted as agreement with caregiver-level labels in this specific household, not as a validated clinical biomarker. In Supplementary Materials, we offer a sample of agitated/neutral recordings for the reader.

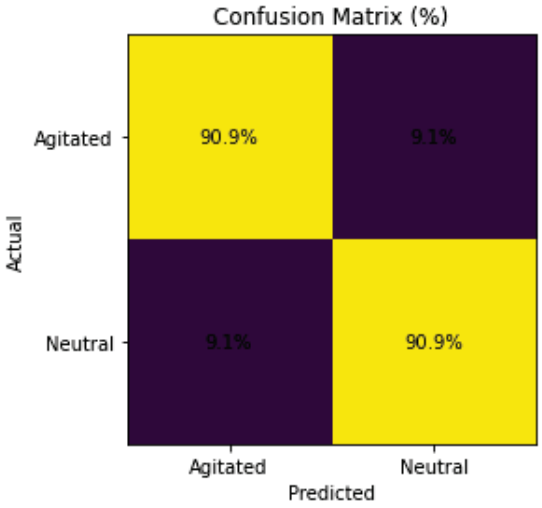


Figure 9. Confusion matrix of emotion recognition between the 'agitated' and 'neutral' class for the emotion recognition module.

3.4. Safety and Emergency Detection

Pharmacologic sleep aids do not reliably ensure uninterrupted nocturnal rest in dementia. Patients may still engage in nighttime wandering or purposeless activity, such as searching for food, watching television, or repeated trips to the toilet. These unattended nocturnal trips are dangerous for their safety as they may result in a fall. This has happened repeatedly for this couple, resulting in bone fractures, agony incidents, and hospitalization. Elders falling at this age are much more likely to experience bone fractures due to osteoporosis. Caregiving during nights is exceptionally problematic because it raises abruptly the cost of this service, and reliable service is hard to find. Our audio-based monitoring device provided unobtrusive nocturnal activity surveillance to detect out-of-bed movement and wandering. The benefits of visualizing activity proxies during night are presented in detail

in Section 4. The device can record events that indicate pain and discomfort, are sometimes identifiable through groans, moans, or sharp exclamations and may overlap acoustically with anger or anxiety. Finally, prolonged silence detection during active hours prompts a check-in. The LLM can integrate the information before it reaches a decision on probable falls from characteristic impact transients (thuds/crashes) followed by silence or distress vocalizations, and recognize shouting or calls for help to trigger immediate alerts to relatives and caregivers. The program at the server can subsequently send push notifications via e-mail and short messages to relatives and caregivers for a probable emergency.

3.5. Caregiver Support and Reporting

We extract the audio events classified by the AST audio recognizer into a JSON file format. The structure of each event is as follows: {"eventid": "evt171390", "timestamp": "2025-08-14 15:43:20", "audio_events": [{"class": "Cough", "probability": 0.37}, {"class": "Silence", "probability": 0.21}, {"class": "Throat clearing", "probability": 0.15}], "rms": 38.69}. The Python 3.11 program time-stamps the event and the AST registers the first three higher in rank classes and their corresponding probabilities and the RMS intensity of the audio event. The LLM was given the JSON file for the current week, the context of the application, the context of the subjects and was asked to prepare a report summary of the previous day for the caregiver for the following day. The structure of the report would be a short log of their activity, emotional state, and detected issues as well as flagging important audio snippets for medical review. Full reports are included in the Appendix A.6. In Table 5 we give a short extract.

Table 5. The LLM prepares a report summarizing key points of health-relevant notes for supporting the caregiver based on a daily file of recognized and timestamped audio events, the context of the application, and a short history of the couple. At this stage of analysis, speech and conversations are not transcribed.

Health-relevant signals (audio only)

- Cough burden (afternoon spike). Six high-confidence cough events occurred 13:45–15:48, aligning with the known chronic cough. Consider hydration, inhaler timing, or irritant exposure review.
- Breath/vocal strain (AM). One gasp/breathing snippet at 08:47 without follow-up distress markers; continue to watch for clustering.
- Mobility/impact sounds. Short bouncing/wood-tap series around 16:55–16:57 may reflect cane contacts or small object taps; no escalations or falls detected.

Engagement & orientation

- Conversation density. Peak conversational activity occurred 13:35–13:48 and 17:47–17:51 near the sensor, suggesting caregiver interaction or couple dialogue proximal to bedroom/bath corridor.
- Night-time/bedtime not confirmed: last record 18:14; no data later in the evening to verify final bathroom/bedroom transition.

The deployment of LLMs in interactions with vulnerable populations necessitates robust mitigation strategies against hallucinations and departures from predefined safety guardrails. In the following section, we develop a technique where LLM interactions are augmented with domain-certified knowledge extracts (e.g., from validated professional texts or medical guidelines). This process effectively grounds the LLM's responses within established safety and care protocols.

3.6. Retrieval-Augmented Generation

Large language models (LLMs) generate continuations from task context and prompts. In extended dialog interactions, initial constraints (e.g., 'don't suggest medicine or outdoor activities') may drift in time and lead to suboptimal or impermissible suggestions. To reduce this risk, we employ retrieval-augmented generation (RAG) to promote factuality, personalization, and auditability. RAG pairs a generator with a retrieval step that, at query time, selects relevant documents or records, and the model is instructed to ground its output in these sources. This design reduces unsupported inferences, incorporates patient and home context (e.g., logs, protocols and prior episodes), and enables citation-backed outputs for care summaries, alerts, and decision support. In this proof of concept, we indexed a source document on dementia care [56]. In operational deployments the corpus can be augmented to include entire books, guides, and manuals. The system ingests daily Audio Spectrogram Transformer event tags in JSON form, caregiver guidance PDF files, and a small household history, and then executes an indexing–retrieval–generation pipeline to produce shift-level reports. During indexing, text is segmented, embedded, and stored in a vector database for semantic search and retrieval supplies the most relevant passages and generation constrains outputs to the retrieved evidence. The code inspects the JSON file for what audio events happened in a day (cough, thuds, speech, and audio related to cane, etc.). It then builds a dynamic retrieval query from the books on treating patients with dementia concerning these events (on top of your base query). Finally, it builds the final prompt that is directed to the LLMs. The LLM is instructed to create a caregiver report with the following sections: 1. summary of the day; 2. behavioral interpretation; 3. caregiver action items (specific; practical); 4. environment/routine adjustments; 5. safety flags; 6. what to monitor tomorrow; and 7. evidence used (cite pages from corpora). In Appendix A, the code link describes an implementation using ChromaDB. Table 6 presents a report with page-level references, and the reader is encouraged to contrasted it with Table 5. Since the JSON is file is different every day the daily report to caregivers remains up to date. Hereinafter, we give a more analytic view of what happens during RAG which is a query targeting communication strategies, safety, and routine management is embedded and used to select the most relevant passages from the indexed corpus. During generation, the system integrates three inputs—retrieved guidance text, a structured summary of daily audio event annotations with timestamps and confidence scores, and a fixed household profile describing history and constraints of the care recipients—to produce a shift-level report for caregivers. The report synthesizes observed behaviors with guideline-concordant responses, highlights potential risks, proposes environment or routine adjustments, and specifies near-term monitoring items. The design emphasizes transparent evidence use by passing the retrieved excerpts and their page references to the language model, enabling traceable outputs without modifying model parameters.

Table 6. A RAG based system prepares a report after vectorizing books, manuals and directives for best treatment of patients with dementia. Note that this report changes daily based on the JSON file that includes the audio recognition events of the previous day.

Daily updated caregiver report (RAG version)
(1) SUMMARY OF THE DAY:
The couple had a relatively active day with 82 speech events, 37 coughs, and 17 sneezes. Distress-like events were detected around 15 h, 05 h, and 04 h.
(2) BEHAVIORAL INTERPRETATION:
The high number of speech events suggests they are communicating, which is positive. However, the distress-like events, especially during the early hours, could indicate sleep disturbances, a common issue in dementia patients. The frequent coughing and sneezing may suggest a health issue, possibly a cold or respiratory infection.
(3) CAREGIVER ACTION ITEMS:

- Monitor their health closely, especially for signs of a cold or respiratory infection.
- Ensure they are keeping hydrated, particularly if they are unwell.
- Encourage them to communicate and engage in conversations.
- Try to maintain a calm and quiet environment, especially during the night to promote better sleep.

(4) ENVIRONMENT/ROUTINE ADJUSTMENTS:

- Introduce relaxing activities in the late afternoon or early evening to help reduce sundowning.
- Ensure their bedroom is quiet, dark, and cool, with a clock showing whether it's day or night next to the bed.
- Limit their screen time towards the end of the day and ensure they get plenty of daylight.

(5) SAFETY FLAGS:

- The distress-like events during the early hours could indicate they are getting up during the night, which could increase the risk of falls.
- The frequent coughing and sneezing could indicate a health issue that needs medical attention.

(6) WHAT TO MONITOR TOMORROW:

- Monitor their sleep patterns and any signs of distress during the night.
- Keep an eye on their health, particularly for any signs of a cold or respiratory infection.
- Observe their communication and social engagement.

(7) EVIDENCE USED:
Guidance from "Caring for a person with dementia: A practical guide" (pages: 43, 63, 75, 79, 139, 140) was used to interpret the data and suggest practical actions.

3.7. Delusion Detection

Detecting delusions and/or hallucinations or even suicidal ideation [57] is valuable information for caregiving, as they can track the development of dementia and detect a possibly harmful action. A delusion is a fixed false belief held despite clear evidence to the contrary. *Example:* the person insists "this isn't my home—I need to go to my other house," starts packing, and rejects reassurance. A hallucination is a sensory perception without an external stimulus (seeing, hearing, or feeling something that is not there). Example: the person says "the kids are sitting on the sofa" when no one is present. What we suggest in this work is that these events, if manifested verbally, can be recognized by the LLM and, therefore, counted. A more elaborate report than that in Table 6, considering these issues, can be asked if we feed the LLM with the transcribed speech of the subjects and features of their emotional state. The aim is to identify events of distress, pain, increasing confusion, agitation, and delusions. In our case, the female has developed two recurrent delusional patterns: (a) she needs to abandon the current home and move back to the first house of her youth. Although she has been living in her last home for over 30 years, she does not recognize it as her own house. (b) A form of delusion that may develop into a hallucination is the one where another 'lady' is inside the house or addressing her husband as her father sporadically. Besides the caregivers sporadic visits by relatives, there are no other people living in the house. The validity of such events has been verified by the caregivers themselves and the IP cameras installed in the rooms. In the Appendix A.6, we have such a report, an extract of which we give in Table 7 for clarity. Note that the system currently does *not* perform automatic delusion detection in real time due to restrictions on ASR reliability. However, if patients do not have dysarthria and a small, but concise history of the patient is given, then it is feasible.

We set the filters to sort speech incidents on the dashboard described in Figure 5, and we listened to discussions, sampling 50 dialogs to include 10 episodes that caregivers and the author had flagged as containing delusional content and 40 routine conversations. Manual Greek transcripts were produced and used as ground truth for the LLM experiment. The 10 "delusional" cases were selected from periods where caregivers or family members documented clear delusional content (place misidentification, phantom outdoors activities, etc.), and the 40 "non-delusional" dialogs were randomly sampled from routine conversations without such flags.

Regarding the annotation protocol, for each dialog, manual Greek transcripts were produced with diarization tags ([M]/[F]) and agitation labels ([A]/[N]). The author then assigned a binary label "Delusional" vs. "Non-delusional" based on the definition of delusion (fixed false belief held despite contrary evidence). We describe the decision rules (e.g., confusion or memory gaps without a fixed false belief are not labeled as delusion). The system discovers delusional schemes by cross-referencing the transcribed narrative against the short family's background to locate factual contradictions. The corpus is small, household-specific, and enriched for clear-cut cases. The detection rate should be interpreted as an existence proof of reasoning ability under idealized conditions as regards ASR, not as an estimate of sensitivity/specificity.

Table 7. Prompting the LLM with (a) small history of the couple (Appendix A.6), (b) the context of the application (prompt), (c) transcribed speech, diarization (male or female), emotion tagging (neutral or agitated). Diarization and emotion tags are in brackets.

PROMPT From now on, act as my expert assistant with access to all your knowledge in the domain of elderly care, elderly psychology, medicine, dementia and best nursing practices of patients with Alzheimer's. Always provide: A clear, direct answer to my request. A step-by-step explanation of how you reached a conclusion. A practical summary or action plan I can apply immediately. Never give vague answers and act like a professional. Push your reasoning to 100% of your capacity. <Small history of the couple> TRANSCRIBED DIALOGUES WITH AUTOMATED DIARIZATION AND EMOTIONAL TAGGING 1. [F][N] Maki did she leave? [M][N] Who? [F][N] Gogo, did she leave? [M][A] I don't know leave me alone! 2. [F][N] I am going to pack our things? [M][N] Why? [F][N] To go home. [M][N] What are you talking about? [F][N] The owners may come back 3. [M][N] Did we eat today? [F][N] I don't remember RESULTS Dialogue 2 shows a delusional theme (place misidentification/"this isn't our home") with exit-seeking intent. The lines "I am going to pack our things. . . To go home. . . The owners may come back" indicate a fixed false belief that they are in someone else's property and must leave. This contradicts the history (they have lived in this home for ~30 years) and fits the female's known occasional delusions. Conclusion: Delusion detected (high likelihood). Dialogue 1 suggests anxiety/checking behavior, not a delusion. Asking if the caregiver "Gogo" left is a reality-check; no false belief is asserted. The male's irritable reply does not introduce delusional content. Conclusion: No delusion detected (low likelihood). Dialogue 3 reflects memory impairment, not a delusion. "Did we eat today?"/"I don't remember" is short-term memory loss. No false belief is stated. Conclusion: No delusion detected (very low likelihood).

4. Discussion

4.1. Practical Achievements So Far

Our proposal is an affordable plug-and-play device whose data is classified at multiple levels, with results jointly interpreted by an LLM. Caregiving is typically an arrangement under budget constraints, and the configuration studied here—without loudspeakers or actuators—targets practical gains from passive audio monitoring. Hereinafter, we gather practical gains after using the system for almost five months continuously: (a) Our approach for overnight surveillance lowers staffing needs for night shifts, which are costly and often inefficient. The system can report the accident (usually a fall) in real time to guardians, reducing the response time and relieving the pain and agony of the elderly that are typically unable to rise. Continuous fall monitoring at night also reduces family anxiety, given the difficulty of elders calling for help after a fall. The system successfully detected a real fall at night and recognized it as such using only audio (see Supplementary Materials). (b) The recognition of audio and its hourly distribution revealed that after caregivers left, the subjects frequently rose, walked, and conversed, leading to daytime exhaustion that had been misattributed. Inspection of the system's "speech" and "conversation" labels between 00:00 and 07:00 prompted a medication adjustment by a doctor that resolved this pattern. (c) The suggested pipeline quantifies delusional episodes over time, enabling trend

tracking, and measures cough burden before and after treatment to evaluate treatment's efficacy. (d) Systematic analysis of conversational content provided actionable insight into psychological status and stressors, enabling individualized care plans that target minor yet recurring practical stressors (e.g., cold bedsheets, arrangement of furniture, room temperature, and lighting conditions) that degrade wellbeing of dementia patients and are commonly not reported the following day due to memory impairment.

We placed particular value on the system's uninterrupted (24/7) reliability, which enables continuous monitoring and provides added support during periods of staffing constraints or caregiver absence, thereby easing the burden on adult children with legal guardianship responsibilities. Our approach is designed to offer a remote yet direct connection between guardians and patients without requiring the patients to use a phone or computer application to send or receive messages. Guardians can quickly review the situation (see Figures 5–7), listen to specific events of interest by clicking on them on the dashboard, and—depending on the individual's cognitive state and acceptance—will soon be able to respond via the device's Wi-Fi loudspeaker to reassure, remind, or gently redirect (duplex communication). This design acknowledges that many older adults do not respond well to handling gadgets: even reaching for a mobile phone, pressing the correct button, or navigating a laptop or tablet application can be difficult or unreliable for them depending on age, mobility, and mental condition.

4.2. Limitations and Generalizability

In interpreting our findings, it is important to recognize the limits of the cohort and analysis. The study focuses on a single couple with stage 2 dementia, both Greek-speaking, living in a specific home environment, and is therefore not representative of the broader dementia population in terms of language, cultural background, disease severity, or living arrangements. All quantitative results are reported at the individual level and are intended as descriptive characterizations of system behavior and within subject patterns rather than as estimates of general performance. That is, the emphasis is on the engineering part. Consequently, we do not make claims about sensitivity, specificity, or other diagnostic metrics at population scale, and we view the present work as a feasibility oriented case study that motivates, rather than replaces, larger and more diverse evaluations.

5. Future Prospects

5.1. Scaling up the Service

Although our primary goal is to deliver affordable, automated in-home services, the architecture scales directly to facility-wide deployments. Field visits to elderly care facilities in the city of Patras, Greece, revealed multi-floor layouts with consecutive rooms housing 2–4 residents in heterogeneous clinical states, often with only a subset of institutions admitting people with dementia or Alzheimer's disease and with wide variation in service level and cost. These characteristics—repeatable room topology, diverse care needs, and operational heterogeneity—make a centralized, multi-room deployment both feasible and valuable, enabling consistent monitoring policies across wards while preserving room-level personalization. The proposed combination of audio signal processing and AST classifiers combined with LLMs stack yields a live "situational awareness board" that could monitor all rooms simultaneously surface deviations (e.g., sudden spikes in cough burden, impact-like thuds, delusional dialogs), enabling oversight across an entire ward. Bedsides evolve into conversational endpoints where the LLM engages residents and staff through brief, goal-directed dialogs and concise screen summaries—probing orientation, interpreting confusion or fixed false beliefs, and continuously "listening" for accidents, pain, distress and alarms—while emitting auditable, structured assessments and

suggested de-escalation or check-in scripts. Cameras equipped with audio and loudspeaker functionalities are viable, and in the Supplementary Materials, we include as an example of what is currently feasible with an analysis of a video depicting real incidents of falls using an audiovisual LLM. However, the audio-only modality alone is more discreet, less invasive, lower cost, and requires no infrastructure changes. In-house deployments and large facilities, processing and inference can run on an on-premises server for security and to adhere to ethical normative.

In the envisioned deployment, LLMs coordinate with assistive robots and screens—both mobile platforms and bedside devices—within a closed-loop control framework. The LLM translates detected needs into verifiable action primitives, monitors execution via on-board sensor feedback, and hands off to human staff when uncertainty or risk thresholds are exceeded. Continuous monitoring and self-adaptation target conservative escalation policies that prioritize precision for non-urgent events and immediate alerts for hazards. Over time, federated learning enables room- and resident-specific personalization while preserving central governance and auditability. This approach is intended to augment—not replace—human care by converting ambient sound into continuously updated, clinically relevant signals and delegating low-risk, routine assistance to restless machines, thereby reserving clinician and caregiver effort for difficult judgment, physical exertion, and empathy that are far from what a machine can yet provide. The next step is a planned multi-site pilot with a larger, more heterogeneous cohort (different dementia stages, living arrangements, and languages) and pre-registered endpoints (e.g., sensitivity to agitation/exit-seeking and caregiver workload).

5.2. Conversation and Cognitive Support Using Actuators

In the present state, we keep a human in the loop as an indispensable part (see Section 4.1). However, we envision that an agentic service will finally autonomously engage in conversations (AI-bots/agents) that stimulate participants' memory, monitor hazardous situations, and mitigate the sense of loneliness. This task is currently technically feasible, especially with the advent of large audiovisual models (e.g., of the class Qwen3-VL) and voice agents in online decoding mode with emotional speech, barge-in capability, and duplex communication. Our aim is to provide this service under the unconditional restriction of financial affordability using autonomous Wi-Fi audio recorders, spare smartphones, and affordable Wi-Fi cameras with audio.

In dementia care, an effective agent must integrate six capabilities within explicit safety and ethics guardrails. First, reasoning must be bounded and evidence-driven: the agent fuses multimodal inputs (e.g., short audio snippets, camera feed, ASR key phrases, affect tags, time of day, and a consented personal profile) to draw low-risk inferences and select pre-approved utterances rather than generate open-ended advice to vulnerable people. Second, acting is constrained and assistive, not clinical: the agent performs digital actions—brief TTS prompts via Wi-Fi loudspeakers, on-screen cues, and caregiver/guardian notifications. This would be useful, especially in the case of the important category of an alone elder that will otherwise not speak a lot. It will operate in supervised autonomy and will generate speech only from curated templates. Third, observing relies on privacy-preserving sensing (on-device preprocessing where feasible, minimal metadata upstream) to maintain situational awareness while protecting dignity. Fourth, planning converts near-term goals (hydration, toileting, calm bedtime, and orientation) into simple, time-appropriate steps with escalation paths to humans when uncertainty or risk is detected. Upon detecting agitation or perseverative questioning—either through upstream classifiers or conversational cues—the agent employs gentle redirection strategies to achieve de-escalation. Fifth, it treats caregivers as partners, exposing transparent logs, ratio-

nales, and easy overrides so human judgment remains central. Finally, self-refining allows personalization via supervised updates based on psychological profiling of the users. The system self-adapts speaking rate, prosody, and linguistic complexity to the user's current cognitive state that is unobtrusively assessed continuously and maintains a consented personal knowledge database to tailor references and improve engagement. This implies that the speech recognizer is already adapted for elderly speech [52–54]. The broad availability of LLMs, ASR, and TTS across most major languages inherently supports the global scalability. The agent can manage telecommunications by autonomously handling incoming calls and initiating contact with guardians or emergency services when necessary. Modern TTS technologies enable the synthesis of expressive, affect-aware speech, enhancing the naturalness of these interactions. While challenges to fully autonomous operation remain, we maintain that these are technically tractable through robust architectural safeguards. Key risks to be mitigated include hallucinatory guidance (highly reduced with RAG), ASR bias against elderly speech patterns (reduced through adaptation), the propagation of false positives from upstream classifiers, and the necessity of enforcing strict topic boundaries (e.g., preventing the LLM from offering medical diagnoses or medication advice).

The system we propose (see Figure 10) differs fundamentally from existing voice assistants such as Alexa/Siri/Cortana, which rely solely on speech-based interaction and reactive responses. In contrast, our system will be proactive, i.e., it will take initiative in starting and maintaining a conversation based on audiovisual feed, continuously integrating audio and visual sensing to monitor older adults for safety-critical events such as falls, choking, or potentially hazardous events such as unattended cooking and handling of objects, etc. This multimodal, context-aware, and memory-preserving framework shifts the paradigm from passive communication to active, personalized care monitoring. In this work we focus only on audio, but in the Supplementary Materials, we offer some direction for vision as well. To safeguard vulnerable users, the design process must be led by a joint medical-technical team that continuously monitors how patients at different stages of disease react to and accept the system.

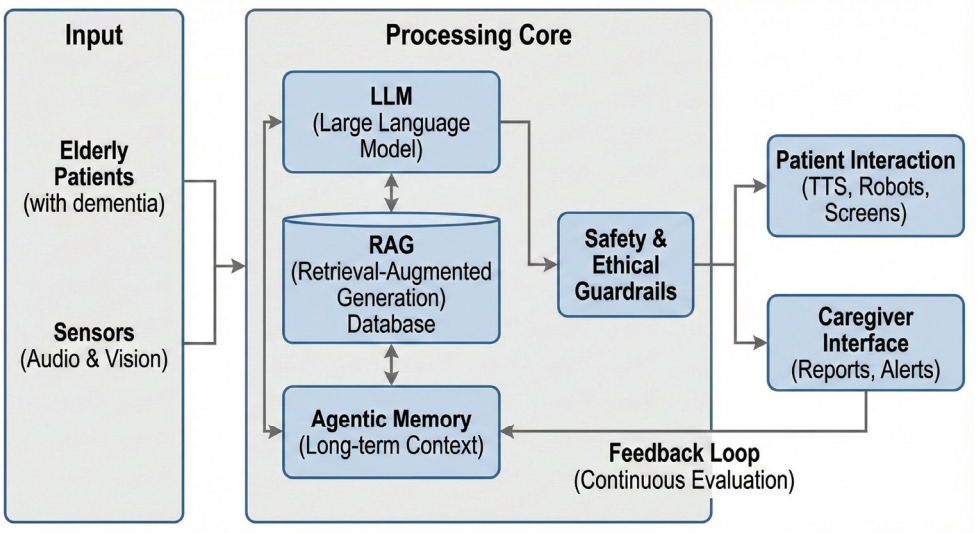


Figure 10. Proposed architecture for future AI-driven dementia care. A Retrieval-Augmented Generation (RAG) enhanced Large Language Model (LLM) engages patients in conversation via a text-to-speech (TTS) module and loudspeakers, continuously re-evaluating the situation in real-time. The system detects emotional speech cues, adapting its voice and responses accordingly. A "long-term memory" database stores patient history, personal data, reports, and audiovisual logs. The system interfaces with external devices (screens, robots, smartphones) and responds to caregiver queries with evidence-backed insights to ensure accountability.

5.3. Adapting Speech Processing Modules

As the older-to-younger demographic imbalance is likely to persist in the Western world—driven by cultural factors and advances in medicine—older adults present a significant opportunity for advanced AI applications. Existing ASR, emotion-recognition, and diarization systems should be adapted to this population. Clinical deployment cannot rely on free-field ASR for safety-critical decisions until such adaptation has been completed and evaluated.

6. Conclusions

Affordable, commodity sensors and AI can supplement—and in underserved cases partially substitute—in-home caregiving for older adults with dementia when dependable services are unaffordable or unavailable. We integrated novel, affordable devices and recent advances in audio processing with LLM-based reasoning and highlighted the limitations of ASR in age-related voice changes, motivating a pipeline that leverages non-speech acoustic events alongside transcripts. Our goals are feasibility, systems integration, and safety/guardrail design rather than diagnostic performance. This manuscript is a proof-of-concept case study, not a definitive clinical study. We prototyped an ESP32-S3 node with an onboard MEMS microphone and note that commodity smartphones and smart TVs are viable, low-cost recorders. Short audio snippets from daily activity are streamed almost in real time to a recognizer with 527 acoustic classes, with speech transcription and emotion inference producing structured cues that are passed to an LLM for situation assessment. We demonstrate end-to-end feasibility in a home with real patients, introduce reusable tooling (open code/data schemas, structured outputs, safety policies), and provide early longitudinal evidence that clinically relevant signals can be tracked. The range of services under the same concept can be significantly expanded with the future incorporation of loudspeakers, robots and screens.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/a18120787/s1, Picture Device.jpg. Dialogues.pdf: Transcribed human dialogues. sound_events_single_day.json: A JSON log of audio events over a single day from the automated system. Videos: male_fall_subs_anonymized.mp4, male_8_11_captioned.mp4, female_fall_subs.mp4, male_fall2_captioned.mp4, female_fall2_short.mp4: Qwen3-VL scans videos for accidents. Video: male_audio_only_fall_subs_anonymized.mp4. Audio only scanning. Emotions folder: Anonymized mp3 audio sample recordings of agitated/neutral state and figures of agitation probability vs. time for each recording.

Funding: This research received no external funding.

Ethical Issues: In Greece, the general research–ethics framework is governed by Law 4521/2018, which requires ethical review for research involving human subjects in biomedical and social sciences. Institutional codes (e.g., the Code of Ethics and Conduct of Research of the University of West Attica, the Code of Conduct of ITE, and other universities) specify that research in the social sciences must protect anonymity and personal data and may proceed with consent if minimal-risk and non-interventional and do not fall under clinical/biomedical intervention. According to the Guide of the National Bioethics Commission, while ethics-committee review is recommended for research on human behavior, it is not mandated in every case. In this study the data collection consisted of audio-recorded conversations with older adults in a non-interventional, minimal-risk format; participants gave informed consent, data were pseudonymized, and no clinical intervention was involved. Therefore, as per the above legislative and institutional context, we determined that a full ethics-committee formal approval was not required while appropriate ethical safeguards (consent, anonymity, data protection) were adopted in accordance with local regulations and the ethical principles that have their origin in the principles of the Declaration of Helsinki.

Institutional Review Board Statement: Not applicable. The researchers analyzed only anonymized event metadata generated on-device; no identifiable audio or video were accessed or transferred.

Informed Consent Statement: Not applicable; no identifiable data or images are included.

Data Availability Statement: Raw recordings remain on participants' devices and are not available. Anonymized event metadata used in this study is not publicly available. A small sample of the dataset of audio events is available at Supplementary Materials.

Conflicts of Interest: The author declares no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

DSP	Digital Signal Processing
RMS	Root, Mean Square Value
VAD	Voice Activity Detection
ASR	Automatic Speech Recognition
ESR	(Automatic) Emotion Speech Recognition
TTS	Text to Speech
AST	Audio Spectrogram Transformer
IoT	Internet of Things
MQTT	Message Queuing Telemetry Transport
RAG	Retrieval-Augmented Generation
VIT	Vision Transformer
JSON	JavaScript Object Notation
MP3	Moving Picture Experts Group Audio Layer III
LLM	Large Language Model
MEMS	Micro Electro-Mechanical Systems

Appendix A

Code used in this work: https://github.com/potamitis123/dementia-care (accessed on 1 November 2025). In Supplementary Materials we analyzed two videos of falls with Qwen3-VL-235B-A22B (https://chat.qwen.ai/, accessed on 1 November 2025) and passed the LLM's output as captions. The system recognized correctly the situation and can be set to notify guardians automatically. This is a concrete example of what audiovisual models can bring to elderly care. The first video has been tagged by our audio-only model and can be found on Supplementary Materials as well. The second video has no audio and in such cases audio only approaches are underpowered. The interested reader needs to know that no injuries happened to the elderly and that there was no way to intervene prior to the fall as the data have been analyzed offline. The videos have been defaced (i.e., blurred the face) when needed to anonymize them using https://github.com/ORB-HD/deface/ (accessed on 1 November 2025). The links of all audio toolboxes used in this work can be found below.

Appendix A.1. General Audio Recognition

Audio Spectrogram Transformer (AST) applies a Vision Transformer to audio by converting waveforms to log-Mel spectrograms, tokenizing them into patches, and feeding them to a transformer encoder for classification. The library provides ASTFeatureExtractor to compute and normalize Mel features (AudioSet mean/std by default) and ASTForAudioClassification/ASTModel for inference or fine-tuning, with configuration parameters like patch size, time/frequency strides, and number of Mel bins exposed via ASTConfig. The

documents highlight practical tips and show ready-to-run examples with the pretrained MIT/ast-finetuned-audioset-10-10-0.4593 checkpoint.

https://huggingface.co/docs/transformers/model_doc/audio-spectrogram-transformer (accessed on 1 November 2025).

Appendix A.2. VAD

TEN is an open-source ecosystem for creating, customizing, and deploying real-time conversational AI agents with multimodal capabilities, including voice, vision, and avatar interactions.

https://github.com/TEN-framework/ten-vad/ (accessed on 1 November 2025).

Appendix A.3. ASR

Whisper is a state-of-the-art model for automatic Speech Recognition (ASR) and speech translation. The Whisper large-v3 model was trained on 1 million hours of weakly labeled audio and 4 million hours of pseudo-labeled audio collected using Whisper large-v2. The model was trained for 2.0 epochs over this mixture dataset. The large-v3 model shows improved performance over a wide variety of languages, showing 10% to 20% reduction in errors compared to Whisper large-v2.

https://huggingface.co/openai/whisper-large-v3 (accessed on 1 November 2025).

Appendix A.4. Diarization

pyannote.audio is an open-source toolkit written in Python for speaker diarization. Based on PyTorch 2.7.1+cu118 machine learning framework, it comes with state-of-the-art pretrained models and pipelines that can be further fine-tuned to our data for even better performance.

https://github.com/pyannote/pyannote-audio (accessed on 1 November 2025).

Appendix A.5. ESR

superb/hubert-large-superb-er is a HuBERT-Large (hubert-large-ll60k) model fine-tuned for the SUPERB Emotion Recognition task. It takes 16 kHz mono audio and predicts one of four utterance-level emotions—angry, happy, sad, or neutral—following the standard SUPERB/IEMOCAP setup that drops minority classes to balance the dataset. The checkpoint is a port of the S3PRL implementation, providing a simple sequence classification head on top of HuBERT's self-supervised speech representations. Usage is straightforward via the Transformer's audio-classification pipeline or the underlying model/feature extractor. The model card reports accuracy on the SUPERB demo split and includes example code. In practice, it is a compact, well-documented baseline for categorical speech emotion recognition that you can adapt or fold into application-specific labels (e.g., "agitated" vs. "neutral").

https://huggingface.co/superb/hubert-large-superb-er (accessed on 1 November 2025).

Appendix A.6. Delusion Detection

GPT-5 set at 'thinking mode'. The prompts, the history, and the dialogs should recreate the results on the delusional patterns detection.

"PROMPT: From now on, function as my expert assistant with access to all your knowledge in the domain of elderly care, elderly psychology, medicine, dementia and best nursing practices of patients with Alzheimer's. Always provide: A clear, direct answer to my request. A step-by-step explanation of how you reached a conclusion. A practical summary or action plan I can apply immediately. Never give vague answers and act like a professional. Push your reasoning to 100% of your capacity.

TASK: We have in-house recordings of speech/conversations of a couple with dementia stage 2. I will provide their discussions, and you need to answer queries based on these discussions and the small history of the couple.

SMALL HISTORY OF THE COUPLE

(a) Female: 87; Male: 90+
(b) Diagnosis: both have dementia (stage 2); only the female has occasional delusions.
(c) Mobility: both have mobility impairments; use canes; no capability to leave the home; they can't access the garden due to a steep staircase; they can't cook or make hot beverages.
(d) Memory: acute short-term memory loss/forgetfulness.
(e) Care context: they live in their own home; two daytime caregivers (Eve and Gogo) cover 8-h shifts; caregivers have been working with the couple for two years; one adult daughter and one other relative visit occasionally; no other people are in the home; Their parents died fifty years ago.
(f) Residency & property: they have lived in the current home for ~30 years; before that, only their youth home (>30 years ago); they do not own any other house.

Query #1: Detect delusional expressions in dialogues. [M]; Male, [F] Female, [A] Agitated emotional state. In absence of [A], emotional state Neutral"

DIALOGUES

Dialogues and results on Q1 are appended to the manuscript as supportive material.

Caregivers' daily summary

"Query #2: Prepare a text-only report summarizing the main events of the previous day by parsing the attached JSON file containing tagged audio events of the previous day <day>. Each event receives the three higher rank labels and their corresponding probabilities and a timestamp. Consider all three classes as possible valid classes. Be brief, prioritize events that can be related to health, nocturnal activity and movement patterns."

ChatGPT-5.1 response: "Here's your concise daily audio report for Wednesday, 10 September 2025 (Europe/Athens)—focused on health, nocturnal activity, and movement patterns.

(1) Direct answer—What mattered yesterday
- Nocturnal cough episode from 04:32–05:34 with multiple high-confidence cough events (peaks up to 0.81). Likely disturbed sleep and airway irritation.
- Evening respiratory activity around 21:26–22:36, including a strong cough at 22:35:48 (0.82).
- Door/movement events at typical shift/transition times: 09:17, 10:25, 12:20, 18:12, 18:47, 20:45, 21:30–21:32. No clear fall pattern.
- Single low-confidence impact ("thump/thud") at 11:25:53—not suggestive of a fall given low probability and surrounding context.
- Overall counts: 133 events total. Health-related: 47 (of which 18 nocturnal). Door/movement: 9. No high-confidence alarms/falls.

(2) Practical action plan (apply today)

Health (coughs & sleep):
- Log this as a significant nocturnal cough episode (04:32–05:34). Ask caregivers to note any triggers (dry air, reflux after late meal, recent cold).
- Interventions today: ensure hydration, warm fluids (prepared by caregiver), avoid late heavy meals, keep head elevated during rest, check bedroom humidity (~40–50%).
- If cough clusters repeat $\geq$3 nights/week or intensify, consider GP review (rule out GERD, post-nasal drip, infection).

Night routine:

- Given repeated late-evening/early-morning events, consider a calmer wind-down from 20:30 (reduced TV volume, warm drink, bathroom check, brief reassurance).

Movement & safety:

- Door activity matches shift changes; no fall signature detected. Keep current fall-prevention setup (clear pathways, canes nearby, night lights). Caregivers to quickly check after any future "thud/door-slam" timestamps".

"Query #3: At night, the couple is unattended by caregivers and there is no one other than the couple in home. There are no pets, no music, no tv set on. Pay attention to transcribed audio patterns in the attached JSON file that may involve a hit, fall, followed by vocal expressions of distress, pain and/or anxiety."

References

1. World Health Organization. Available online: https://www.who.int/news-room/fact-sheets/detail/dementia? (accessed on 6 November 2025).
2. EUROSTAT. Available online: https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Population_structure_and_ageing (accessed on 6 November 2025).
3. United Nations. Available online: https://www.un.org/development/desa/pd/sites/www.un.org.development.desa.pd/files/undesa_pd_2024_wpp_2024_advance_unedited_0.pdf (accessed on 6 November 2025).
4. Alzheimer's Association. Available online: https://www.alz.org/alzheimers-dementia/facts-figures? (accessed on 6 November 2025).
5. König, A.; Satt, A.; Sorin, A.; Hoory, R.; Derreumaux, A.; David, R.; Robert, P.-H. Automatic speech analysis for the assessment of patients with predementia and Alzheimer's disease. *Alzheimer's Dement. Diagn. Assess. Dis. Monit.* **2015**, *1*, 112–124. [CrossRef] [PubMed]
6. Haider, F.; De La Fuente Garcia, S.; Luz, S. An Assessment of Paralinguistic Acoustic Features for Detection of Alzheimer's Dementia in Spontaneous Speech. *IEEE J. Sel. Top. Signal Process.* **2020**, *14*, 272–281. [CrossRef]
7. Qi, X.; Zhou, Q.; Dong, J.; Bao, W. Noninvasive automatic detection of Alzheimer's disease from spontaneous speech: A review. *Front. Aging Neurosci.* **2023**, *15*, 1224723. [CrossRef] [PubMed]
8. Saeedi, S.; Hetjens, S.; Grimm, M.O.W.; Barsties, V.; Latoszek, B. Acoustic Speech Analysis in Alzheimer's Disease: A Systematic Review and Meta-Analysis. *J. Prev. Alzheimer's Dis.* **2024**, *11*, 1789–1797. [CrossRef]
9. Martínez-Nicolás, I.; Llorente, T.E.; Martínez-Sánchez, F.; Meilán, J.J.G. Ten years of research on automatic voice and speech analysis of people with Alzheimer's disease and mild cognitive impairment: A systematic review article. *Front. Psychol.* **2021**, *12*, 620251. [CrossRef]
10. van den Berg, R.L.; de Boer, C.; Zwan, M.D.; Jutten, R.J.; van Liere, M.; van de Glind, M.-C.A.; Dubbelman, M.A.; Schlüter, L.M.; van Harten, A.C.; Teunissen, C.E.; et al. Digital remote assessment of speech acoustics in cognitively unimpaired adults: Feasibility, reliability and associations with amyloid pathology. *Alzheimer's Res. Ther.* **2024**, *16*, 176. [CrossRef]
11. Liu, J.; Fu, F.; Li, L.; Yu, J.; Zhong, D.; Zhu, S.; Zhou, Y.; Liu, B.; Li, J. Efficient Pause Extraction and Encode Strategy for Alzheimer's Disease Detection Using Only Acoustic Features from Spontaneous Speech. *Brain Sci.* **2023**, *13*, 477. [CrossRef]
12. Meilán, J.J.; Martínez-Sánchez, F.; Carro, J.; López, D.E.; Millian-Morell, L.; Arana, J.M. Speech in Alzheimer's disease: Can temporal and acoustic parameters discriminate dementia? *Dement. Geriatr. Cogn. Disord.* **2014**, *37*, 327–334. [CrossRef]
13. Xue, C.; Karjadi, C.; Paschalidis, I.C.; Au, R.; Kolachalama, V.B. Detection of dementia on voice recordings using deep learning: A Framingham Heart Study. *Alzheimer's Res. Ther.* **2021**, *13*, 146. [CrossRef]
14. Ding, K.; He, C.; Chen, Z.; Xu, X.; Li, W.; Hu, B. Using acoustic voice features to detect mild cognitive impairment: Machine learning model development study. *JMIR Aging* **2024**, *7*, e57873. [CrossRef]
15. Vincze, V.; Szatlóczki, G.; Tóth, L.; Gosztolya, G.; Pákáski, M.; Hoffmann, I.; Kálmán, J. Telltale silence: Temporal speech parameters discriminate between prodromal dementia and mild Alzheimer's disease. *Clin. Linguist. Phon.* **2021**, *35*, 727–742. [CrossRef]
16. Sefcik, J.S.; Ersek, M.; Hartnett, S.C.; Cacchione, P.Z. Integrative review: Persistent vocalizations among nursing home residents with dementia. *Int. Psychogeriatr.* **2019**, *31*, 667–683. [CrossRef] [PubMed]
17. Kosters, J.; Janus, S.I.M.; Van Den Bosch, K.A.; Zuidema, S.; Luijendijk, H.J.; Andringa, T.C. Soundscape Optimization in Nursing Homes Through Raising Awareness in Nursing Staff With MoSART. *Front. Psychol.* **2022**, *13*, 871647. [CrossRef] [PubMed]
18. Kaur, P.; Wang, Q.; Shi, W. Fall detection from audios with Audio Transformers. *Smart Health* **2022**, *26*, 100340. [CrossRef]
19. Newaz, N.T.; Hanada, E. The Methods of Fall Detection: A Literature Review. *Sensors* **2023**, *23*, 5212. [CrossRef]

20. Li, Y.; Liu, P.; Fang, Y.; Wu, X.; Xie, Y.; Xu, Z.; Ren, H.; Jing, F. A Decade of Progress in Wearable Sensors for Fall Detection (2015–2024): A Network-Based Visualization Review. *Sensors* **2025**, *25*, 2205. [CrossRef]
21. Rocha, I.C.; Arantes, M.; Moreira, A.; Vilaça, J.L.; Morais, P.; Matos, D.; Carvalho, V. Monitoring Wearable Devices for Elderly People with Dementia: A Review. *Designs* **2024**, *8*, 75. [CrossRef]
22. Saradopoulos, I.; Potamitis, I.; Ntalampiras, S.; Rigakis, I.; Manifavas, C.; Konstantaras, A. Real-Time Acoustic Detection of Critical Incidents in Smart Cities Using Artificial Intelligence and Edge Networks. *Sensors* **2025**, *25*, 2597. [CrossRef]
23. Casu, F.; Lagorio, A.; Ruiu, P.; Trunfio, G.A.; Grosso, E. Integrating Fine-Tuned LLM with Acoustic Features for Enhanced Detection of Alzheimer's Disease. *IEEE J. Biomed. Health Inform.* **2025**, 1–14. [CrossRef]
24. Yusupov, I.; Douglas, H.; Lane, R.; Ferman, T. Vocalizations in dementia: A case report and review of the literature. *Case Rep. Neurol.* **2014**, *6*, 126–133. [CrossRef]
25. Fillit, H.; Aigbogun, M.S.; Gagnon-Sanschagrin, P.; Cloutier, M.; Davidson, M.; Serra, E.; Guérin, A.; Baker, R.A.; Houle, C.R.; Grossberg, G. Impact of agitation in long-term care residents with dementia in the United States. *Int. J. Geriatr. Psychiatry* **2021**, *36*, 1959–1969. [CrossRef] [PubMed]
26. Pedersen, S.K.A.; Andersen, P.N.; Lugo, R.G.; Andreassen, M.; Sütterlin, S. Effects of Music on Agitation in Dementia: A Meta-Analysis. *Front. Psychol.* **2017**, *8*, 742. [CrossRef] [PubMed]
27. Miller, S.; Vermeersch, P.E.; Bohan, K.; Renbarger, K.; Kruep, A.; Sacre, S. Audio presence intervention for decreasing agitation in people with dementia. *Geriatr. Nurs.* **2001**, *22*, 66–70. [CrossRef] [PubMed]
28. Shah, R.; Basapur, S.; Hendrickson, K.; Anderson, J.; Plenge, J.; Troutman, A.; Ranjit, E.; Banker, J. Does an Audio Wearable Lead to Agitation Reduction in Dementia: The Memesto AWARD Proof-of-Principle Clinical Research Study. *Res. Sq.* **2025**, rs.3.rs-6008628. [CrossRef]
29. Alsina-Pagès, R.M.; Navarro, J.; Alías, F.; Hervás, M. homeSound: Real-Time Audio Event Detection Based on High Performance Computing for Behaviour and Surveillance Remote Monitoring. *Sensors* **2017**, *17*, 854. [CrossRef]
30. Rashidi, P.; Cook, D.J. Keeping the Resident in the Loop: Adapting the Smart Home to the User. *IEEE Trans. Syst. Man Cybern. —Part A Syst. Hum.* **2009**, *39*, 949–959. [CrossRef]
31. Adlam, T.; Faulkner, R.; Orpwood, R.; Jones, K.; Macijauskiene, J.; Budraitiene, A. The installation and support of internationally distributed equipment for people with dementia. *IEEE Trans. Inf. Technol. Biomed.* **2004**, *8*, 253–257. [CrossRef]
32. Bouchard, B.; Giroux, S.; Bouzouane, A. A Keyhole Plan Recognition Model for Alzheimer's Patients: First Results. *Appl. Artif. Intell.* **2007**, *21*, 623–658. [CrossRef]
33. Mégret, R.; Dovgalecs, V.; Wannous, H.; Karaman, S.; Benois-Pineau, J.; El Khoury, E.; Pinquier, J.; Joly, P.; André-Obrecht, R.; Gaëstel, R.; et al. The IMMED Project: Wearable Video Monitoring of People with Age Dementia. In Proceedings of the 18th ACM International Conference on Multimedia, Firenze, Italy, 25–29 October 2010; Association for Computing Machinery: New York, NY, USA, 2010; pp. 1299–1302.
34. Adami, A.; Pavel, M.; Hayes, T.; Singer, C. Detection of movement in bed using unobtrusive load cell sensors. *IEEE Trans. Inf. Technol. Biomed.* **2010**, *14*, 481–490. [CrossRef]
35. Tapia, E.; Intille, S.; Larson, K. Activity Recognition in the Home Using Simple and Ubiquitous Sensors. In Proceedings of the International Conference on Pervasive Computing, Linz and Vienna, Austria, 21–23 April 2004; Volume 3001, pp. 158–175.
36. LeBellego, G.; Noury, N.; Virone, G.; Mousseau, M.; Demongeot, J. A model for the measurement of patient activity in a hospital suite. *IEEE Trans. Inf. Technol. Biomed.* **2006**, *10*, 92–99. [CrossRef]
37. Barnes, N.; Edwards, N.; Rose, D.; Garner, P. Lifestyle monitoring technology for supported independence. *Comput. Control. Eng.* **1998**, *9*, 169–174. [CrossRef]
38. Tamura, T.; Kawarada, A.; Nambu, M.; Tsukada, A.; Sasaki, K.; Yamakoshi, K. E-healthcare at an experimental welfare techno house in Japan. *Open Med. Inform. J.* **2007**, *1*, 1–7. [CrossRef] [PubMed]
39. Boumpa, E.; Gkogkidis, A.; Charalampou, I.; Ntaliani, A.; Kakarountas, A.; Kokkinos, V. An Acoustic-Based Smart Home System for People Suffering from Dementia. *Technologies* **2019**, *7*, 29. [CrossRef]
40. Periša, M.; Teskera, P.; Cvitić, I.; Grgurević, I. Empowering People with Disabilities in Smart Homes Using Predictive Informing. *Sensors* **2025**, *25*, 284. [CrossRef]
41. Gong, Y.; Chung, Y.-A.; Glass, J. AST: Audio Spectrogram Transformer. *arXiv* **2021**, arXiv:2104.01778. [CrossRef]
42. Gemmeke, J.F.; Ellis, D.P.W.; Freedman, D.; Jansen, A.; Lawrence, W.; Moore, R.C.; Plakal, M.; Ritter, M. Audio Set: An ontology and human-labeled dataset for audio events. In Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017; pp. 776–780.
43. Ilias, L.; Askounis, D. Multimodal Deep Learning Models for Detecting Dementia From Speech and Transcripts. *Front. Aging Neurosci.* **2022**, *14*, 830943. [CrossRef]
44. Llaca-Sánchez, B.A.; García-Noguez, L.R.; Aceves-Fernández, M.A.; Takacs, A.; Tovar-Arriaga, S. Exploring LLM Embedding Potential for Dementia Detection Using Audio Transcripts. *Eng* **2025**, *6*, 163. [CrossRef]

45. Zhang, M.; Pan, Y.; Cui, Q.; Lü, Y.; Yu, W. Multimodal LLM for enhanced Alzheimer's Disease diagnosis: Interpretable feature extraction from Mini-Mental State Examination data. *Exp. Gerontol.* **2025**, *208*, 112812. [CrossRef]
46. Bang, J.; Han, S.; Kang, B. Alzheimer's disease recognition from spontaneous speech using large language models. *ETRI J.* **2024**, *46*, 96–105. [CrossRef]
47. Li, R.; Wang, X.; Berlowitz, D.; Mez, J.; Lin, H.; Yu, H. CARE-AD: A multi-agent large language model framework for Alzheimer's disease prediction using longitudinal clinical notes. *npj Digit. Med.* **2025**, *8*, 541. [CrossRef]
48. Du, X.; Novoa-Laurentiev, J.; Plasek, J.M.; Chuang, Y.W.; Wang, L.; Marshall, G.A.; Mueller, S.K.; Chang, F.; Datta, S.; Paek, H.; et al. Enhancing early detection of cognitive decline in the elderly: A comparative study utilizing large language models in clinical notes. *eBioMedicine* **2024**, *109*, 105401. [CrossRef] [PubMed]
49. Amini, S.; Hao, B.; Yang, J.; Karjadi, C.; Kolachalama, V.B.; Au, R.; Paschalidis, I.C. Prediction of Alzheimer's disease progression within 6 years using speech: A novel approach leveraging language models. *Alzheimer's Dement.* **2024**, *20*, 5262–5270. [CrossRef] [PubMed]
50. Agbavor, F.; Liang, H. Predicting dementia from spontaneous speech using large language models. *PLoS Digit. Health* **2022**, *1*, e0000168. [CrossRef] [PubMed]
51. Chen, L.; Asgari, M. Refining Automatic Speech Recognition System for Older Adults. In Proceedings of the ICASSP 2021—2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 6–11 June 2021; pp. 7003–7007. [CrossRef]
52. Hu, S.; Xie, X.; Geng, M.; Jin, Z.; Deng, J.; Li, G.; Wang, Y.; Cui, M.; Wang, T.; Meng, H.; et al. Self-Supervised ASR Models and Features for Dysarthric and Elderly Speech Recognition. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2024**, *32*, 3561. [CrossRef]
53. Geng, M.; Xie, X.; Ye, Z.; Wang, T.; Li, G.; Hu, S.; Liu, X.; Meng, H. Speaker Adaptation Using Spectro-Temporal Deep Features for Dysarthric and Elderly Speech Recognition. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2022**, *30*, 2597–2611. [CrossRef]
54. El Hajal, K.; Hermann, E.; Hovsepyan, S.; Doss, M.M. Unsupervised Rhythm and Voice Conversion to Improve ASR on Dysarthric Speech. *Proc. Interspeech* **2025**, *2025*, 2760–2764. [CrossRef]
55. Barry, S.J.; Dane, A.D.; Morice, A.H.; Walmsley, A.D. The automatic recognition and counting of cough. *Cough* **2006**, *2*, 8. [CrossRef]
56. Caring for a Person with Dementia A Practical Guide. Available online: https://www.alzheimers.org.uk/sites/default/files/2022-07/Caring-for-a-person-with-dementia-a-practical-guide.pdf (accessed on 20 November 2025).
57. Brodaty, H. The practice and ethics of dementia care. *Int. Psychogeriatr.* **2015**, *27*, 1579–1581. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

algorithms

MDPI

Article

Development and Evaluation of a Real-Time Home Monitoring Application Utilising Long Short-Term Memory Integrated in a Smartphone

Abdussalam Salama [1,*], Reza Saatchi [2,*], Maryam Bagheri [1], Mahpara Saleem [1] and Muhammad Usman Shad [1]

[1] School of Computing and Digital Technologies, Sheffield Hallam University, Sheffield S1 1WB, UK; maryam.bagheri@shu.ac.uk (M.B.); mahparasaleem42@gmail.com (M.S.); shad.muhammadusman@student.shu.ac.uk (M.U.S.)
[2] School of Engineering and Built Environment, Sheffield Hallam University, Sheffield S1 1WB, UK
* Correspondence: abdussalam.salama@shu.ac.uk (A.S.); r.saatchi@shu.ac.uk (R.S.)

Abstract

A novel real-time home monitoring application was developed that utilises long short-term memory (LSTM) and is integrated in a smartphone. Its personalised LSTM accurately learns to detect abnormal movement patterns. The application locally processes the smartphone's accelerometery data in the form of a signal magnitude vector (SMV) to analyse and interpret the movement patterns. The LSTM was conceptualised by a univariate time-series regression model. It adaptively updates its training parameters by processing the individual's last seven days of movement data, thus providing a stable, individualised, and dynamic activity baseline. It then quantifies the normal and abnormal movement patterns by continuously comparing the learnt information against the current accelerometery readings. An abnormal movement pattern, e.g., a fall or an unexpected period of inactivity triggers multi-channel alerts to care givers using SMS and email. The application's performance was evaluated using the data collected from 25 adult volunteers, aged 40–70 years. By interpreting their movement patterns, the application demonstrated a detection accuracy quantified by the coefficient of determination (R^2) = 0.93 and an absolute error of 0.05. This precision highlighted a low false positive rate in a real-world evaluation. The study successfully demonstrated a robust, cost-effective, and privacy-preserving home monitoring technology.

Keywords: remote patient monitoring; smartphone applications; Long Short-Term Memory (LSTM); elderly home monitoring; activity recognition; fall detection

1. Introduction

The welfare and safety of the elderly are paramount social and governmental concerns. A significant portion of the elderly live alone in their homes, often because they cannot afford specialised residential care facilities or they prefer their own home environment. This situation has resulted in incidents where a significant number of individuals are discovered deceased in their homes long after their passing [1]. Research indicated that the incidence of unobserved deaths has risen considerably in recent decades [2], underscoring an urgent need for effective home monitoring facilities. There is a technological need to detect and interpret abnormal or prolonged inactivity to alert care givers to perform a timely intervention [3].

Patient monitoring technologies have undergone considerable improvements in recent years, mainly due to advancements in sensing and data interpretation technologies. Smartphones are now ubiquitous and are utilised by individuals across all age groups. They contain integrated movement detection sensors, such as accelerometers, that can be conveniently adapted for remote detection of abnormal movement activities by the elderly. Despite their widespread adaptation, the potential of these devices for continuous health monitoring has remained underutilised. This gap is partly attributable to the extensive focus on dedicated Internet of Things (IoT) devices in healthcare. While IoT devices have extensive potential, they often have limitations concerning power consumption, user acceptability, and requirements for complex installation [4,5].

Existing artificial intelligence (AI)-based solutions increasingly rely on deep learning models. Although these models are powerful, they frequently encounter difficulties with personalised behavioural analysis in time-critical emergency detection scenarios [6,7]. Furthermore, these models are conventionally trained on extensive and centralised datasets which inevitably introduce significant user privacy concerns. This study directly addresses these technological challenges by presenting a novel remote patient home monitoring application, specifically designed to improve the safety of the elderly living independently in their own homes [7,8].

The innovative feature of this application is its effective integration of the smartphone's built-in accelerometer with AI interpretation of the user's movement patterns in real-time. Its on-device (i.e., integrated in the smartphone) deep learning LSTM model is specifically devised to distinguish between the user's normal movements and routine inactivity (such as sleeping), thus ensuring that alerts are only issued during true emergencies. Crucially, the LSTM model automatically trains on the individualised movement patterns after just one week of initial data collection, which significantly reduces the false alarms. The application also includes an option for follow-on update training when this becomes necessary. This intelligent, on-line training approach not only enhances the user privacy but also provides an essential reassurance to the caregivers by ensuring a timely intervention in emergency situations.

The primary contribution of this study is the development of an adaptive, personalised, real-time home monitoring application. The application effectively leverages accelerometery data and an LSTM-based AI model to reliably detect inactivity patterns. Its further contributions include the precise methodology for the LSTM's training and its evaluation using a smartphone. This approach has potential to improve response time and consequently enhance patient safety in emergencies, particularly for the elderly who may be at a higher risk of medical emergency. By implementing an on-line training protocol for the AI model, the application both guarantees enhanced privacy and reduces reliance on an external server, ensuring a timely intervention, even in a resource-limited environment. Therefore, this work provides a significant contribution to the field, supporting the creation of a personalised, scalable, and efficient monitoring application that safeguards patient's well-being and facilitates a rapid reaction during emergencies. In the following sections, the related literature is reviewed, the methodology used to design and evaluate the application is explained, and the results are discussed.

2. Literature Review

This section provides a review of the existing body of literature relevant to the current study. It aims to highlight the key contributions of the earlier studies, identify the existing limitations, and delineate the critical research gaps that necessitated the development of the application reported in this article.

2.1. Remote Patient Monitoring

The decision-making process in remote patient monitoring (RPM) is crucial for accurately recognising various activities, such as jogging, sitting, standing, and walking. Fuzzy logic is a technique previously used for this purpose, proving valuable for managing uncertainty when training data is limited [9]. However, a significant drawback is that the inherent rules in fuzzy systems are fixed and cannot adapt to evolving user behaviours, which can restrict their real-world flexibility. This inflexibility constitutes a key research gap, one that can be effectively addressed by employing adaptive AI models. The models like convolutional neural networks (CNNs) and LSTM networks can learn movement patterns over time, thereby potentially offering superior performance and individualization [10].

AI-based applications for patient home monitoring represent a paradigm shift in healthcare by enabling continuous tracking from remote settings. The ability of AI models to interpret complex information can provide informative decision making [11]. Integration of deep learning with the Internet of Medical Things (IoMT) has potential to assist with healthcare diagnostics, treatment, and patient monitoring [12]. Transformer-based deep learning models have been devised for real-time activity and fall detection of the elderly [13]. An adaptive dilated transformer bidirectional long short-term memory (Bi-LSTM) was used as part of a secure wearable sensor-based remote healthcare monitoring device [14]. The integration of smart devices, along with their built-in sensors, facilitates a real-time decision support environment. When these are combined with LSTM models, they facilitate an adaptive analysis that measurably enhances clinical decision-making [15,16]. By analysing individual patient datasets, these applications become capable of providing early detection of a patient's health deterioration. A primary strength of AI in remote patient monitoring (RPM) is its ability to learn and implement personalised healthcare behaviours, which significantly improves outcomes in managing chronic diseases and providing dedicated elderly care.

The benefits and risks in utilising RPM in different patient groups have received varied levels of investigation [17]. For example, RPM is cancer patients has been associated with a significant reduction in the length of hospital stay, emergency department visits and hospital readmissions [18]. RPM deployment should consider care transition from an inpatient hospital environment to a home setting to facilitate its successful adaptation [19]. The considerations include patient's safety, adherence to the prescribed medical guidance, quality of life and the financial implications. RPM development should involve affected patients through a codesign process. The codesign ensures an active participation of stakeholders, e.g., technologists, patients, and health providers, in the conceptualisation, design, and deployment of RPM devices, systems, and services [20]. The importance of a patient-centred approach to RPM development has been emphasised in a study involving high-risk oncology patients [21]. Retention of patients included in an RPM programme is an important consideration. In a study involving diabetic patients, it was reported that being younger or having higher blood glucose or blood pressure at the baseline increased the risk of withdrawing from an RPM programme [22]. A study has explored patients' experiences of RPM for chronic management focusing on health literacy and therapeutic relationships [23]. RPM deployment in rural and regional areas can have its own specific challenges that could include inadequate literacy, language barriers, disparity in IT infrastructure and devices, reliability of services, training opportunities, and funding models [24]. The devices used for RPM should have a carefully designed user interface for the intended patient group to ensure acceptability across a wide range of patients with varying IT skills [25].

The complexities in RPM have resulted in integrated care practices that accommodate technology design, service design, and organisational structures and policies [26]. A review

article has examined the pros and cons of RPM [27]. Its pros included issues such as providing continuous monitoring, facilitating early detection, cost efficiency, convenience, and comfort of the patients. For its limitations the article highlighted issues such as data accuracy and reliability, privacy and security risks, technical barriers, and compliance factors. A study has recommended considerations to assist in the design and implementation of RPM [28]. These included the codesign of RPM devices with the intended users, interactive capability, and having dedicated professional responsible for monitoring data and communication.

To successfully construct an intelligent application that can adaptively track a patient's unique movement activities across time, LSTM networks are an ideal choice. These networks are a specialised form of recurrent neural networks (RNNs), specifically designed to model sequential data [29]. When an LSTM is trained using accelerometery data, it gains the ability to analyse time-series activity patterns. It then uses this data to monitor the patient's overall activities and to respond appropriately to emergencies. However, these methods are not without their limitations. Traditional AI models are typically trained on powerful, centralised servers using extensive datasets. This conventional workflow necessitates that all personal data to be transmitted over the internet to remote systems, raising serious concerns regarding privacy breaches and potential security threats [30–32]. A substantially safer and more patient-centric approach involves training these models directly on the user's own smartphone. This approach keeps the individual's data confidential locally, ensuring that the users retain greater control over their collected personal information.

2.2. On-Device Training to Mitigate Data Risks

The development of on-device AI models has attracted considerable interest for its capability to facilitate real-time processing and enhance data privacy by keeping all computations on edge devices [33]. This approach effectively addresses the requirement for highly efficient health tracking solutions. In contrast to the cloud-based systems, for on-devices, their AI handles data processing directly on the smart device, thereby reducing reliance on external servers. This feature is especially pertinent for immediate-response situations, such as detecting falls or prolonged inactivity. A primary benefit of on-device AI approaches is their ability to offer continuous monitoring without requiring an active internet connection, significantly increasing reliability, particularly in remote areas. Moreover, by ensuring data is processed locally, an on-device solution substantially reduces the privacy concerns typically associated with server-oriented architectures [34]. This localised approach also reduces overall computational demands, which makes personalised health monitoring more widely accessible [35].

2.3. Activity Monitoring Using Smartphone Sensors Integrated with AI

Smartphone-based accelerometer sensors have rapidly emerged as promising tools for RPM. They can provide a low-cost, scalable, and user-friendly solutions for continuous health assessments. These integrated sensors facilitate the detection of physical activities without the requirement of an extra hardware. Recent research has demonstrated that an accelerometer-based data can classify human activities with a high degree of accuracy, strongly supporting their viability in mobile health applications [36].

Studies have reported that employing smartphone accelerometers for patient monitoring is both practical and practical to administer. As smartphones are already a routine part of daily life, patients are not compelled to wear intrusive, supplementary devices or fundamentally alter their daily routines. A study indicated that smartphones could reliably pick up changes in activity levels, thereby making healthcare more accessible and considerably less expensive, which is especially important for long-term care scenarios [35].

We summarise this literature review by highlighting four major challenges in the current RPM approaches that can limit their effectiveness. These include difficulties in operation by the elderly users; reliance on rigid, fixed fuzzy rules; the transmission of patient data over the internet, which raises security concerns; and a dependency on external sensors, which increases the cost and maintenance complexity due to required cloud services. To systematically address these limitations, our proposed solution offers several advantages. Firstly, it introduces a lightweight, user-friendly application that demands minimal interaction from the elderly. Secondly, it incorporates an adaptive deep neural network model based on LSTM to significantly enhance both accuracy and flexibility. Additionally, the entire system, including model training, operates exclusively on a smartphone that eliminates the need for data transmission and greatly enhances data privacy. Lastly, by utilising only the smartphone's built-in accelerometer, the solution removes the need to purchase supplementary equipment, thereby reducing the overall cost significantly.

3. Design of the RPM Architecture

The proposed application consists of an Android-based RPM software application, specifically designed to identify abnormal physical activities using the smartphone's integrated motion sensors. This includes an embedded LSTM-based machine learning inference engine that runs entirely on the smartphone. The overall architecture was optimised to operate autonomously on commercially available Android handsets, effectively removing the necessity for external devices or continuous connectivity to cloud services. The application functions by continuously monitoring the user's movements, capturing their real-time motion data, and analysing it locally using an LSTM model that identifies potential fall incidents or other abnormal movement patterns. Once detected, the application automatically initiates real-time alerts directed to designated caregivers.

Figure 1 illustrates the overall application's architecture and the flow of data. This diagram includes the sensor input, the processes of on-device data analysis, the classification stage, and the final alert notification module.

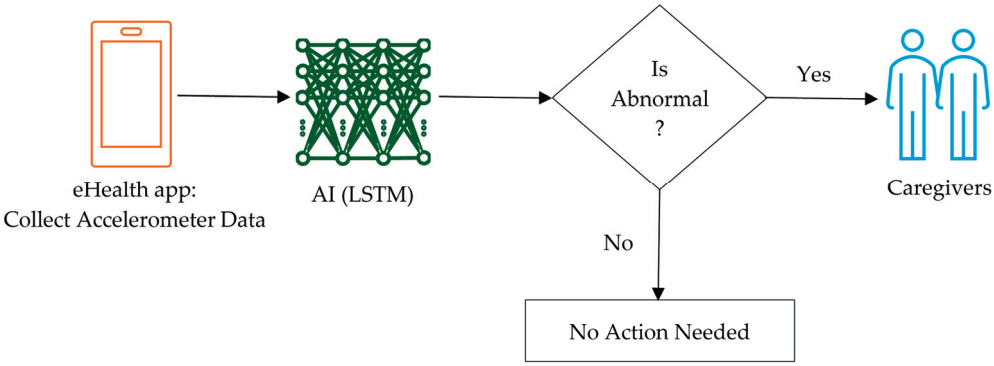


Figure 1. System architecture of the proposed system.

The core of the application is an LSTM artificial neural network, specifically implemented to detect unusual movement patterns and accurately identify fall incidents and unusual periods of inactivity. Once these events are detected, the application automatically triggers and sends real-time alerts to pre-designated caregivers through an email and an SMS. This section provides a comprehensive description of the application architecture, the process used for data collection, the configuration of the LSTM model, the alert mechanisms, and the final evaluation of the application's performance.

The core functionality of the application is delivered through a native Android application developed in Kotlin (version 1.9.0) utilising the Android Studio Hedgehog (2023.1.1).

To guarantee compatibility across current devices, the application was rigorously tested on a selection of smartphones operating on the Android 12 or a higher model. Testing devices included the Samsung Galaxy A53 5G, A73 5G, and M54 5G (Samsung Electronics, Suwon, Republic of Korea), the Huawei P30 and P40 (Huawei Technologies, Shenzhen, China), and the Google Pixel 7a (Google LLC, Mountain View, CA, USA).

The application captures motion data using the smartphone's integrated accelerometer sensor. This data is then pre-processed directly on the smartphone and temporarily stored within a local SQLite database. This strategy of using a lightweight offline data processing allows the application to remain functional even without continuous access to the internet. A machine learning inference engine, written in Python (version 3.10.x) and implemented using TensorFlow Lite (TFLite version 2.13.0), is fully integrated into the application to facilitate efficient, on-device monitoring. This engine employs a trained LSTM model to analyse motion patterns in real time and successfully identify potential fall incidents or abnormal movement patterns.

When an unusual movement is detected, the application immediately activates alerts across several mechanisms. These include email notifications, local alerts on the device, and emergency messages sent to caregivers whose details have been pre-configured using either SMS or internet-based messaging services. When no abnormal activity occurs, the application continues its silent operation in the background, ensuring the user's routine use of the smartphone is not interfered. This architecture delivers a solution for continuous real-time monitoring that is scalable, cost-effective, and conscious of user privacy. It is highly suitable for deployment in both home-based care and clinical environments. By deliberately removing the requirement for external hardware or persistent reliance on cloud connectivity, the system supports practical and efficient deployment across a wide range of real-world settings.

4. Materials and Methods

This study employed a quantitative experimental approach for the design, development, and evaluation of a smartphone-based home monitoring application underpinned by AI. The application was evaluated using data recorded from 25 healthy adult volunteers, comprising males and females aged 40 to 70 years. This diversity in age and gender ensured the application could be assessed across a wider range of user profiles, thereby enhancing the generalisability and real-world applicability of the findings. The study had ethics approval from Sheffield Hallam University Research Ethics Committee, the ethics application code is ER76195762.

4.1. Data Collection and Preprocessing

The application continuously gathered movement data from the smartphone's integrated triaxial accelerometer to monitor user activity in real time. The recorded raw sensor readings consisted of acceleration values along the x, y, and z axes of the phone's accelerometer, each paired with its corresponding timestamp. To effectively standardise the data, each axis value underwent normalisation by dividing its values by the standard gravitational acceleration (9.81 m/s^2), resulting in the normalised values x_g, y_g, and z_g. Based on these derived values, a signal magnitude vector (SMV) was subsequently generated to provide a measure of the overall intensity of movement, determined as follows:

$$SMV = \sqrt{x_g^2 + y_g^2 + z_g^2} \quad (1)$$

This approach is widely adopted in human activity recognition systems that rely on smartphones and has been proven effective at capturing meaningful motion patterns [37]. The raw accelerometer values were initially sampled at 200 Hz to ensure the capture of

fine-grained motion details, particularly sudden impacts. However, to ensure efficient on-device processing and long-term pattern analysis, this raw data underwent aggregation. The data points were temporarily collected in 1 s intervals and stored in an in-memory list over a 60 s window. At the end of each 60 s window, the highest SMV peak value recorded within the period was identified and saved, along with its corresponding timestamp, in a local SQLite database. This sequence of peak SMV values, representing a single data point per minute, served as the univariate time-series input for the LSTM model, thus enabling sequential analysis of user movements for the accurate detection of falls and other unusual motion patterns [38]. Using SQLite for local data storage provided several distinct advantages. It eliminated any reliance on cloud-based services, thereby ensuring that sensitive user data remained secure and private directly on the device. The application was specifically configured to retain up to three months of historical activity data, which facilitates both long-term movement pattern analysis and trend monitoring.

4.2. LSTM Model and Activity State Classification

At the end of every 60 s window, the application identifies the highest SMV (i.e., SMV_{actual}) from the recorded accelerometer readings. This observed value is then immediately compared to the corresponding predicted value generated by the on-device LSTM model $\left(SMV_{predicted}\right)$. This model has been trained to learn and recognise the user's normal movement patterns [39,40]. This critical comparative approach allows the system to effectively distinguish between the expected day-to-day fluctuations in movement activities and genuine, significant anomalies. The classification system is driven by two key derived metrics and a set of predefined thresholds (ϵ, τ_1, τ_2):

Prediction Error (E): This metric quantifies the absolute deviation of the observed movement activity from the LSTM's prediction:

$$E = \left|SMV_{actual} - SMV_{predicted}\right| \tag{2}$$

Relative Anomaly Score (A_R): This score is particularly sensitive to rapid, high-magnitude changes in movement intensity, calculated as the percentage increase in the actual SMV over the predicted SMV:

$$AR = (SMV_{actual}/SMV_{predicted}) - 1 \tag{3}$$

The application employs the pre-defined thresholds (ϵ, τ_1, τ_2) and applies them to these calculated metrics to subsequently classify the movement into one of five distinct states. The tolerance threshold for confidently declaring a "Normal" state is specifically defined as $\epsilon = 0.05$ g. Based on these metrics and the established thresholds, the application categorises each time window into one of five clear movement pattern states that are formally outlined in Table 1. The integration of LSTM-based pattern prediction with these dynamic, rule-based heuristics facilitates a highly fine-grained differentiation of real-world movement behaviour. Figure 2 visually represents the overall system flow, where the decision block labelled 'Is Abnormal?' encapsulates the entirety of the multi-state classification logic detailed within this section and Table 1, ultimately leading either to the initiation of an alert or to no further action.

Table 1. Activity state classification and formalised criteria.

State	Formalised Criteria	Description
Normal	$E \leq \epsilon$ (where $\epsilon = 0.05$ g)	Actual SMV matches the predicted value within tolerance, indicating expected routine movement.

Table 1. *Cont.*

State	Formalised Criteria	Description
Unexpected Activeness	$A_R > \tau_1$ (where $\tau_1 = 80\%$) AND $SMV_{predicted} \leq 1.0$ g	High anomaly score despite low predicted activity, indicating sudden unplanned movement.
Idleness	$SMV_{predicted} > 1.0$ g AND $E > \epsilon$ AND $SMV_{actual} < 1.0$ g	Predicted motion is moderate/high but actual SMV is low, indicating unexpected inactivity.
Fall Detected	$SMV_{predicted} \leq \tau_2$ (where $\tau_2 = 0.5$ g) AND $A_R > \tau_1$ (where $\tau_1 = 80\%$)	High-impact anomaly following a near-free-fall prediction, reducing false positives for fall events.
Abnormal State	$SMV_{actual} \leq 1.0$ g persists continuously for 12 h.	Long-term low activity, indicating prolonged inactivity.

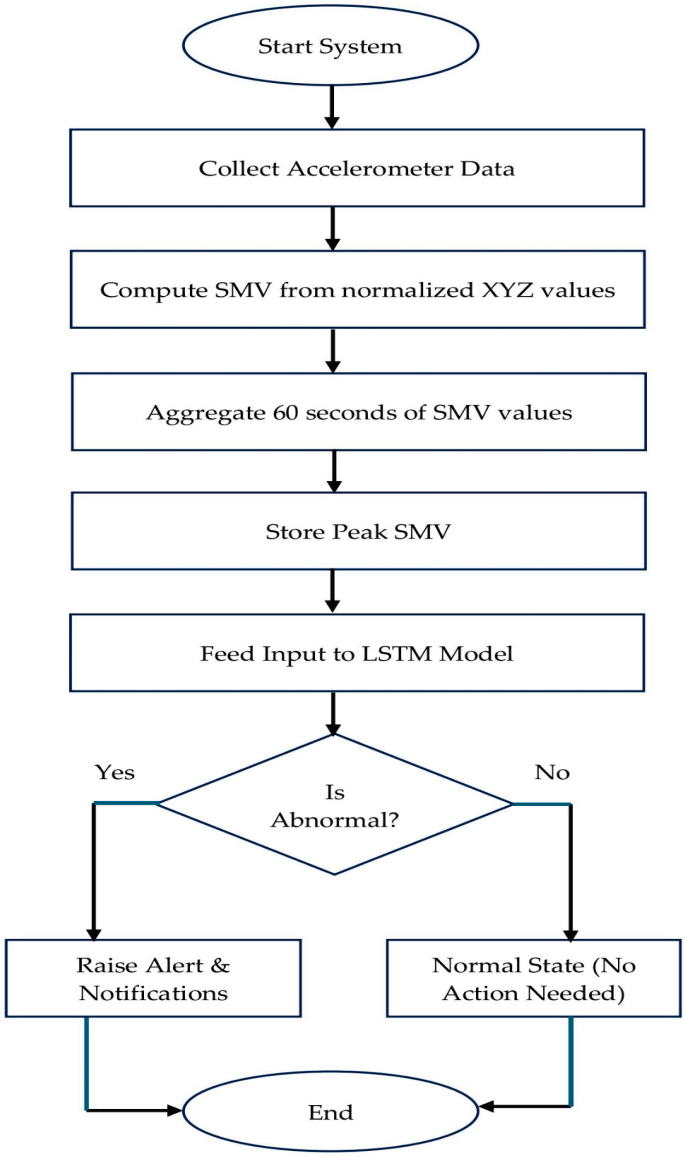


Figure 2. System flowchart.

4.3. LSTM Model Architecture and Training

The LSTM model, a variant of the recurrent neural network (RNN), was specifically chosen for its outstanding capability to recognise time-dependent patterns within sequential data. Unlike standard feedforward networks, the LSTM is equipped with an internal memory cell that allows it to selectively retain or discard information over various time

steps. This capability is essential for effectively analysing accelerometer signals, where the sequential order of a user's movements is necessary for correctly identifying activity patterns, such as walking, resting, or detecting a fall [41].

In this application, the LSTM is structured as a univariate time-series regression task focused on abnormal movement detection. Its core objective is to predict the expected peak signal magnitude vector (SMV) value for the upcoming minute $SMV_{predicted}$ by analysing the movement intensity patterns that preceded it. The input sequence consists of the last N = 60 peak SMV values (representing the preceding one hour of activity, with one value recorded per minute). This hour-long context is important for generating a reliable prediction. The output is a single predicted SMV value, which is then continuously compared against the actual observed peak SMV_{actual} within the current minute to determine the user's activity state, as outlined in Section 4.2 [39,40]. The architecture itself comprises an LSTM layer with 128 hidden units, followed by a Dropout layer with a 30% rate to mitigate overfitting. This feeds into a fully connected dense layer (64 neurons, (Rectified Linear Unit activation function), and finally, a single-neuron output layer with a linear activation function used for the regression task.

A key element of this system is its personalised, on-line training protocol, designed explicitly to counter the high false-alarm rates common in generic models. The application initiates a learning phase upon initial deployment, collecting the first seven days of the user's aggregated SMV data to establish a stable, personalised baseline. Following this, the model is fine-tuned on a rolling basis: once every 24 h, the LSTM undergoes a brief re-training process using only the most recent seven days of the user's data to adjust its internal weights. This daily fine-tuning involves a limited number of epochs (typically 5) and employs the Adam optimizer. By consistently updating its knowledge, the model adapts successfully to the routine changes, significantly minimising the prediction error for normal activities and reducing false positive alerts. The entire procedure is performed locally on the smartphone using TensorFlow Lite (TFLite), which guarantees both data privacy and operational autonomy.

To maintain high sensitivity to genuine anomalies whilst preventing the model from over-fitting to routine fluctuations, this daily re-training protocol is highly constrained. The system employs only the most recent seven days of the user's data for fine-tuning, restricting the adjustment to a maximum of 5 epochs. The Mean Squared Error (MSE) is used as the loss function. This choice strongly penalises larger prediction deviations that helps the model swiftly anchors itself to the individual's stable activity baseline. This approach achieves essential personalization without compromising predictive stability on the mobile hardware. Training was conducted using the Adam optimizer. The MSE was the primary loss function, with the Mean Absolute Error (MAE) tracked as a secondary evaluation metric. The overall protocol included a maximum of 50 epochs, a mini-batch size of 32, and an 80:20 split between the training and validation data. To further prevent overfitting and reduce unnecessary computation time, early stopping was implemented with a patience threshold set at 10 epochs. This configuration represents a practical compromise between achieving high accuracy, maintaining computational efficiency, and ensuring real-time performance on resource-constrained smartphone devices.

4.4. Alert Generation and Notification System

Once the user's activity has been classified based on the comparison between the actual and the predicted movement magnitudes, the application proceeds to determine whether the resulting state necessitates intervention. In the event of an abnormal state, e.g., a detected fall, unexpected idleness or sudden unusual activity, an alert is promptly generated and sent to the user's pre-configured emergency contacts using both email

and SMS communication channels. The alert was deliberately devised to be proactive, highly responsive, and unobtrusive. It functions asynchronously, which ensures that critical notifications are dispatched without interfering with the primary application workflow. When an emergency state is identified, the application automatically drafts an alert message, incorporating personalised details like the user's name and the contact's name. The email alert utilises a structured HTML template that clearly outlines the nature of the detected abnormal activity and suggests appropriate actions, such as checking on the individual and reviewing their recent data via the mobile application. This email is directed to the designated caregivers. Simultaneously, a concise SMS message with emergency wording is sent to ensure the fastest possible delivery and visibility. All generated alerts are systematically recorded in a local database, capturing the contact details, the precise time of the alert, and the reason for the notification to support future auditing and application evaluation.

If the user subsequently confirms that they are safe, for example, by interacting with a safety confirmation feature within the application, the application transmits a follow-up notification to the same contacts. This subsequent message reassures caregivers that no immediate action is required thereby effectively reducing unnecessary worry or panic. The entire alert mechanism is designed with fail-safety and redundancy as its core principles, supporting both internet-based communication (Email) and cellular networks SMS to maximise delivery reliability, even in situations where connectivity may be limited. This multi-channel, real-time notification application constitutes a critical element of a broader remote patient monitoring framework, ensuring a timely response to potentially life-threatening events.

5. Results

This section presents the experimental findings and the performance validation of the developed smartphone-based remote monitoring application. The application underwent rigorous evaluations with 25 volunteer participants. The participants were university staff, PhD students, and MSc students. Each participant installed and utilised the application on the provided mobile phone (all an identical model) over a continuous period of two months. These comprehensive evaluations were specifically designed to assess the application's effectiveness in detecting both critical abnormal events (such as simulated falls) and prolonged periods of inactivity, ensuring that timely alerts were reliably triggered when necessary.

5.1. Evaluation and Performance

The proposed LSTM-based activity classification application was evaluated through a combination of standard model performance metrics and rigorous real-world testing involving human participants. The comprehensive dataset used for this testing was gathered from 25 volunteers (both male and female, aged 40–70), who performed routine daily activities (for instance, walking, sitting, and standing) and simulated fall scenarios under controlled indoor conditions. These varied evaluation tasks ensured that the model was assessed across a wide range of activity profiles and abnormal motion patterns.

While the analysis involved all 25 participants, the specific time-series activity plots and the critical alert performance graph presented in this paper are derived solely from the two-month monitoring data of a single, representative individual (Participant-1, a male, 62 years old). This focused, case-study methodology was adopted because the personalisation trends observed, specifically the marked reduction in false positives during the first week of training were consistently representative of the overall group's general

pattern. Presenting this detailed, individualised data offers a clearer illustration of the LSTM's adaptive learning process than a potentially diluted, averaged group analysis.

For the regression-based fall detection model, performance was evaluated on a held-out validation set using standard regression metrics: Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R^2 (Coefficient of Determination). These results are shown in Table 2.

Table 2. Performance based on standard regression metrics.

Metric	Value
MAE (Mean Absolute Error)	0.045
MSE (Mean Square Error)	0.012
RMSE (Root Mean Square Error)	0.110
Score (Coefficient of Determination, R^2)	0.931

An R^2 value of 0.931 indicates that the model can account for 93.1% of the variance present in the actual SMV values that clearly demonstrates its strong generalisation ability. The observed low values for both MAE and RMSE further confirm that the predicted values remain consistently close to the observed values, even during periods when motion fluctuates rapidly. These results collectively prove the model's effectiveness in accurately predicting movement intensity and capturing the temporal dynamics inherent in human motion.

To rigorously evaluate the application's capacity to classify the five distinct activity states, we assessed its performance based on the formalised classification logic detailed in Section 4.2. Although the application achieved a strong overall classification accuracy of approximately 93% across all activity types, performance was primarily gauged using metrics that focus specifically on patient safety and the crucial need to minimise alert fatigue: sensitivity (true positive rate) and specificity (true negative rate).

The regression metrics presented in Table 2 form the foundational basis for the application's operation. The high (close to 1) R^2 value of 0.931 confirms that the personalised LSTM model successfully learns and replicates 93.1% of the variance observed in the individual user's movement patterns. This high level of predictive accuracy is essential, as the system's overall effectiveness is directly reliant on the model's capacity to accurately forecast normal movement. When the disparity between the predicted and actual SMV values exceeds the established thresholds (Equations (2) and (3)), it reliably signals a true anomaly. The subsequent section evaluates the results of the on-device personalisation against real-world data from the testing phase, which confirms the application's effectiveness in minimising false alarms, a significant limitation of typical generalised monitoring systems.

A visual comparison between the actual SMV values and the LSTM model's estimated SMV is displayed in Figure 3.

This specific example is derived from the data collected from a 62-year-old participant during the validation phase. While the figure focuses on a single individual, a similar close alignment was consistently observed across the other participants. Rather than merely forecasting future values, the LSTM model utilises sequential SMV inputs to identify patterns that are indicative of either normal or abnormal motion. The close tracking observed between the actual and estimated SMV values demonstrates the model's ability in recognising real-time variations in movement, thereby supporting accurate activity classification and timely alert generation. The predicted SMV values (illustrated in orange) generated by the LSTM model track the actual SMV values (illustrated in blue) very closely. This clearly demonstrates the model's robust ability to learn and accurately replicate real

motion patterns. This visualisation was generated using a Python script based on test sequences gathered from the human participants.

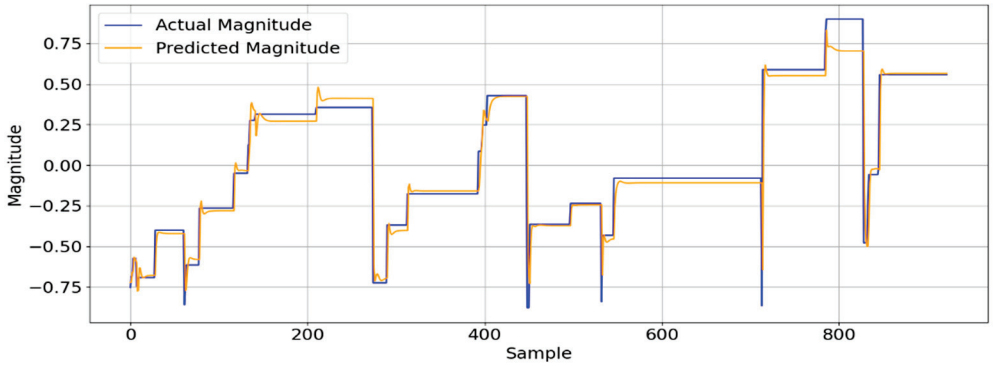


Figure 3. A typical plot comparing actual SMV (blue) and LSTM-estimated SMV (orange) for a 62-year-old participant.

5.2. Activity Monitoring and Behavioural Trends

This section presents the real-world operational results of the application, focusing specifically on how the monitored activity data is both collected and visualised. Most critically, it examines the direct influence of the on-device LSTM's personalised learning on alert reliability. The initial configuration of the mobile application was deliberately devised to be straightforward and user-friendly, requiring only minimal input from the user. The participant details form, shown in Figure 4, facilitates the entry of the monitored individual's demographic and contact information, including their full name, age, and location.

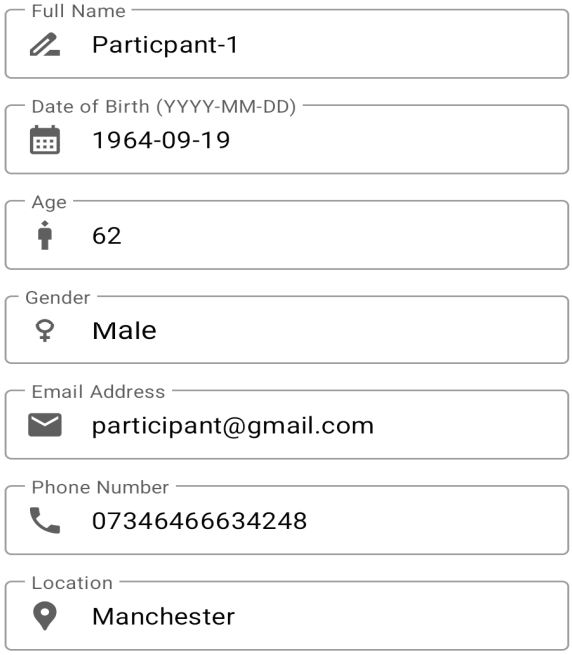


Figure 4. Participant details input form.

Complementing this, the Add/Update Caregiver interface, illustrated in Figure 5, provides a structured form for inputting caregiver details. The design of this form follows a sequential and minimal-interference methodology, necessitating that users only enter basic demographic and contact information. When previously stored caregiver details require

modification, the same interface is automatically repurposed into an updated mode, with fields pre-filled with the existing data to minimise the user's learning curve.

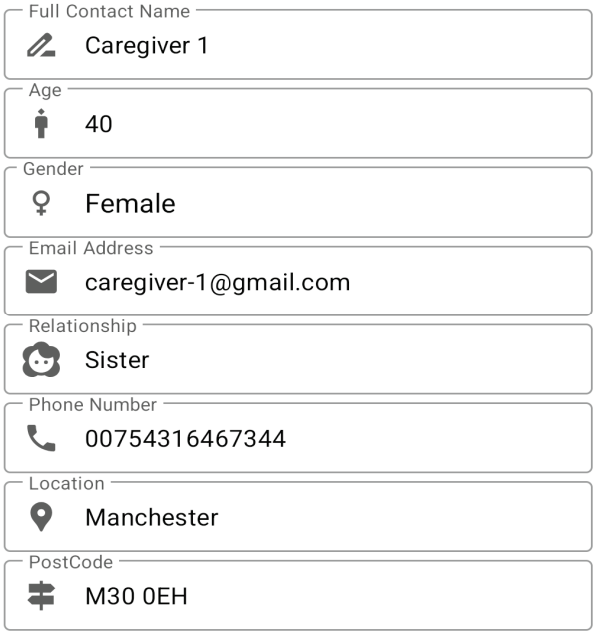


Figure 5. Caregiver details input form.

The Participant and Caregiver Details Summary Dashboard, shown in Figure 6, serves as a central hub that links the monitored individual with their caregiver. The screen provides a clear overview of the participant's key demographic information at the top, while a dedicated list below shows a summary of each registered caregiver's contact details and relationship to the participant.

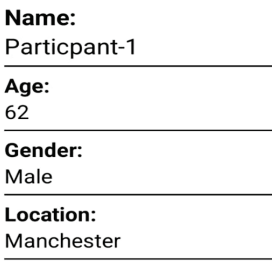


Figure 6. Participant and caregiver details summary dashboard.

The monitoring interface provides a clear, easily interpretable view of a user's activity levels across different time scales, which the on-device LSTM model then uses to learn the user's habitual, normal routine. As Figure 7 demonstrates, the "Hourly Activities" graph displays the recorded movement magnitude, highlighting the periods when activities are heightened.

These data points, visible as sharp spikes, indicate that the user is actively moving, probably using the phone, walking, or engaging in other movements while carrying the device. The model analysed these specific patterns to gain an understanding of the user's hourly movement patterns. The "Daily Activities" graph, represented in Figure 8, offers a broader view by summarising the maximum magnitude of movement recorded over each 24 h period.

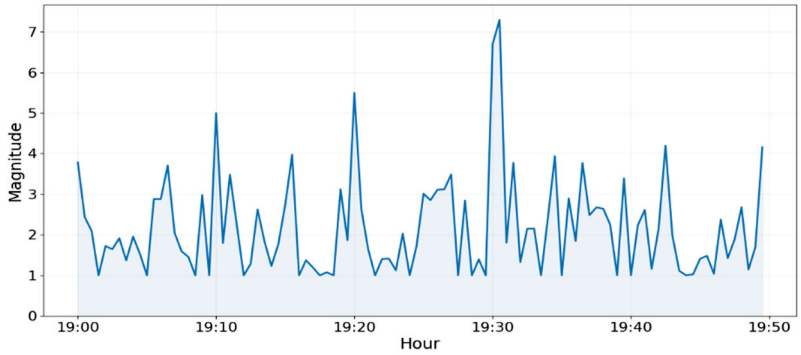


Figure 7. Hourly activity magnitude plot.

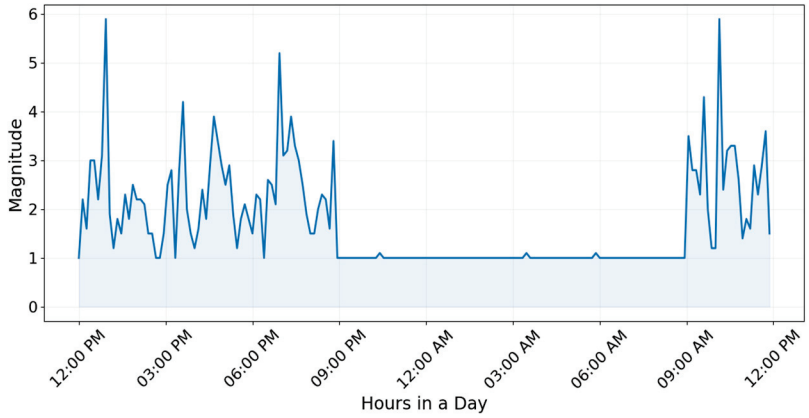


Figure 8. Daily activity magnitude plot.

The application utilises these daily trends as a core input for the LSTM model, enabling it to learn and appropriately account for natural variations in physical activities across days. By analysing these visualised plots, the application builds a robust understanding of the user's customary activity patterns, which is critical for identifying significant deviations that justify generating an alert.

The application continuously gathers and analyses motion data utilising the smartphone's integrated accelerometer. This data is processed by the on-device LSTM model that is trained to recognise the user's unique patterns of physical activity across time. By closely monitoring the daily and weekly variations in movement intensity, the model successfully establishes a personalised behavioural baseline for every user, making it far more effective at detecting anomalies.

The "Weekly Activities" bar chart (shown in Figure 9) visualises the maximum movement magnitude recorded each day over the past seven days.

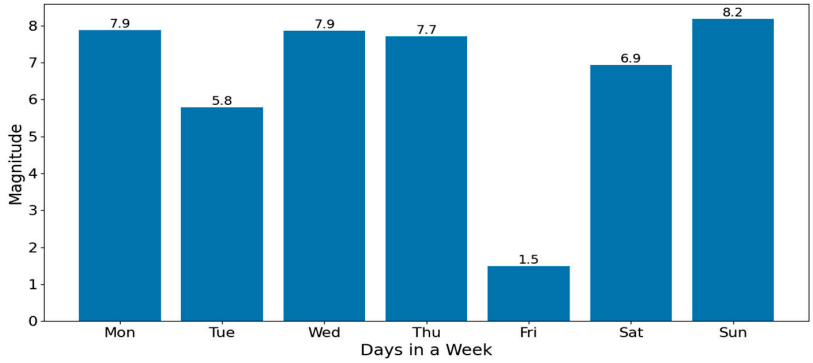


Figure 9. Weekly activity magnitude plot.

These values, which ranged from 1.48 to 8.18 in our evaluations, represent key indicators of the user's daily physical intensity. Higher values reflect more active days, while conversely, lower values suggest minimal movements. The LSTM model strategically uses these daily peaks to reinforce its understanding of the short-term fluctuations present in the user's routine.

Similarly, the "Monthly Activities" chart (shown in Figure 10) offers a broader perspective on long-term behaviour by displaying the maximum daily magnitude for every day of the month.

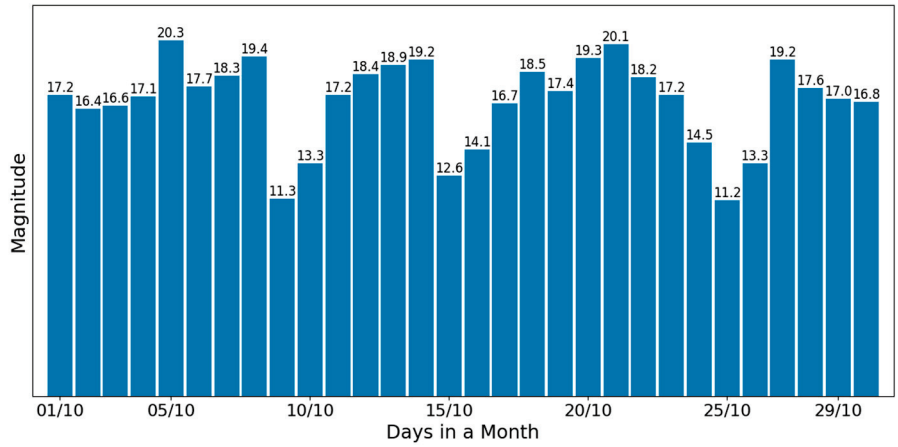


Figure 10. Monthly activity magnitude plot.

This extended history allows the model to further refine its baseline and successfully identify prolonged deviations from typical activity levels, e.g., sustained inactivity or unexpected surges in movement that could potentially signal emergencies requiring a caregiver's urgent attention.

The monitoring interface offers critical insights into both the system's operational performance and its decision-making process. Figure 11 displays the "Number of Alerts by Day" bar chart, which tracks the system's activity across the testing period.

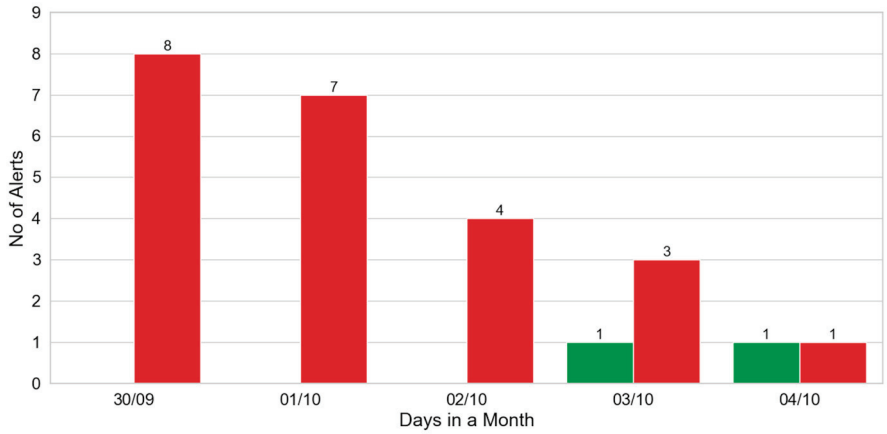


Figure 11. Alerts sent to caregiver.

The red bars on the chart represent the number of alerts successfully transmitted to a caregiver following the detection of an abnormal movement pattern. This figure visually confirms the success of the on-device fine-tuning protocol that was defined in Section 4.3 The x-axis indicates the days of the month during the test period, and the y-axis shows the corresponding number of alerts. initially, the red bars ('Not Okay') largely represent system-generated alerts that were effectively false positives. Crucially, as the LSTM

model has learnt the participant's unique movement patterns over the first few days (from 30 September 2025 to 4 October 2025), the number of these false alarms progressively decreased from 8 on the first day to just 1 on the final day. This outcome clearly demonstrates the model's ability to reduce false alarms and significantly enhance its predictive specificity through continuous, adaptive learning.

The application provides an Emergency Assistance interface to allow participants to communicate their status directly to their caregivers. This screen presents two clear, distinct buttons for immediate user action. If a genuine emergency arises, the participant can press the prominent 'Emergency!' button to instantly transmit both SMS and email alerts to the designated caregivers, ensuring a prompt notification and response. Alternatively, if the participant secures their safety, or if a system-generated alert was triggered unintentionally, they can press the 'I Am Okay!' button. This action sends a follow-up notification to the caregivers to confirm the user is safe and alleviates any unnecessary concern. This purposeful design supports timely communication, successfully reduces caregiver anxiety, and ensures that genuine emergencies are given the necessary priority for rapid intervention.

5.3. Mobile Application Notifications

Upon anomaly detection, the application consistently dispatches timely alerts to designated emergency contacts via email (using SMTP) and SMS to the primary carer. The alert content is structured for immediate action. The automated email carries the subject "Urgent: Abnormal Activity Detected" and explicitly instructs the caregiver to check on the individual and review the application's activities, providing three distinct required actions. The corresponding SMS message, designed for rapid delivery, provides a concise, urgent warning: "This is an automated message! [Participant-1] needs emergency! From eHealth."

Should the 'I Am Okay' button be selected, the user's status is immediately updated. Concurrently, both an SMS and an email notification containing the message 'All is well' are dispatched to the designated caregiver. Furthermore, when the 'I Am Okay' button is selected, the application registers this event locally in the SQLite database (referencing Tables 3 and 4), marking it as an instance where the alert required user override. This logging process is crucial, as it provides audit data on specific false positive instances which will be used to train future, more advanced versions of the system.

Table 3. The record identifier (ID), timestamp (Date/Time), tri-axial accelerometer values (*X*-axis, *Y*-axis, *Z*-value), acceleration magnitude, and participant's identifier (User ID). These evaluations took place on 15 August 2025.

ID	Date/Time	*X*-Axis	*Y*-Axis	*Z*-Value	Magnitude	User ID
8257	2025-08-15T21:59:55.70	−0.746	0.059	0.836	1.122	1
8258	2025-08-15T22:00:15.73	−0.901	0.310	0.635	1.145	1
8259	2025-08-15T22:00:35.74	−1.004	0.317	−0.209	1.074	1
8260	2025-08-15T22:00:55.75	−0.886	−0.0830	0.669	1.113	1
8261	2025-08-15T22:01:15.75	−0.995	−0.190	0.401	1.089	1
8262	2025-08-15T22:01:35.76	−0.984	0.086	0.352	1.049	1
8263	2025-08-15T22:01:56.21	−0.917	−0.150	0.539	1.074	1
8264	2025-08-15T22:02:16.79	−1.245	−0.055	0.446	1.324	1
8265	2025-08-15T22:02:16.79	−1.245	−0.055	0.446	1.324	1

Table 4. Log of real-time system alerts.

Alert ID	Date/Time of Alert	User ID	Caretaker Email	Alert Reason
1	2025-08-14T08:27:18	1	caretaker-1@gmail.com	Unexpected active state
2	2025-08-14T08:27:47.88	1	caretaker-1@gmail.com	Fall detected state
3	2025-08-14T08:28:02.88	1	caretaker-1@gmail.com	Fall detected state
4	2025-08-14T08:28:32.93	1	caretaker-1@gmail.com	Unexpected active state
5	2025-08-14T08:29:02.96	1	caretaker-1@gmail.com	Fall detected state
6	2025-08-14T08:29:17.98	1	caretaker-1@gmail.com	Fall detected state
7	2025-08-14T08:30:18.01	1	caretaker-1@gmail.com	Unexpected active state
8	2025-08-14T08:31:18.05	1	caretaker-1@gmail.com	Fall detected state
9	2025-08-14T08:31:18.05	1	caretaker-1@gmail.com	Fall detected state
10	2025-08-14T08:38:55.40	1	caretaker-1@gmail.com	Unexpected active state

5.4. Local Storage—SQLite Dataset

This section details the application's local storage design, which utilises an SQLite database. This design is fundamental to guaranteeing both operational efficiency and user data privacy by entirely removing any reliance on external cloud services. Table 3 provides a conceptual overview of the four critical components within this database schema. The following outlines the Data Schema Components.

The component displayed in Table 3, specifically the sensor data, functions as the primary data source for the LSTM model. It continually logs raw accelerometer readings, where the *X*-axis, *Y*-axis, and *Z*-axis columns record the normalised triaxial acceleration values. Crucially, the magnitude column stores the calculated signal magnitude vector (SMV). This time-series data, specifically the sequence of peak SMV values extracted from the magnitude column, is the direct input used for the on-device LSTM model to learn individual behavioural patterns (as discussed in Section 4.2). The integrity and volume of this data are essential for effective predictive regression.

The alerts information indicating the log of all real-time notifications is shown in Table 4. It indicates that the system classification (detailed in Section 4.2) has triggered an alert.

Each stored record includes details such as the Date/Time of Alert and, most importantly, the Alert Reason. The reasons, such as 'Fall detected state' and 'Unexpected active state,' are the direct result of the LSTM's anomaly detection combined with the subsequent heuristic ruleset. This complete log serves as a critical component for system auditing and is essential for verifying the effectiveness of the false positive reduction strategy (as was visually demonstrated in Figure 11). The Participant Information data shown in Table 5 indicate the manner of securely storing user-specific demographic data, including the individual's age, gender, and location. This data is utilised during the initial setup phase and is essential for ensuring that the entire monitoring and alert system is correctly personalized and accurately linked to the unique individual.

Table 5. The participant information data.

User ID	Name	DoB	Age	Gender	Email	Phone	Location
1	Particpant-1	19 September 1964	62	Male	participant@gmail.com	7346466634248	Manchester

The caregiver contact data shown in Table 6 indicates the manner of storing the designated emergency contacts. This data is essential as it ensures the system can efficiently retrieve necessary details like email and phone number to transmit automated alerts via SMS, email, or push notification in the event of a detected emergency.

Table 6. The caregiver contact data.

ID	Name	Relationship	Age	Gender	Phone	Email	Location	Postcode
1	CareTaker-1	Son	26	Male	74296057201	ms61@gmail.com	Manchester	M1 0AA

Tables 3–6 provide an overview of the four core components in the SQLite local storage schema. The local storage design enables secure data handling, seamless LSTM processing, responsive real-time monitoring through alert logging and contact retrieval.

6. Discussion and Future Work

There were studies utilisting smartphones for human activity detection (e.g., [42]). Our study however demonstrated an on-device, LSTM-based application offering a highly effective and practical solution for remote patient monitoring. It successfully addressed the key limitations associated with existing approaches. The quantitative results, particularly the high R^2 score of 0.931 and the low Mean Absolute Error (MAE), confirm the LSTM model's robust capability to accurately predict movement patterns based on raw accelerometer data. This high predictive accuracy is crucial, as it allows the application to establish a highly reliable baseline of an individual's normal behaviour.

The most significant contribution of this research is the application's inherent adaptability, which directly tackles the prevalent issue of false alarms often found in generic AI-based solutions. The "Number of Alerts by Day" chart (Figure 11) visually confirms that as the LSTM model has learnt each participant's unique daily routine over a week, the number of alerts progressively decreased. This on-device, personalised learning mitigates the need for large, pre-trained datasets and successfully eliminates the significant privacy risks associated with transmitting sensitive health data to external servers. Furthermore, the application's reliance on a smartphone's integrated sensors offers a cost-effective and scalable alternative to expensive, dedicated IoT devices.

The main limitation of the current study was that the evaluations were based on healthy adult volunteers. Patients with diverse medical conditions can have different movement pattens than healthy adults. Brain disorders such as Parkinson's disease that affect movement can have symptoms such as involuntary shaking, slow movement, and stiff muscles. These can affect the manner in which the application reacts and learns. Therefore, while this study effectively establishes the efficacy of the proposed application, several areas exist for future enhancement of its capabilities and real-world applicability. Firstly, to validate the application's overall generalizability, future studies should aim to validate it on a substantially larger and more diverse cohort of participants, including individuals with varying health conditions and mobility challenges. Secondly, the application could be enhanced by integrating data from other on-device sensors, such as a gyroscope, to further improve the accuracy of both activity classification and fall detection. Finally, future research could explore advanced machine learning techniques, such as reinforcement learning, to allow the model to learn more effectively from the user's "I am okay" overrides, which would make the system even more efficient and reduce false alarms more rapidly. It could also concentrate on expanding the model's capabilities to not only detect anomalies but also to classify a wider spectrum of daily activities, thereby providing a richer, more detailed log for proactive health monitoring. Finally, future studies will implement additional application-level protections, such as data encryption for the local SQLite database, to safeguard stored sensor data against device-level threats.

7. Conclusions

This study successfully developed and rigorously evaluated a novel, on-device remote patient monitoring application that offers a privacy-preserving and highly cost-effective solution specifically tailored for elderly individuals residing alone. By strategically leveraging the readily available, built-in accelerometer of a smartphone, the application eliminated the dependency on costly, dedicated hardware, and external cloud servers. The central and most impactful contribution of this work lie in the application's innovative use of a personalised Long Short-Term Memory (LSTM) model, which is trained exclusively and directly on the individual user's unique behavioural data.

The experimental results definitively demonstrated the substantial importance of adopting this on-device learning approach. As the LSTM model continuously learned and adapted to the user's motion patterns across the monitoring period, the number of false alarms, clearly visualized by a progressive decrease in unnecessary alerts sent to caregivers, was significantly mitigated. This critical adaptability enabled the application to reliably distinguish between normal daily routines and genuine, urgent anomalies, such as unexpected falls or extended periods of inactivity. By thoughtfully combining this intelligent, learning-based model with a user-override feature (the "I Am Okay" option), the application provided a notably reliable and trustworthy monitoring solution. Ultimately, this research offers a scalable and practical application that successfully improves patient safety in their home environment while simultaneously upholding their privacy, thereby proving the efficacy and readiness of on-device AI for critical, real-world health monitoring applications.

Author Contributions: A.S. provided overall project management, leadership, and administrative oversight. M.S. was primarily responsible for developing the methodology, designing the system architecture, and implementing the core LSTM functionality. M.S. specifically contributed to the design and refinement of the application's user features. The coding and implementation of all technical components were jointly handled by A.S., M.S., M.U.S. and M.B. took the lead on conducting the critical literature review to support the research work. M.B. also managed the gathering of user feedback and optimization efforts based on those real-world insights. R.S. performed the final validation and critical review of the manuscript. All named authors contributed to the writing, editing, and finalisation of the paper. All authors have read and agreed to the published version of the manuscript.

Funding: Financial support for this research was provided by Sheffield Hallam University (internal university funding, project reference N160).

Institutional Review Board Statement: Ethical approval for conducting this study was secured from Sheffield Hallam University. The protocol was thoroughly reviewed by the University's Research Ethics Committee. The ethics application code, ER76195762, received approval on 25 May 2025. All experimental procedures were strictly carried out in compliance with established institutional guidelines and regulatory requirements.

Informed Consent Statement: Informed consent was formally obtained from every subject who participated in this study.

Data Availability Statement: All data supporting the reported findings are entirely contained within this article.

Acknowledgments: The authors would like to thank all participants who took part in the evaluations undertaken in this study.

Conflicts of Interest: The authors declare that this research was conducted completely in the absence of any commercial or financial relationships that could be interpreted as potential conflicts of interest.

References

1. Salama, A.; Saatchi, R.; Bagheri, M.; Shebani, K.; Javed, Y.; Balaraman, R.; Adhikari, K. A Fuzzy Logic-Based eHealth Mobile App for Activity Detection and Behavioral Analysis in Remote Monitoring of Elderly People: A Pilot Study. *Symmetry* **2025**, *17*, 988. [CrossRef]
2. Hiam, L.; Estrin-Serlui, T.; Dorling, D.; McKee, M.; Minton, J. A deathly silence: Why has the number of people found decomposed in England and Wales been rising? *J. R. Soc. Med.* **2023**, *117*, 172–181. [CrossRef] [PubMed]
3. Kim, H.-J.; Park, J.-H.; Park, N.-Y.; Moon, H.J. Single-Person Households among the Socially Vulnerable Elderly Population: A Comparative Study of Housing, Living Habits, and Social Environments in Selected Neighborhoods of Seoul for the Prevention of Solitary Deaths. *Sustainability* **2024**, *16*, 7740. [CrossRef]
4. Stampfler, T.; Elgendi, M.; Fletcher, R.R.; Menon, C. Fall detection using accelerometer-based smartphones: Where do we go from here? *Front. Public Health* **2022**, *10*, 996021. [CrossRef]
5. Karar, M.E.; Shehata, H.I.; Reyad, O. A Survey of IoT-Based Fall Detection for Aiding Elderly Care: Sensors, Methods, Challenges and Future Trends. *Appl. Sci.* **2022**, *12*, 3276. [CrossRef]
6. Zhang, Y.; Tong, G.; Lin, C. Home Activity Recognition for Rural Elderly Based on Deep Learning and Smartphone Sensors. *J. Organ. End User Comput.* **2024**, *36*, 1–28. [CrossRef]
7. Wang, C.-Y.; Lin, F.-S. AI-Driven Privacy in Elderly Care: Developing a Comprehensive Solution for Camera-Based Monitoring of Older Adults. *Appl. Sci.* **2024**, *14*, 4150. [CrossRef]
8. Perez, A.J.; Siddiqui, F.; Zeadally, S.; Lane, D. A review of IoT systems to enable independence for the elderly and disabled individuals. *Internet Things* **2022**, *21*, 100653. [CrossRef]
9. Saatchi, R. Fuzzy logic concepts, developments and implementation. *Information* **2024**, *15*, 656. [CrossRef]
10. Vallabhaneni, A.S.; Perla, A.; Regalla, R.R.; Kumari, N. The power of personalization: AI-driven recommendations. In *Minds Unveiled*; Productivity Press: New York, NY, USA, 2024; pp. 111–127.
11. Shaik, T.; Tao, X.; Higgins, N.; Li, L.; Gururajan, R.; Zhou, X.; Acharya, U.R. Remote patient monitoring using artificial intelligence: Current state, applications, and challenges. *WIREs Data Min. Knowl. Discov.* **2023**, *13*, e1485. [CrossRef]
12. Mulo, J.; Liang, H.; Qian, M.; Biswas, M.; Rawal, B.; Guo, Y.; Yu, W. Navigating challenges and harnessing opportunities: Deep learning applications in internet of medical things. *Future Internet* **2025**, *17*, 107. [CrossRef]
13. Zafar, R.O.; Zafar, F. Real-time activity and fall detection using transformer-based deep learning models for elderly care applications. *BMJ Health Care Inform.* **2025**, *32*, e101439. [CrossRef]
14. Vishnupriya, G.; Anusha, S.; Kayikci, S. An efficient and secure wearable sensor based remote healthcare monitoring system using adaptive dilated transformer Bi-LSTM with gated recurrent unit. *Trans. Emerg. Telecommun. Technol.* **2024**, *35*, e4932. [CrossRef]
15. Rama, E.; Zuberi, S.; Aly, M.; Askari, A.; Iqbal, F.M. Clinical Outcomes of Passive Sensors in Remote Monitoring: A Systematic Review. *Sensors* **2025**, *25*, 3285. [CrossRef]
16. Amrani, H.; Micucci, D.; Napoletano, P. Personalized Models in Human Activity Recognition using Deep Learning. In Proceedings of the 2020 25th International Conference on Pattern Recognition (ICPR), Milan, Italy, 10–15 January 2021; pp. 9682–9688. [CrossRef]
17. Wartenberg, C.; Elden, H.; Frerichs, M.; Jivegård, L.L.; Magnusson, K.; Mourtzinis, G.; Nyström, O.; Quitz, K.; Sjöland, H.; Svanberg, T.; et al. Clinical benefits and risks of remote patient monitoring: An overview and assessment of methodological rigour of systematic reviews for selected patient groups. *BMC Health Serv. Res.* **2025**, *25*, 133. [CrossRef]
18. Rockey-Bartletthtt, C.; Morelli, J.; Coffel, M.; Geracitano, J.; Lafata, J.E.; Khairat, S. Effect of Remote Patient Monitoring on Healthcare Use Among Patients with Cancer: A Systematic Review. *Digit. Health (Sage J.)* **2025**, *11*, 1–16. [CrossRef]
19. Tan, S.Y.; Sumner, J.; Wang, Y.; Yip, A.W. A systematic review of the impacts of remote patient monitoring (RPM) interventions on safety, adherence, quality-of-life and cost-related outcomes. *npj Digit. Med.* **2024**, *7*, 192. [CrossRef] [PubMed]
20. Sumner, J.; Tan, S.Y.; Wang, Y.; Keck, C.H.S.; Lee, E.W.X.; Chew, E.H.H.; Yip, A.W. Co-designing remote patient monitoring technologies for inpatients: Systematic review. *J. Med. Internet Res.* **2024**, *26*, e58144. [CrossRef]
21. Mazzella-Ebstein, A.; Dunnack-Yackel, H.; Daly, R.M.; Salner, A.L.; Eanniello, M.K.; Fontaine, S.; Huang, J.; Bernal, C.; Wilhelm, C.; Ackerman, J.; et al. Looking ahead! The feasibility of implementing remote patient monitoring for high-risk oncology patients. *BMC Health Serv. Res.* **2025**, *25*, 918. [CrossRef]
22. Su, D.; Michaud, T.L.; Ern, J.; Li, J.; Chen, L.; Li, Y.; Shi, L.; Zhang, D.; Andersen, J.; Pagán, J.A. Diabetes management through remote patient monitoring: A mixed-methods evaluation of program enrollment and attrition. *Healthcare* **2025**, *13*, 698. [CrossRef]
23. Stevens, J.; Ghapanchi, A.H.; Purarjomandlangrudi, A.; Bruce, S. Patient experiences of remote patient monitoring: Implications for health literacy and therapeutic relationships. *Technologies* **2025**, *13*, 464. [CrossRef]
24. Tagne, J.F.; Burns, K.; O'Brein, T.; Chapman, W.; Cornell, P.; Huckvale, K.; Ameen, I.; Bishop, J.; Buccheri, A.; Reid, J.; et al. Challenges for remote patient monitoring programs in rural and regional areas: A qualitative study. *BMC Health Serv. Res.* **2025**, *25*, 374. [CrossRef]

25. Baumann, S.; Stone, R.T.; Abdelall, E. Introducing a remote patient monitoring usability impact model to overcome challenges. *Sensors* **2024**, *24*, 3977. [CrossRef]
26. Londral, A. Features of remote patient monitoring systems that implement integrated care: A perspective aligned with current challenges for digital health technologies: Comment on "Towards a framework for implementing remote patient monitoring from an integrated care perspective: A scoping review". *Int. J. Health Policy Manag.* **2025**, *14*, 8724.
27. Gouveia, A.J.; Ribeiro, J.; Sousa, G.; Barbosa, D.; Machado, M. Advancing healthcare through remote patient monitoring: A brief literature review. *Procedia Comput. Sci.* **2025**, *256*, 893–899. [CrossRef]
28. Thomas, E.E.; Taylor, M.L.; Banbury, A.; Snoswell, C.L.; Haydon, H.M.; Rejas, V.M.G.; Smith, A.C.; Caffery, L.J. Factors influencing the effectiveness of remote patient monitoring interventions: A realist review. *BMJ Open* **2021**, *11*, e051844. [CrossRef]
29. Kader, N.I.A.; Yusof, U.K.; Khalid, M.N.A.; Husain, N.R.N. A Review of Long Short-Term Memory Approach for Time Series Analysis and Forecasting. In *Lecture Notes in Networks and Systems, Proceedings of the 2nd International Conference on Emerging Technologies and Intelligent Systems, Chi Minh City, Vietnam, 29–30 September 2023*; Al-Sharafi, M.A., Al-Emran, M., Al-Kabi, M.N., Shaalan, K., Eds.; Springer: Cham, Switzerland, 2023; Volume 573. [CrossRef]
30. Kumar, M.; Kumar, A.; Verma, S.; Bhattacharya, P.; Ghimire, D.; Kim, S.-H.; Hosen, A.S.M.S. Healthcare Internet of Things (H-IoT): Current Trends, Future Prospects, Applications, Challenges, and Security Issues. *Electronics* **2023**, *12*, 2050. [CrossRef]
31. Alharbi, S.H.; Alzahrani, A.M.; Syed, T.A.; Alqahtany, S.S. Integrity and Privacy Assurance Framework for Remote Healthcare Monitoring Based on IoT. *Computers* **2024**, *13*, 164. [CrossRef]
32. Abbas, S.R.; Abbas, Z.; Zahir, A.; Lee, S.W. Federated Learning in Smart Healthcare: A Comprehensive Review on Privacy, Security, and Predictive Analytics with IoT Integration. *Healthcare* **2024**, *12*, 2587. [CrossRef] [PubMed]
33. Wang, X.; Tang, Z.; Guo, J.; Meng, T.; Wang, C.; Wang, T.; Jia, W. Empowering Edge Intelligence: A Comprehensive Survey on On-Device AI Models. *ACM Comput. Surv.* **2025**, *57*, 228. [CrossRef]
34. Andriulo, F.C.; Fiore, M.; Mongiello, M.; Traversa, E.; Zizzo, V. Edge Computing and Cloud Computing for Internet of Things: A Review. *Informatics* **2024**, *11*, 71. [CrossRef]
35. Abdellatif, A.A.; Mohamed, A.; Chiasserini, C.F.; Tlili, M.; Erbad, A. Edge Computing for Smart Health: Context-Aware Approaches, Opportunities, and Challenges. *IEEE Netw.* **2019**, *33*, 196–203. [CrossRef]
36. Torres-Huitzil, C.; Alvarez-Landero, A. Accelerometer-based human activity recognition in smartphones for healthcare services. In *Mobile Health*; Adibi, S., Ed.; Springer: Cham, Switzerland, 2015; Volume 5, pp. 87–105. [CrossRef]
37. Zhang, Y.; Yao, X.; Fei, Q.; Chen, Z. Smartphone sensors-based human activity recognition using feature selection and deep decision fusion. *IET Cyber-Phys. Syst. Theory Appl.* **2023**, *8*, 76–90. [CrossRef]
38. Bartczak, N.; Glanowska, M.; Kowalewicz, K.; Kunin, M.; Susik, R. Fall Detection Based on Recurrent Neural Networks and Accelerometer Data from Smartphones. *Appl. Sci.* **2025**, *15*, 6688. [CrossRef]
39. Mohammad, Z.; Anwary, A.R.; Mridha, M.F.; Shovon, S.H.; Vassallo, M. An Enhanced Ensemble Deep Neural Network Approach for Elderly Fall Detection System Based on Wearable Sensors. *Sensors* **2023**, *23*, 4774. [CrossRef]
40. Yu, Y.S.; Wie, S.; Lee, H.; Lee, J.; Kim, N.H. Long Short-Term Memory-Based Fall Detection by Frequency-Modulated Continuous Wave Millimeter-Wave Radar Sensor for Seniors Living Alone. *Appl. Sci.* **2025**, *15*, 8381. [CrossRef]
41. Li, Y.; Wang, L. Human Activity Recognition Based on Residual Network and BiLSTM. *Sensors* **2022**, *22*, 635. [CrossRef]
42. Bayat, A.; Pomplun, M.; Tran, D.A. A Study on Human Activity Recognition Using Accelerometer Data from Smartphones. *Procedia Comput. Sci.* **2014**, *34*, 450–457. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

algorithms

MDPI

Article

Chatting Ain't Diagnosing: Diagnostic Variability and Fundamental Errors in Multimodal LLM Interpretation in Radiology

Sungjoon Hong [1], Mihir Matalia [2] and Milan Toma [1,*]

[1] Department of Osteopathic Manipulative Medicine, College of Osteopathic Medicine, New York Institute of Technology, Old Westbury, NY 11568, USA; shong09@nyit.edu
[2] Academic Technologies Group, College of Osteopathic Medicine, New York Institute of Technology, Old Westbury, NY 11568, USA; mihir.matalia@nyit.edu
* Correspondence: mtoma@nyit.edu or tomamil@tomamil.com

Abstract

Recent studies investigating the diagnostic capabilities of large language models (LLMs) have attracted significant media attention, often resulting in headlines claiming that AI systems can match or even outperform physicians. As LLMs have rapidly proliferated, this has fueled a widespread misconception that they represent the cutting edge of artificial intelligence in all contexts. This narrative tends to overshadow the continued importance of task-specific machine learning models, which were developed and validated for particular diagnostic applications well before the rise of LLMs. This single-case study evaluated the reliability of five leading multimodal LLMs (GPT-5, Gemini 3 Pro, Llama 4 Maverick, Grok 4, and Claude Opus 4.5 Extended) for radiological image interpretation by presenting each model with an identical non-contrast head CT demonstrating intracranial pathology, complemented by a novel cross-evaluation protocol wherein each model graded all responses. The deliberate use of a straightforward case (rather than diagnostically challenging pathology) aimed to establish minimum competency thresholds; if LLMs cannot reliably interpret obvious pathology, their deployment on ambiguous cases becomes indefensible. The study intentionally excluded human radiologist ground truth to avoid generating comparative accuracy metrics that could be selectively cited for commercial purposes, focusing instead on demonstrating class-wide limitations rather than ranking individual products. Results revealed a 20% rate of fundamental diagnostic error, with one model misidentifying ischemic stroke as intracerebral hemorrhage with incorrect lateralization. Even among concordant models, clinically meaningful variability persisted in acuity characterization, anatomical localization, and differential diagnoses. Cross-evaluation exposed ground truth disagreement between models, self-evaluation bias, inconsistent grading stringency, and divergent evaluation philosophies. Only one model included appropriate safety disclaimers. These findings demonstrate that current multimodal LLMs exhibit unacceptable diagnostic variability and evaluative inconsistency for autonomous clinical deployment. The appropriate clinical role for LLMs should be distinguished by deployment context: autonomous diagnosis requires validated task-specific models; decision support applications demand rigorous radiologist oversight protocols; and educational summarization represents the most appropriate current use case, with mandatory disclaimers. Healthcare applications requiring reliable image interpretation should prioritize validated, task-specific machine learning systems over general-purpose language models.

Keywords: large language models; multimodal artificial intelligence; medical image interpretation; radiology; computed tomography; diagnostic accuracy; inter-model variability; ischemic stroke; artificial intelligence in healthcare; AI safety

1. Introduction

The public discourse surrounding artificial intelligence (AI) has become increasingly synonymous with large language models (LLMs), i.e., systems such as ChatGPT, Claude, Gemini, and their contemporaries that have captured global attention since 2022. This conflation, while understandable given the unprecedented visibility these tools have achieved, obscures a fundamental reality: AI as a field has been under active development for over half a century, and LLMs represent merely the most recent (and in many respects, most narrowly specialized) manifestation of a far broader technological landscape.

1.1. A Brief History of Artificial Intelligence

Understanding why large language models are ill-suited for medical image interpretation requires situating them within the broader landscape of artificial intelligence, a field of which seven-decade history has produced diverse approaches to machine cognition, each with distinct strengths and limitations. The current public fascination with LLMs obscures this rich technological heritage and, more critically, the specialized diagnostic tools that emerged from it.

The foundations of artificial intelligence were laid in the 1950s, with Alan Turing's seminal 1950 paper "Computing Machinery and Intelligence" [1] and the 1956 Dartmouth Conference [2] that formally established AI as a field of academic inquiry. In the decades that followed, AI research progressed through multiple paradigms: symbolic reasoning and expert systems in the 1960s–1980s, neural network revivals in the 1980s–1990s, and the machine learning renaissance of the 2000s–2010s driven by increased computational power and data availability [3–5]. Deep learning architectures, particularly convolutional neural networks (CNNs), achieved breakthrough performance in image recognition tasks by the early 2010s [6], years before the transformer architecture that underlies modern LLMs was even conceived.

The transformer architecture was introduced in 2017, and foundational models such as BERT and GPT emerged in 2018 [7,8]. However, it was not until GPT-3 in 2020 (with its 175 billion parameters) that LLMs began to demonstrate the conversational fluency that would later captivate public attention [9,10]. The release of ChatGPT in November 2022 marked the inflection point at which LLMs transitioned from technical curiosities to mainstream phenomena [11]. In this context, the entire history of publicly prominent LLMs spans barely three years; which is a tiny fraction of AI's seven-decade evolution [11,12].

1.2. The Misconception: LLMs as the Totality of AI

The explosive visibility of LLMs has fostered a pervasive misconception: that "AI" and "LLM" are interchangeable terms, and that these language-centric systems represent the state of the art across all AI applications. Such conflation complicates ongoing efforts to define AI's transformative potential in healthcare and navigate its ethical implications [13]. Misunderstanding the distinction can have serious consequences, particularly in domains where task-specific machine learning models (developed and validated long before the LLM era) remain the appropriate technological choice.

The very nomenclature is instructive. Large language models are, by design and training objective, optimized for linguistic tasks: text generation, summarization, translation,

https://doi.org/10.3390/a19030170

and conversational interaction. Their architecture (i.e, the transformer) was developed specifically for sequence-to-sequence modeling in natural language processing. While multimodal variants have emerged that can process images alongside text, with applications extending to domains such as orthodontic biomechanical reasoning [14], these capabilities represent extensions grafted onto a fundamentally language-centric architecture, not systems purpose-built for visual analysis.

The viral launch of ChatGPT in November 2022 marked an inflection point in public awareness of artificial intelligence capabilities (Figure 1). Within months, large language models transitioned from specialized research tools to household technologies, fundamentally altering public expectations regarding AI performance across diverse domains. This rapid mainstreaming, however, has fostered a pervasive misconception: that general-purpose LLMs represent the state of the art across all artificial intelligence applications, including medical image interpretation.

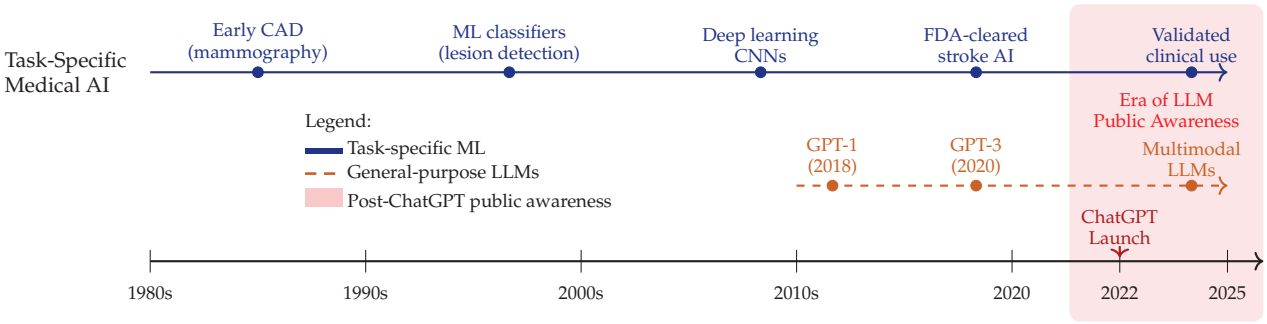


Figure 1. Timeline illustrating the divergent development trajectories of task-specific medical machine learning and general-purpose large language models. Task-specific ML systems for medical imaging (blue) have undergone decades of development, validation, and regulatory clearance, culminating in FDA-cleared diagnostic AI deployed in clinical practice. General-purpose LLMs (orange, dashed) emerged more recently and achieved widespread public awareness following the ChatGPT launch in November 2022 (shaded region).

As illustrated in Figure 1, this perception obscures a critical distinction. Task-specific machine learning models (developed, validated, and in many cases regulatory-cleared for defined diagnostic imaging tasks) have been advancing through detailed clinical validation pathways for decades. These systems remain the appropriate technological choice for applications requiring reliable, reproducible diagnostic accuracy. The conflation of LLM capabilities with validated medical AI risks inappropriate deployment of unvalidated tools in contexts where patient safety demands proven performance.

Recent studies examining LLM diagnostic capabilities have garnered substantial media attention, often generating headlines that proclaim AI systems can match or exceed physician performance [15,16]. Figure 2 illustrates findings from a randomized controlled trial by Goh et al., in which GPT-4 alone achieved a diagnostic reasoning score of 92%, i.e., significantly higher than physicians using conventional resources (74%, $p = 0.03$) [15]. Such results, when reported without context, fuel public perception that LLMs represent reliable diagnostic tools ready for clinical deployment. However, it is necessary to note that the Goh study evaluated LLM performance on text-based clinical vignettes, i.e., a task that leverages the core linguistic reasoning capabilities for which these models were designed. Medical diagnosis in practice frequently requires interpretation of imaging studies, ranging from cross-sectional imaging to specialized reports such as visual field tests [17], a fundamentally different task that tests the multimodal visual capabilities of these systems rather than their text-processing strengths.

https://doi.org/10.3390/a19030170

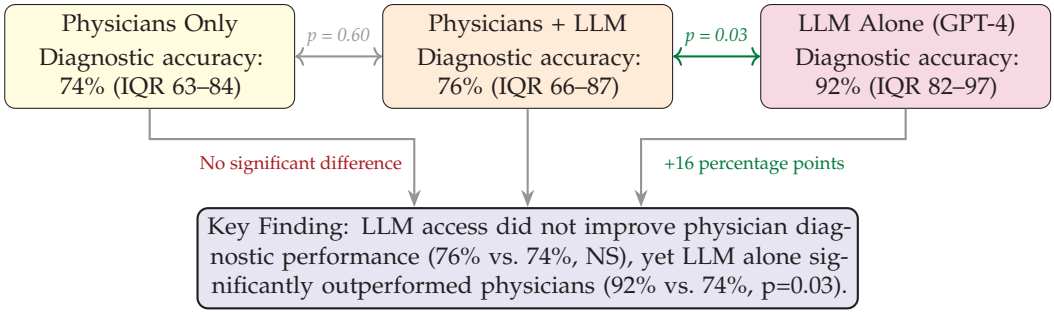


Figure 2. Summary of key findings from a randomized controlled trial examining physician diagnostic reasoning with and without LLM access [15]. Despite the LLM alone achieving significantly higher scores (92%) than physicians using conventional resources (74%, $p = 0.03$), physician access to the LLM produced no significant improvement (76% vs. 74%, $p = 0.60$).

1.3. Machine Learning in Radiology: A Parallel History

The application of machine learning to medical imaging predates the LLM era by decades, as illustrated in Figure 1. Foundational work began in the early 1960s, when Lodwick et al. introduced the first computational analysis of chest X-rays for lung cancer classification using manual feature coding and rule-based computer analysis [18]. Lodwick subsequently coined the term "computer-aided diagnosis" (CAD) and outlined a research plan for automated classification of radiological images [19]. Throughout the 1970s, these methods advanced significantly: Kruger et al. applied automated feature extraction and statistical classification to chest radiographs for cardiac analysis [20], Toriwaki et al. employed image processing and pattern recognition to automatically measure the cardiothoracic ratio [21], and further work by Kruger et al. advanced pattern recognition techniques for automated abnormality detection in chest radiographs [22,23]. These early efforts laid the groundwork for CAD systems for mammography, which received FDA approval as early as 1998 [24]. Deep learning approaches to radiological image analysis began demonstrating expert-level performance in the mid-2010s, with landmark studies showing CNN-based systems matching or exceeding dermatologist accuracy in skin lesion classification (2017) [25], radiologist performance in chest X-ray interpretation (2017–2018) [26], and ophthalmologist accuracy in diabetic retinopathy screening (2016) [27]. Recent comprehensive reviews have further documented the expanding applications of machine learning across medical imaging modalities [28,29].

These systems share a critical characteristic: they are task-specific. A CNN trained to detect acute ischemic stroke on non-contrast head CT has been optimized exclusively for that task, with training datasets comprising thousands to millions of expert-annotated images of that specific pathology in that specific imaging modality [30]. This focused training enables the model to learn the precise visual features (subtle hypodensities, loss of gray–white differentiation, insular ribbon signs, hyperdense vessel signs) that distinguish pathology from normalcy in that defined context. Similar validation strategies have been applied successfully to automated MRI-based classification tasks in neurological diagnosis, demonstrating the efficacy of task-specific approaches [31].

By contrast, multimodal LLMs acquire their image interpretation capabilities as a byproduct of training on heterogeneous image–text pairs scraped from diverse internet sources. Medical imaging content constitutes a small and potentially unrepresentative fraction of this training corpus [32]. The resulting systems possess broad but shallow capabilities: they can describe what they "see" in an image using medically plausible language, but their visual representations have not been optimized for the discriminative features that matter for specific diagnostic tasks. Comparative studies examining multi-

https://doi.org/10.3390/a19030170

modal LLMs against purpose-built feature extraction methods in clinical contexts, such as autism prediction from home videos, further illuminate these architectural limitations [33].

1.4. The Critical Distinction: Language Generation Versus Diagnostic Accuracy

The core competency of an LLM is generating contextually appropriate text [34]. When presented with a medical image, a multimodal LLM generates a textual description that is linguistically coherent and stylistically consistent with radiological reporting conventions. This output may appear authoritative (e.g., formatted in bullet points, employing correct anatomical terminology, offering differential diagnoses) yet the relationship between this generated text and the actual pixel-level content of the image is fundamentally different from that of a purpose-trained diagnostic model.

A task-specific stroke detection algorithm processes the image through layers of learned convolutional filters optimized to detect the visual signatures of ischemic injury. Standardized validation frameworks for clinically actionable machine learning models emphasize the importance of detailed performance characterization across diverse datasets [35]. Its output (i.e., stroke present/absent, with localization) is a direct function of learned visual features validated against confirmed diagnoses. An LLM, by contrast, generates a textual response that is statistically consistent with the kinds of reports that accompanied similar-seeming images in its training data. The LLM is not "diagnosing" in any meaningful sense; it is predicting what text a radiologist might have written, based on pattern matching against its training distribution.

This distinction has profound implications for reliability. A well-validated task-specific model has quantified performance characteristics: known sensitivity, specificity, positive and negative predictive values, and characterized failure modes. Its errors are systematic and discoverable through rigorous validation. An LLM's errors, by contrast, are emergent properties of a system not designed for the task at hand, potentially unpredictable, and difficult to characterize systematically.

1.5. The Present Study: Empirical Demonstration of LLM Limitations

This study was designed to empirically demonstrate the limitations of current multimodal LLMs for medical image interpretation, thereby illustrating why the broader AI ecosystem (particularly task-specific machine learning models) remains essential for clinical diagnostic applications.

We evaluated five leading multimodal LLMs (GPT-5, Gemini 3 Pro, Llama 4 Maverick, Grok 4, and Claude Opus 4.5 Extended) on a standardized radiological interpretation task: analysis of a non-contrast head CT demonstrating intracranial pathology. Beyond assessing diagnostic accuracy, we implemented a novel cross-evaluation protocol in which each LLM graded the responses of all five models, enabling assessment of evaluative consistency and the reliability of LLM-based quality assurance.

Our findings reveal that current LLMs exhibit unacceptable variability in diagnostic interpretation, including fundamental errors (misidentifying ischemic stroke as hemorrhage with incorrect lateralization) and categorical disagreements about basic radiological findings (acute versus chronic infarction). Even among models that reached the correct general diagnosis, clinically meaningful variability persisted in acuity characterization, anatomical localization, and differential diagnosis generation. The cross-evaluation protocol further revealed that LLMs cannot reliably agree on radiological ground truth, undermining proposals for their use in automated quality assurance or educational assessment.

Importantly, this study was designed to demonstrate the general unreliability of multimodal LLMs for diagnostic imaging interpretation, not to rank or compare the performance of individual models against one another or against human radiologists. The authors have

https://doi.org/10.3390/a19030170

no relationship (financial, professional, or otherwise) with any AI system or developer, including those whose products were evaluated (OpenAI, Google DeepMind, Meta AI, xAI, or Anthropic), nor with any other AI systems or developers not included in this study. We deliberately avoided comparative accuracy assessments against human radiologist ground truth because such comparisons risk shifting focus toward determining "which LLM performs best," which could inadvertently serve as product endorsement. Our objective is not to identify superior products within an unreliable class of tools. Instead, we aim to demonstrate that the entire class of general-purpose multimodal LLMs exhibits fundamental limitations that render them unsuitable for autonomous diagnostic imaging interpretation.

These findings support the position that LLMs, despite their impressive linguistic capabilities, are not appropriate tools for medical image interpretation requiring diagnostic accuracy. The appropriate AI technology for such applications remains task-specific machine learning models: systems designed, trained, and validated specifically for defined imaging modalities and pathological conditions. LLMs may serve valuable roles in healthcare (clinical documentation, patient communication, literature synthesis) but the pattern recognition demands of diagnostic radiology require purpose-built solutions that the current generation of language models cannot reliably provide.

The sections that follow detail our methodology, present our findings, and discuss the implications of this work for the responsible integration of AI technologies into clinical practice. We argue that the current enthusiasm for LLMs must be tempered by recognition of their fundamental limitations and that healthcare institutions, AI developers, and regulatory bodies must maintain clear distinctions between general-purpose language systems and validated diagnostic AI tools.

2. Materials and Methods

To empirically evaluate the reliability of multimodal large language models for radiological image interpretation, we designed a two-phase experimental protocol. The first phase assessed diagnostic performance by presenting five leading multimodal LLMs with an identical medical imaging task and systematically comparing their outputs across clinically relevant dimensions. The second phase implemented a novel cross-evaluation protocol wherein each LLM graded the responses generated by all five models (including its own, presented anonymously), enabling assessment of evaluative consistency, self-evaluation bias, and inter-model agreement on radiological ground truth. This dual-phase approach allowed examination of LLM reliability from two complementary perspectives: their capacity to generate accurate diagnostic interpretations and their capacity to evaluate the quality of such interpretations; both of which are prerequisite competencies for any proposed clinical or educational deployment. The following subsections detail the selection criteria for models and test images, the standardized prompts employed, the grading rubric and cross-evaluation methodology, and the analytical framework used to assess inter-model variability and agreement.

2.1. Selection of Large Language Models

Five LLMs with vision capabilities were selected for this comparative analysis. The selection criteria included: (1) demonstrated ability to process and interpret medical imaging data, (2) public availability or API access for research purposes, (3) representation among the most widely used and cited multimodal AI systems as of late 2025, and (4) diverse model architectures and training paradigms to capture variability across different development approaches. The selected models were: GPT-5 (OpenAI), Gemini 3 Pro (Google DeepMind), Llama 4 Maverick (Meta AI), Grok 4 (xAI), and Claude Opus 4.5 Extended (Anthropic).

https://doi.org/10.3390/a19030170

2.2. Image Selection and Prompt Design

A non-contrast axial CT scan of the head demonstrating intracranial pathology was selected as the test image (Figure 3). No external radiologist ground truth labeling was incorporated into the study design. This methodological choice was deliberate: the selected image depicts unambiguous intracranial pathology that does not require subspecialty expertise to recognize as abnormal. The relevant clinical question is not whether expert radiologists can correctly interpret this straightforward case (they can), but whether state-of-the-art LLMs can achieve consistent, accurate interpretations when presented with a case designed to be easy. By selecting an obviously pathological image rather than a diagnostically challenging case, we aimed to establish a minimum competency threshold. If LLMs cannot reliably interpret straightforward pathology, their deployment on ambiguous or challenging cases becomes indefensible. Furthermore, incorporating human radiologist grading would have introduced a comparative dimension (LLM vs. human accuracy) that risks framing certain models as "better" or "closer to human performance," which is a characterization we specifically sought to avoid given our lack of commercial relationships with any AI developer and our objective of demonstrating class-wide limitations rather than identifying superior products.

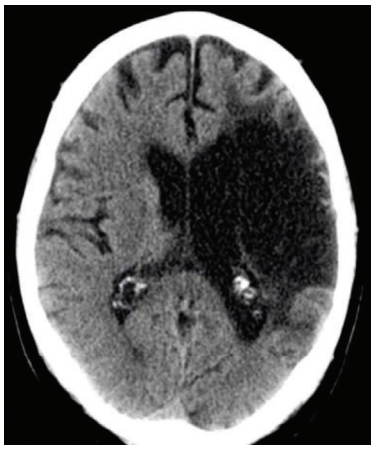


Figure 3. Non-contrast axial CT scan of the head at the level of the lateral ventricles and basal ganglia, demonstrating intracranial pathology. This image served as the standardized test case for all five LLM evaluations. Ground truth interpretation and clinical implications are discussed in the Results section.

The same standardized prompt was submitted to all five LLMs to ensure consistency in evaluation parameters. The prompt was designed to elicit responses across multiple clinically relevant dimensions:

A standardized prompt was submitted to all five LLMs instructing them to analyze the image as a board-certified radiologist across five domains: imaging modality, anatomical region, primary diagnosis, supporting features, and differential diagnoses. The complete prompt text is provided in Appendix A.1, and all technical parameters relevant to reproducibility are documented in Section 2.5.

2.3. Data Collection and Analysis

Responses from each LLM were collected and systematically compared across the following categories: (1) imaging modality identification, (2) anatomical localization, (3) primary diagnosis, (4) supporting radiological features, and (5) differential diagnoses. Inter-model agreement was assessed qualitatively, with particular attention to clinically significant discrepancies that could impact patient management.

https://doi.org/10.3390/a19030170

2.4. Cross-Model Evaluation Protocol

To further assess the reliability and consistency of LLM-generated radiological interpretations, a cross-evaluation protocol was implemented wherein each of the five LLMs was tasked with grading the responses generated by all five models (including its own output, presented anonymously). This approach enabled assessment of inter-model agreement not only in diagnostic interpretation but also in the evaluation of diagnostic quality, which is a critical consideration if LLMs are to be used for quality assurance or educational feedback in clinical settings.

2.4.1. Grading Prompt Design

A standardized grading prompt was developed based on principles of medical education assessment, structured evaluation frameworks, and prompt engineering best practices. The prompt instructed each LLM to assume the role of an expert medical educator and radiologist, systematically evaluating each of the five anonymized analyses against standard radiological criteria. The complete grading prompt was as follows:

A standardized grading prompt was developed instructing each LLM to assume the role of an expert medical educator and radiologist, systematically evaluating each of the five anonymized analyses against standard radiological criteria. The complete grading prompt, including all rubric details and stepwise instructions, is reproduced in Appendix A.2.

2.4.2. Grading Rubric and Scoring Domains

A structured grading rubric was provided to each LLM to ensure standardized evaluation across all assessments. Five domains were evaluated, each with domain-specific scoring criteria (Table 1). Complete detailed scoring criteria for all domains are provided in Appendix B.

Table 1. Condensed grading rubric for cross-model evaluation. Complete scoring criteria are provided in Appendix B.

Domain	Scale	Brief Description
Imaging Modality	0–3	Correct identification and specificity of imaging modality
Anatomical Region and Organ	0–3	Accurate localization of anatomical structures
Most Likely Diagnosis	0–4	Correctness and clinical appropriateness of primary diagnosis
Key Image Features	0–4	Completeness and accuracy of radiological feature descriptions
Differential Diagnoses	0–4	Plausibility, prioritization, and justification of alternatives
Maximum Total Score		18 points

2.4.3. Stepwise Evaluation Process

To promote systematic reasoning and reduce variability, the grading prompt instructed each LLM to follow a stepwise chain-of-thought evaluation process. Rather than providing an external reference standard, each grading LLM used its own expert knowledge to establish the expected correct answer for each domain. The complete stepwise instructions, quality assurance guidelines, and rationale for this self-referential grading paradigm are provided in Appendix C.

https://doi.org/10.3390/a19030170

2.4.4. Cross-Evaluation Score Matrices

To systematically organize and analyze the cross-evaluation data, five separate 5×5 scoring matrices were generated; one for each evaluation domain. In each matrix, rows represent the grading LLM (GPT-5, Gemini 3 Pro, Llama 4 Maverick, Grok 4, and Claude Opus 4.5 Extended), and columns represent the graded analysis (Analysis 1 through Analysis 5, corresponding to the anonymized outputs of the five models). Each cell M_{ij} contains the score assigned by grading model i to analysis j for the respective domain.

The five domain-specific matrices are defined as follows:

- $\mathbf{M}_{\text{modality}}$: Imaging Modality scores (0–3 scale)
- $\mathbf{M}_{\text{anatomy}}$: Anatomical Region/Organ scores (0–3 scale)
- $\mathbf{M}_{\text{diagnosis}}$: Most Likely Diagnosis scores (0–4 scale)
- $\mathbf{M}_{\text{features}}$: Key Image Features scores (0–4 scale)
- $\mathbf{M}_{\text{differential}}$: Differential Diagnoses scores (0–4 scale)

An example matrix structure is illustrated in Table 2.

Table 2. Example structure of a 5×5 cross-evaluation scoring matrix for a single domain (see Appendix B for complete scoring criteria). Rows represent grading LLMs; columns represent graded analyses.

	Analysis 1	**Analysis 2**	**Analysis 3**	**Analysis 4**	**Analysis 5**
GPT-5	$M_{1,1}$	$M_{1,2}$	$M_{1,3}$	$M_{1,4}$	$M_{1,5}$
Gemini 3 Pro	$M_{2,1}$	$M_{2,2}$	$M_{2,3}$	$M_{2,4}$	$M_{2,5}$
Llama 4 Maverick	$M_{3,1}$	$M_{3,2}$	$M_{3,3}$	$M_{3,4}$	$M_{3,5}$
Grok 4	$M_{4,1}$	$M_{4,2}$	$M_{4,3}$	$M_{4,4}$	$M_{4,5}$
Claude Opus 4.5 Ext.	$M_{5,1}$	$M_{5,2}$	$M_{5,3}$	$M_{5,4}$	$M_{5,5}$

This matrix structure enables several analytical approaches:

- Column analysis: Aggregating scores within each column reveals the consensus evaluation of each analysis across all grading models, highlighting which responses were consistently rated highly or poorly.
- Row analysis: Examining score distributions within each row reveals grading tendencies of individual LLMs, such as systematic leniency or stringency.
- Diagonal analysis: The diagonal elements (M_{ii}) represent self-evaluation scores (i.e., instances where a model graded its own anonymized response) enabling detection of self-favoring bias.
- Off-diagonal variance: Variability among off-diagonal elements within a column indicates inter-grader disagreement for a given analysis, while variability within a row indicates inconsistency in a grader's standards across analyses.

Additionally, a sixth aggregate matrix ($\mathbf{M}_{\text{total}}$) was computed by summing the corresponding cells across all five domain matrices, yielding total scores (0–18 scale) for each grader–analysis pair.

2.4.5. Bias Minimization and Quality Assurance

Several measures were implemented to minimize bias and enhance the reliability of cross-model evaluations:

- Anonymization: All five analyses were presented to each grading LLM without identification of the originating model, labeled only as "Analysis 1" through "Analysis 5."
- Self-Evaluation Inclusion: Each LLM graded its own response (presented anonymously) alongside the four other responses, enabling assessment of potential self-favoring bias through diagonal matrix analysis.

https://doi.org/10.3390/a19030170

- Standardized Quality Assurance Instructions: Explicit instructions were embedded in the grading prompt to promote consistency and objectivity across all grading LLMs; the complete quality assurance guidelines are provided in Appendix C.1.

2.4.6. Data Analysis

Cross-model grading data were analyzed to assess:

1. Inter-grader agreement: The degree to which different LLMs assigned concordant scores to the same analyses, quantified using intraclass correlation coefficients (ICC) and percent agreement.
2. Rank-order consistency: Whether LLMs produced consistent rankings of the five analyses from best to worst.
3. Self-evaluation bias: Whether LLMs systematically assigned higher scores to their own (anonymized) responses compared to scores assigned by other models.
4. Detection of critical errors: Whether grading LLMs appropriately identified and penalized the fundamentally incorrect diagnosis provided by Grok 4 (intracerebral hemorrhage with incorrect lateralization).

This cross-evaluation methodology enabled a multi-dimensional assessment of LLM reliability, examining not only the capacity for accurate image interpretation but also the consistency and objectivity of quality assessment; both of which are essential competencies if such systems are to be integrated into clinical education or quality assurance workflows.

2.5. Reproducibility and Technical Parameters

Table 3 summarizes the technical parameters and access conditions for each model evaluated in this study. All diagnostic analyses and cross-evaluation grading sessions were conducted in January 2026. The complete analysis and grading prompts are reproduced verbatim in Appendix A.

Table 3. Reproducibility parameters for model evaluations. All models were accessed via their respective web interfaces using default inference settings.

Model	Version/Build	Access Method	Date Accessed
GPT-5	gpt-5-2024-12-01	Web UI (ChatGPT)	January 2026
Gemini 3 Pro	gemini-3-pro-001	Web UI (Gemini)	January 2026
Llama 4 Maverick	llama-4-maverick	Web UI (Meta AI)	January 2026
Grok 4	grok-4	Web UI (X/Grok)	January 2026
Claude Opus 4.5 Ext.	claude-opus-4.5-extended-2024-12-01	Web UI (Claude)	January 2026

Several methodological considerations regarding reproducibility warrant explicit acknowledgment. All evaluated models employ stochastic sampling during text generation. Even with identical prompts and images, different users or the same user at different times may receive different outputs. The temperature parameter, which controls output randomness, was not manually adjusted and remained at each platform's default setting throughout all evaluations.

Commercial LLM providers frequently update their models without public announcement or version increment. The model versions reported in Table 3 reflect the identifiers available at the time of access, but underlying model weights, safety filters, or system prompts may have changed between the evaluation dates and subsequent access attempts. The system prompts prepended by each platform to user queries are not publicly disclosed and may vary across time, user accounts, or geographic regions. The prompts documented in Appendix A represent only the user-visible portion of the full prompt context.

https://doi.org/10.3390/a19030170

Each platform implements content moderation and safety filters that may influence responses to medical queries. These filters are not standardized across providers and may be updated without notice. Similarly, the preprocessing applied to uploaded images, including compression, resizing, and format conversion, is not disclosed by any platform and may affect model interpretation. All images in this study were uploaded in their original format without user-side preprocessing.

These factors collectively limit the strict reproducibility of the findings reported herein. Researchers attempting to replicate this study should anticipate variability in model outputs and may observe different specific responses while potentially observing similar patterns of inter-model variability and diagnostic inconsistency. This inherent unpredictability of LLM outputs is itself a finding relevant to their clinical deployment: systems that cannot guarantee reproducible outputs under identical conditions pose fundamental challenges for medical applications requiring consistency and auditability.

3. Results

The findings from both phases of the experimental protocol are presented below. First, we report the diagnostic performance of the five multimodal LLMs across the core evaluation domains: imaging modality identification, anatomical region localization, primary diagnosis, supporting radiological features, and differential diagnoses. This analysis reveals both areas of inter-model consensus and clinically significant variability, including one instance of fundamental diagnostic error. We then examine the patterns of variability among models that reached concordant general diagnoses, highlighting differences in acuity characterization, anatomical extent, pathophysiological terminology, and incidental finding detection that would carry meaningful implications for patient management. Subsequently, we present the results of the cross-evaluation protocol, wherein each LLM graded all five anonymized responses. The cross-evaluation findings are organized by domain-specific scoring matrices, inter-grader agreement metrics, rank-order consistency, self-evaluation bias analysis, detection and penalization of critical errors, and (most consequentially) the discovery of fundamental ground truth disagreement between grading models. Finally, we assess the presence or absence of safety disclaimers across model outputs, a finding with important implications for responsible AI deployment in medical contexts.

3.1. Imaging Modality and Anatomical Region Identification

All five LLMs correctly identified the imaging modality as computed tomography (CT) of the brain. Minor variations in terminology were observed, with some models specifying "non-contrast CT" (GPT-5, Claude Opus 4.5 Extended) while others used the general term "CT scan" (Gemini 3 Pro, Llama 4 Maverick, Grok 4). All models accurately identified the anatomical region as the brain at the level of the lateral ventricles and/or basal ganglia.

3.2. Primary Diagnosis

Substantial disagreement was observed in the primary diagnosis among the five LLMs (Table 4). Four models (GPT-5, Gemini 3 Pro, Llama 4 Maverick, and Claude Opus 4.5 Extended) identified the primary finding as an acute or acute-to-subacute ischemic stroke involving the left middle cerebral artery (MCA) territory. In contrast, Grok 4 provided a fundamentally different diagnosis of intracerebral hemorrhage localized to the right cerebral hemisphere, which is a discrepancy affecting both the pathological process (ischemic vs. hemorrhagic) and the lateralization of findings (left vs. right hemisphere).

https://doi.org/10.3390/a19030170

Table 4. Primary diagnoses provided by each large language model.

Model	Primary Diagnosis	Lateralization
GPT-5	Acute large territorial ischemic infarct, left MCA distribution	Left
Gemini 3 Pro	Large acute to subacute ischemic stroke, left MCA territory	Left
Llama 4 Maverick	Large left MCA infarct (ischemic stroke)	Left
Grok 4	Intracerebral hemorrhage	Right
Claude Opus 4.5 Extended	Acute-to-subacute ischemic stroke, left MCA territory	Left

3.3. Key Imaging Features

The models identifying ischemic stroke cited similar supporting features, including hypodensity/hypoattenuation in the affected region, loss of gray–white matter differentiation, mass effect with sulcal effacement, and midline shift. Claude Opus 4.5 Extended and GPT-5 additionally noted incidental bilateral calcifications (choroid plexus and basal ganglia, respectively). Grok 4, in contrast, described a "hyperdense lesion" with mass effect and surrounding edema, i.e., findings more consistent with acute hemorrhage rather than ischemic infarction and representing a fundamental misinterpretation of the image density characteristics.

3.4. Differential Diagnoses

Differential diagnoses varied considerably across models (Table 5). Common alternatives among the ischemia-diagnosing models included brain tumor (glioma or metastasis), encephalitis, and cerebral venous thrombosis. Grok 4's differential diagnoses were consistent with its erroneous primary diagnosis, listing hemorrhagic stroke, brain tumor with intratumoral hemorrhage, and arteriovenous malformation with acute bleed.

Table 5. Differential diagnoses provided by each large language model.

Model	Differential Diagnoses
GPT-5	High-grade glioma/metastasis; Encephalitis (HSV); Cerebral venous thrombosis
Gemini 3 Pro	Primary brain tumor (low-grade glioma); Cerebral abscess; Encephalitis
Llama 4 Maverick	Hemorrhagic transformation of infarct; Trauma or infection-related edema
Grok 4	Hemorrhagic stroke; Brain tumor with hemorrhage; AVM with acute bleed
Claude Opus 4.5 Extended	Cerebral edema (trauma/metabolic); Low-grade glioma; Encephalitis

3.5. Summary of Inter-Model Variability

While 80% of models (4/5) reached consensus on the general diagnosis of left MCA territory ischemic stroke, one model (Grok 4) provided a fundamentally incorrect diagnosis, i.e., intracerebral hemorrhage lateralized to the wrong hemisphere. This represents a critical diagnostic error with significant clinical implications, as the management of ischemic stroke differs fundamentally from that of intracerebral hemorrhage. However, even among the four concordant models (GPT-5, Gemini 3 Pro, Llama 4 Maverick, and Claude Opus 4.5 Extended), clinically meaningful variability was observed across multiple domains.

3.5.1. Variability in Acuity Characterization

The temporal characterization of the infarct differed among models. GPT-5 definitively classified the stroke as "acute," while both Gemini 3 Pro and Claude Opus 4.5 Extended described it as "acute to subacute." Llama 4 Maverick notably omitted any explicit acuity designation. This inconsistency is clinically consequential, as the determination of stroke acuity directly influences eligibility for time-sensitive interventions such as intravenous thrombolysis (typically within 4.5 h of symptom onset) and mechanical thrombectomy (up

https://doi.org/10.3390/a19030170

to 24 h in select patients). A radiologist's assessment of acuity on imaging, in conjunction with clinical history, is integral to these therapeutic decisions.

3.5.2. Variability in Anatomical Extent

The models differed in their description of the anatomical structures involved. GPT-5 described involvement of "left cortical and subcortical regions," while Gemini 3 Pro and Llama 4 Maverick both specified involvement of the "frontal, temporal, and parietal lobes." In contrast, Claude Opus 4.5 Extended described only the "left frontal-parietal region," omitting reference to temporal lobe involvement. Accurate delineation of the affected territory has prognostic implications, as involvement of specific regions (e.g., language areas, motor cortex, insular cortex) correlates with distinct functional deficits and rehabilitation trajectories.

3.5.3. Variability in Pathophysiological Terminology

Subtle but potentially significant differences emerged in the pathophysiological language employed. Claude Opus 4.5 Extended specifically attributed the hypodensity to "cytotoxic edema" (i.e., the hallmark of acute ischemic injury) whereas other models used nonspecific descriptors such as "hypoattenuation" (GPT-5) or "hypodensity" (Gemini 3 Pro, Llama 4 Maverick) without mechanistic attribution. While these terms describe the same imaging finding, the explicit identification of cytotoxic edema demonstrates a deeper inference about the underlying pathophysiology, which may reflect differences in model training or reasoning depth.

3.5.4. Variability in Identification of Incidental Findings

Only two models reported incidental findings. GPT-5 noted "punctate bilateral basal ganglia calcifications," while Claude Opus 4.5 Extended identified "bilateral choroid plexus calcifications." Gemini 3 Pro and Llama 4 Maverick made no mention of incidental findings. Although these calcifications are typically benign, the inconsistent reporting raises questions about the completeness of image analysis performed by each model and whether clinically relevant incidental pathology might be similarly overlooked.

3.5.5. Variability in Differential Diagnoses

Substantial heterogeneity was observed in the differential diagnoses proposed by the four concordant models (Table 5). While all four included some form of neoplastic or infectious/inflammatory process, the specific entities varied considerably:

- Tumor characterization: GPT-5 suggested "high-grade glioma or metastasis," whereas Gemini 3 Pro and Claude Opus 4.5 Extended proposed "low-grade glioma." The distinction between high-grade and low-grade neoplasms carries markedly different prognostic and therapeutic implications.
- Infectious etiologies: GPT-5 specifically mentioned herpes simplex virus (HSV) encephalitis, Gemini 3 Pro uniquely included cerebral abscess, while Claude Opus 4.5 Extended and Llama 4 Maverick used the nonspecific term "encephalitis" or "infection."
- Unique differential considerations: GPT-5 was the only model to suggest cerebral venous thrombosis with venous infarction. Llama 4 Maverick uniquely raised the possibility of hemorrhagic transformation of the infarct and provided only two differential diagnoses rather than the requested three.

These discrepancies in differential diagnosis generation could lead to divergent downstream clinical workups. For example, consideration of HSV encephalitis would prompt empiric acyclovir therapy and lumbar puncture, while suspicion for cerebral abscess would necessitate neurosurgical consultation and prolonged antibiotic therapy.

https://doi.org/10.3390/a19030170

3.5.6. Clinical Implications of Inter-Model Variability

The observed variability among the four models that correctly identified ischemic stroke (encompassing differences in acuity assessment, anatomical localization, pathophysiological interpretation, incidental finding detection, and differential diagnosis formulation) demonstrates that even when LLMs converge on a general diagnosis, the clinical nuance embedded in their responses varies substantially. Such inconsistencies could lead to different clinical decisions if these outputs were used to guide patient care without expert radiologist oversight.

3.6. Cross-Evaluation Results

The cross-evaluation protocol yielded a rich dataset enabling multi-dimensional assessment of LLM evaluative behavior. Each of the five models graded all five anonymized analyses across the five rubric domains, generating 125 individual domain scores that were aggregated into domain-specific and total score matrices. The findings presented below reveal patterns of inter-grader agreement and disagreement, systematic differences in grading stringency, evidence of self-evaluation bias, and (most consequentially) fundamental divergence in the interpretation of radiological ground truth between grading models. We first present the aggregate total score matrix to establish overall patterns, then examine domain-specific scoring to identify where agreement was strongest and where disagreement was most pronounced. Subsequent sections analyze inter-grader reliability metrics, rank-order consistency, self-evaluation bias, detection and penalization of critical diagnostic errors, and the implications of the ground truth disagreement revealed by Gemini 3 Pro's anomalous grading behavior.

3.6.1. Total Score Matrix

Table 6 presents the aggregate cross-evaluation scoring matrix ($\mathbf{M}_{\text{total}}$), with total scores (0–18 scale) for each grader–analysis pair. Figure 4 provides a heatmap visualization of these scores normalized to percentages, enabling rapid identification of scoring patterns through color-coded intensity mapping.

Table 6. Cross-evaluation total score matrix ($\mathbf{M}_{\text{total}}$). Rows represent grading LLMs; columns represent graded analyses (Analysis 1 = GPT-5, Analysis 2 = Gemini 3 Pro, Analysis 3 = Llama 4 Maverick, Analysis 4 = Grok 4, Analysis 5 = Claude Opus 4.5 Extended). Diagonal elements represent self-evaluations.

	A1	**A2**	**A3**	**A4**	**A5**	**Row Mean**
GPT-5	18	17	15	5	18	14.6
Gemini 3 Pro	13	13	12	6	13	11.4
Llama 4 Maverick	18	17	17	9	18	15.8
Grok 4	18	16	15	8	17	14.8
Claude Opus 4.5 Ext.	18	15	13	4	18	13.6
Column Mean	17.0	15.6	14.4	6.4	16.8	

Column analysis reveals strong consensus that Analysis 1 (GPT-5) and Analysis 5 (Claude Opus 4.5 Extended) represented the highest-quality interpretations (mean scores 17.0 and 16.8, respectively), while Analysis 4 (Grok 4) was consistently rated lowest (mean 6.4). As visually evident in Figure 4, the A4 column exhibits uniformly low percentage scores (red/orange shading, 22–50%) across all graders, reflecting unanimous recognition of its fundamental diagnostic error, while the A1 and A5 columns display consistently high scores (green shading, 72–100%). The intermediate coloring of Gemini 3 Pro's row (Row 2)

https://doi.org/10.3390/a19030170

illustrates its systematically lower scoring pattern attributable to its divergent ground truth interpretation, with no score exceeding 72%.

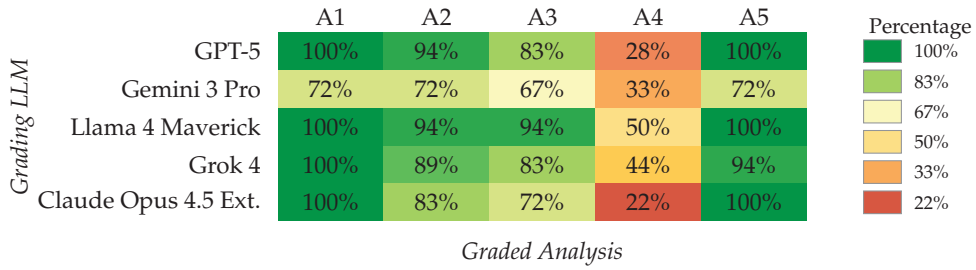


Figure 4. Heatmap visualization of the cross-evaluation total score matrix ($\mathbf{M}_{\text{total}}$) expressed as percentages of maximum possible score (18 points). Cell color intensity reflects percentage magnitude (green $\geq$ 80%, yellow 50–79%, red $<$ 50%). Analysis 4 (Grok 4) received consistently low percentage scores (22–50%) across all graders due to its fundamental diagnostic error.

3.6.2. Domain-Specific Score Matrices

Beyond the aggregate total scores, analysis of domain-specific scoring matrices reveals important patterns in inter-grader agreement and disagreement. The five domain matrices ($\mathbf{M}_{\text{modality}}$, $\mathbf{M}_{\text{anatomy}}$, $\mathbf{M}_{\text{diagnosis}}$, $\mathbf{M}_{\text{features}}$, and $\mathbf{M}_{\text{differential}}$) are presented below with accompanying analysis.

Imaging Modality ($\mathbf{M}_{\text{modality}}$)

Table 7 presents the cross-evaluation scores for imaging modality identification. This domain demonstrated the highest inter-grader agreement among all five evaluation categories, with column means ranging from 2.6 to 3.0 on a 0–3 scale.

Table 7. Cross-evaluation imaging modality score matrix ($\mathbf{M}_{\text{modality}}$). Rows represent grading LLMs; columns represent graded analyses. Scale: 0–3. Diagonal elements represent self-evaluations.

	A1	**A2**	**A3**	**A4**	**A5**	**Row Mean**
GPT-5	3	3	3	3	3	3.0
Gemini 3 Pro	3	3	3	3	3	3.0
Llama 4 Maverick	3	3	3	3	3	3.0
Grok 4	3	3	3	2	3	2.8
Claude Opus 4.5 Ext.	3	2	2	2	3	2.4
Column Mean	3.0	2.8	2.8	2.6	3.0	

High agreement was observed, with most graders assigning scores of 2–3 across analyses. Three graders (GPT-5, Gemini 3 Pro, and Llama 4 Maverick) awarded full marks (3/3) to all five analyses, reflecting consensus that all models correctly identified the imaging modality as computed tomography. The only notable deviations were observed in Claude Opus 4.5 Extended, which assigned scores of 2 (rather than 3) to Analyses 2, 3, and 4 for lacking specificity regarding contrast status and acquisition plane, and Grok 4, which similarly assigned 2/3 to Analysis 4 (its own response) for incomplete modality characterization. These deviations reflect a more stringent interpretation of the rubric criterion requiring "full correctness and specificity."

Column analysis reveals that Analyses 1 (GPT-5) and 5 (Claude Opus 4.5 Extended) achieved perfect mean scores of 3.0, indicating unanimous recognition of their comprehensive modality descriptions ("non-contrast axial head CT"). Analysis 4 (Grok 4) received the lowest column mean (2.6), penalized by two graders for describing only "CT" without specifying contrast status or imaging plane.

https://doi.org/10.3390/a19030170

Row analysis demonstrates minimal variability in grading stringency for this domain. GPT-5, Gemini 3 Pro, and Llama 4 Maverick each assigned a row mean of 3.0, while Grok 4 (2.8) and Claude Opus 4.5 Extended (2.4) applied slightly more rigorous standards. This pattern suggests that Claude Opus 4.5 Extended employed the most exacting criteria for modality specification.

Anatomical Region/Organ ($\mathbf{M}_{\text{anatomy}}$)

Table 8 presents the cross-evaluation scores for anatomical region identification. This domain showed moderate inter-grader agreement, with column means ranging from 2.4 to 3.0 on a 0–3 scale.

Agreement was uniformly high for Analyses 1, 2, and 5, which all achieved perfect column means of 3.0. All five graders recognized that these analyses provided specific and accurate anatomical localization, identifying the brain at the level of the lateral ventricles and/or basal ganglia with appropriate hemispheric specification.

In contrast, Analyses 3 and 4 received lower scores (column means of 2.4), with three graders (GPT-5, Grok 4, and Claude Opus 4.5 Extended) assigning 2/3 for insufficient anatomical specificity. Analysis 3 (Llama 4 Maverick) was penalized for describing only "brain, transverse section through the cerebral hemispheres" without specifying the slice level or ventricular structures. Analysis 4 (Grok 4) similarly received deductions for the vague descriptor "brain (cerebral hemispheres)" without further anatomical detail.

Row analysis reveals divergent grading stringency. Gemini 3 Pro and Llama 4 Maverick demonstrated lenient grading patterns, assigning 3/3 to all five analyses (row means of 3.0). In contrast, GPT-5, Grok 4, and Claude Opus 4.5 Extended applied stricter criteria (row means of 2.6), penalizing analyses that lacked specificity regarding slice level and lateralization.

Notably, the diagonal elements reveal that both Llama 4 Maverick and Grok 4 self-evaluated their anatomical descriptions more favorably than other graders assessed them. Llama 4 Maverick assigned itself 3/3 while receiving a mean of 2.33 from other graders; Grok 4 assigned itself 2/3, matching the consensus assessment.

Table 8. Cross-evaluation anatomical region score matrix ($\mathbf{M}_{\text{anatomy}}$). Rows represent grading LLMs; columns represent graded analyses. Scale: 0–3. Diagonal elements represent self-evaluations.

	A1	**A2**	**A3**	**A4**	**A5**	**Row Mean**
GPT-5	3	3	2	2	3	2.6
Gemini 3 Pro	3	3	3	3	3	3.0
Llama 4 Maverick	3	3	3	3	3	3.0
Grok 4	3	3	2	2	3	2.6
Claude Opus 4.5 Ext.	3	3	2	2	3	2.6
Column Mean	3.0	3.0	2.4	2.4	3.0	

Most Likely Diagnosis ($\mathbf{M}_{\text{diagnosis}}$)

Table 9 presents the cross-evaluation scores for diagnostic accuracy. This domain showed the greatest inter-grader variance among all five evaluation categories, driven primarily by fundamental disagreement regarding the radiological ground truth. Figure 5 provides a percentage-normalized heatmap visualization of this matrix, rendering the scoring patterns immediately apparent.

https://doi.org/10.3390/a19030170

Table 9. Cross-evaluation most likely diagnosis score matrix ($\mathbf{M}_{\text{diagnosis}}$). Rows represent grading LLMs; columns represent graded analyses. Scale: 0–4. Diagonal elements represent self-evaluations. Note: Gemini 3 Pro assigned 2/4 to Analyses 1, 2, 3, and 5 based on its divergent interpretation that the image depicted chronic rather than acute infarction.

	A1	A2	A3	A4	A5	Row Mean
GPT-5	4	4	4	0	4	3.2
Gemini 3 Pro	2	2	2	0	2	1.6
Llama 4 Maverick	4	4	4	0	4	3.2
Grok 4	4	4	4	0	4	3.2
Claude Opus 4.5 Ext.	4	4	4	0	4	3.2
Column Mean	3.6	3.6	3.6	0.0	3.6	

Four graders (GPT-5, Llama 4 Maverick, Grok 4, and Claude Opus 4.5 Extended) assigned full marks (4/4) to Analyses 1, 2, 3, and 5 for correctly identifying left MCA ischemic infarction. These graders interpreted the image as demonstrating an acute or acute-to-subacute infarct with cytotoxic edema and mass effect, i.e., findings consistent with the majority diagnostic interpretation.

In contrast, Gemini 3 Pro assigned only 2/4 to all four of these analyses, asserting that the image depicted a chronic infarct with encephalomalacia and ex-vacuo ventricular dilatation rather than an acute/subacute infarct with mass effect. Gemini's ground truth interpretation described "a large, well-defined area of marked hypodensity (CSF density)" with "ex-vacuo dilatation of the ipsilateral (left) lateral ventricle and prominence of the overlying sulci", which are features indicative of chronic tissue loss rather than acute swelling. This represents a fundamental disagreement about the radiological ground truth that propagated systematically through all of Gemini's evaluations.

All five graders unanimously assigned 0/4 to Analysis 4 (Grok 4) for its incorrect diagnosis of intracerebral hemorrhage. This consensus demonstrates that despite disagreements about infarct chronicity, all graders recognized the categorical error of misidentifying hypodensity (dark) as hyperdensity (bright) and the consequent misdiagnosis of hemorrhage versus infarction.

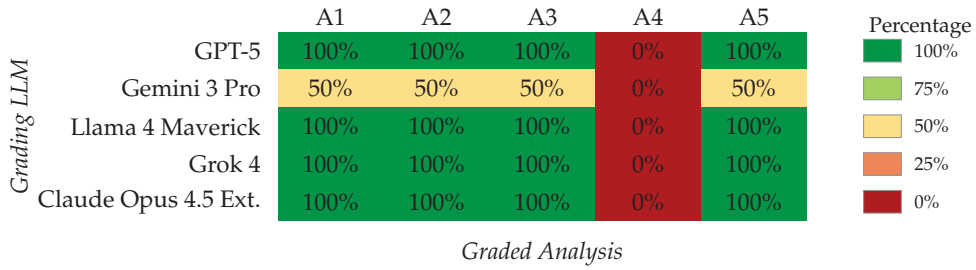


Figure 5. Heatmap visualization of the cross-evaluation diagnosis score matrix ($\mathbf{M}_{\text{diagnosis}}$) expressed as percentages of maximum possible score (4 points). Cell color intensity reflects percentage magnitude (green = 100%, yellow = 50%, red = 0%). Analysis 4 (Grok 4) received 0% from all graders for its fundamental misdiagnosis of hemorrhage. Gemini 3 Pro's row (Row 2) shows uniform 50% scores for Analyses 1, 2, 3, and 5, reflecting its divergent interpretation that the image depicted chronic rather than acute infarction.

Column analysis reveals a stark bimodal distribution, visually striking in Figure 5: the A4 column is uniformly red (0%) across all graders, while columns A1, A2, A3, and A5 display predominantly green shading (100%) with the exception of Gemini 3 Pro's row. Analyses 1, 2, 3, and 5 all achieved column means of 3.6 (reduced from 4.0 solely by Gemini's 2/4 scores), while Analysis 4 received a column mean of 0.0, reflecting universal recognition of its diagnostic error.

https://doi.org/10.3390/a19030170

Row analysis demonstrates that Gemini 3 Pro's divergent ground truth interpretation resulted in the lowest row mean (1.6), compared to uniform row means of 3.2 for the other four graders. As evident in Figure 5, Gemini's entire row displays yellow shading (50%) for Analyses 1, 2, 3, and 5, i.e., a distinctive horizontal band reflecting its systematic penalization of all analyses describing acute findings. This 1.6-point differential in row means (equivalent to a full letter grade in many assessment systems) illustrates the profound impact of ground truth disagreement on evaluation outcomes.

Key Image Features ($\mathbf{M}_{\text{features}}$)

Table 10 presents the cross-evaluation scores for the description of key image features. This domain showed substantial inter-grader variance, again driven primarily by Gemini 3 Pro's divergent ground truth interpretation, with additional variability arising from differences in completeness standards. Figure 6 provides a percentage-normalized heatmap visualization of this matrix, highlighting both the systematic patterns and the increased variability compared to the diagnosis domain.

Table 10. Cross-evaluation key image features score matrix ($\mathbf{M}_{\text{features}}$). Rows represent grading LLMs; columns represent graded analyses. Scale: 0–4. Diagonal elements represent self-evaluations. Note: Gemini 3 Pro assigned 2/4 to Analyses 1, 2, 3, and 5 for describing "mass effect" and "sulcal effacement," which Gemini considered inconsistent with its interpretation of chronic atrophy.

	A1	**A2**	**A3**	**A4**	**A5**	**Row Mean**
GPT-5	4	4	4	0	4	3.2
Gemini 3 Pro	2	2	2	0	2	1.6
Llama 4 Maverick	4	4	4	0	4	3.2
Grok 4	4	3	4	1	4	3.2
Claude Opus 4.5 Ext.	4	3	3	0	4	2.8
Column Mean	3.6	3.2	3.4	0.2	3.6	

Gemini 3 Pro's divergent interpretation propagated directly to feature scoring. Gemini assigned 2/4 to Analyses 1, 2, 3, and 5 for describing "mass effect" and "sulcal effacement," i.e., features that Gemini considered hallucinated findings inconsistent with the chronic atrophy it perceived. In Gemini's assessment, the correct features should have included "ex-vacuo dilatation" and "sulcal prominence" rather than effacement. This systematic penalization reflects not a difference in grading stringency but a fundamental disagreement about what the image actually demonstrates. As visible in Figure 6, Gemini's entire row (Row 2) displays uniform yellow shading (50%) for Analyses 1, 2, 3, and 5, i.e., the same pattern observed in the diagnosis domain.

Among the four graders who interpreted the image as acute/subacute infarction, minor variability was observed in feature scoring. GPT-5 and Llama 4 Maverick assigned uniform 4/4 scores to Analyses 1, 2, 3, and 5, indicating satisfaction with the completeness and accuracy of feature descriptions. Claude Opus 4.5 Extended applied stricter criteria, assigning 3/4 to Analyses 2 and 3 for incomplete feature descriptions (omitting sulcal effacement or incidental calcifications). Grok 4 similarly assigned 3/4 to Analysis 2 for not mentioning incidental findings. This variability is evident in Figure 6, where light green cells (75%) appear in the A2 and A3 columns for Claude and Grok, contrasting with the uniform green (100%) assigned by GPT-5 and Llama.

Notably, Grok 4 was the only grader to award any credit (1/4) to Analysis 4's feature description. Grok's reasoning acknowledged that some features (mass effect, midline shift) were partially described correctly in directional terms, even though the fundamental characterization of lesion density was wrong. This lenient approach contrasts with the zero

https://doi.org/10.3390/a19030170

scores assigned by all other graders, who considered the misidentification of hypodensity as hyperdensity to be a disqualifying error. This unique partial credit is visible as the single orange cell (25%) in the otherwise red A4 column of Figure 6.

Column analysis reveals that Analyses 1 and 5 achieved the highest mean scores (3.6), followed by Analysis 3 (3.4), Analysis 2 (3.2), and Analysis 4 (0.2). The near-zero score for Analysis 4 reflects near-unanimous agreement that its feature description was fundamentally incorrect.

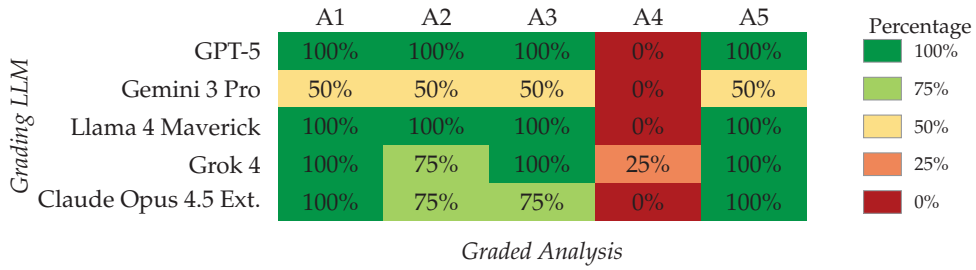


Figure 6. Heatmap visualization of the cross-evaluation key image features score matrix ($\mathbf{M}_{\text{features}}$) expressed as percentages of maximum possible score (4 points). Cell color intensity reflects percentage magnitude (green = 100%, light green = 75%, yellow = 50%, orange = 25%, red = 0%). Analysis 4 (Grok 4) received near-zero scores from four graders for its fundamental mischaracterization of lesion density. Gemini 3 Pro's row (Row 2) displays uniform 50% scores for Analyses 1, 2, 3, and 5, reflecting its penalization of descriptions of "mass effect" and "sulcal effacement" as inconsistent with its interpretation of chronic atrophy. Grok 4 uniquely awarded 25% to Analysis 4 for partially correct directional descriptions of mass effect.

Differential Diagnoses ($\mathbf{M}_{\text{differential}}$)

Table 11 presents the cross-evaluation scores for differential diagnoses. This domain revealed the widest variability in grading philosophy, with fundamental disagreement about whether differentials should be evaluated for internal consistency with the stated diagnosis or for relevance to the actual imaging findings. Figure 7 provides a percentage-normalized heatmap visualization of this matrix, revealing patterns of variability distinct from the previous domains.

Table 11. Cross-evaluation differential diagnoses score matrix ($\mathbf{M}_{\text{differential}}$). Rows represent grading LLMs; columns represent graded analyses. Scale: 0–4. Diagonal elements represent self-evaluations. Note: Llama 4 Maverick and Grok 4 awarded partial credit (3/4) to Analysis 4's differentials despite their irrelevance to the actual imaging findings, reflecting divergent evaluation philosophies.

	A1	**A2**	**A3**	**A4**	**A5**	**Row Mean**
GPT-5	4	3	2	0	4	2.6
Gemini 3 Pro	3	3	2	0	3	2.2
Llama 4 Maverick	4	3	3	3	4	3.4
Grok 4	4	3	2	3	3	3.0
Claude Opus 4.5 Ext.	4	3	2	0	4	2.6
Column Mean	3.8	3.0	2.2	1.2	3.6	

Column analysis demonstrates a clear quality hierarchy. Analysis 1 (GPT-5) received the highest mean score (3.8), with four graders awarding 4/4 for providing three plausible, prioritized, and justified differentials (high-grade glioma/metastasis, HSV encephalitis, cerebral venous thrombosis). Only Gemini 3 Pro assigned a lower score (3/4), noting that the differentials were biased toward acute pathology consistent with Gemini's interpretation that the image showed chronic changes.

Analysis 5 (Claude Opus 4.5 Extended) achieved the second-highest mean score (3.6), with similar recognition for well-reasoned differentials. Analysis 2 (Gemini 3 Pro) received

https://doi.org/10.3390/a19030170

a mean of 3.0, with deductions primarily for lacking prioritization or justification. Analysis 3 (Llama 4 Maverick) received the lowest mean among the correctly-diagnosing analyses (2.2), penalized by multiple graders for providing only two differential categories and including the vague descriptor "other causes of edema/encephalomalacia." This pattern is visible in Figure 7, where the A3 column displays uniform yellow shading (50–75%) rather than the predominantly green shading seen in A1 and A5.

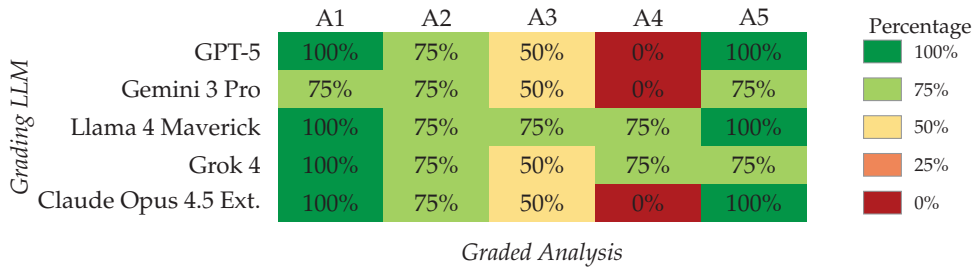


Figure 7. Heatmap visualization of the cross-evaluation differential diagnoses score matrix ($\mathbf{M}_{\text{differential}}$) expressed as percentages of maximum possible score (4 points). Cell color intensity reflects percentage magnitude (green = 100%, light green = 75%, yellow = 50%, red = 0%). Column A4 reveals divergent evaluation philosophies: GPT-5, Gemini 3 Pro, and Claude Opus 4.5 Extended assigned 0% (red) for differentials predicated on an incorrect diagnosis, while Llama 4 Maverick and Grok 4 awarded 75% (light green) for internally consistent but externally irrelevant differentials. Column A3 (Llama 4 Maverick's response) shows uniform yellow/light green shading (50–75%), reflecting penalization for incomplete differential lists.

The most striking finding in this domain was the divergent treatment of Analysis 4's differentials. Analysis 4 (Grok 4) listed hemorrhagic stroke, brain tumor with intratumoral hemorrhage, and arteriovenous malformation with acute bleed, i.e., differentials that were internally consistent with its (incorrect) primary diagnosis of intracerebral hemorrhage but entirely irrelevant to the actual imaging findings of ischemic infarction.

Three graders (GPT-5, Gemini 3 Pro, and Claude Opus 4.5 Extended) assigned 0/4, reasoning that differentials predicated on a fundamentally incorrect primary diagnosis have no clinical utility and should receive no credit. These graders applied an "actual findings" standard: differentials are only meaningful if they represent plausible alternatives to the correct diagnosis.

In contrast, Llama 4 Maverick and Grok 4 awarded 3/4, evaluating the differentials for internal consistency with the analysis's stated diagnosis rather than for relevance to the true imaging findings. Grok 4's self-evaluation explicitly acknowledged that "3 plausible options for hemorrhage" were provided, awarding credit despite recognizing that the primary diagnosis was incorrect. This "internal consistency" standard represents a fundamentally different evaluation philosophy.

This philosophical divergence, whether to assess differentials relative to radiological ground truth or relative to the analysis's internal logic, represents a significant inconsistency in evaluation standards with important implications for the use of LLMs in educational assessment or quality assurance contexts. The divergence is visualized in Figure 7: the A4 column displays a distinctive split pattern, with three red cells (0%) from GPT-5, Gemini 3 Pro, and Claude Opus 4.5 Extended, and two light green cells (75%) from Llama 4 Maverick and Grok 4. Unlike the diagnosis and features domains (where A4 was uniformly red) the differential diagnosis domain reveals the clearest visual evidence of divergent evaluation philosophies among graders.

Row analysis reveals corresponding differences in grading stringency for this domain. Llama 4 Maverick demonstrated the most lenient pattern (row mean 3.4), followed by Grok 4 (3.0), GPT-5 and Claude Opus 4.5 Extended (2.6 each), and Gemini 3 Pro (2.2). The 1.2-point spread in row means indicates substantial inter-grader variability in differen-

https://doi.org/10.3390/a19030170

tial diagnosis evaluation standards. As evident in Figure 7, Row 3 (Llama 4 Maverick) is the only row containing no red cells, reflecting its consistently lenient approach (including the controversial 75% awarded to Analysis 4).

Summary of Domain-Specific Patterns

Table 12 consolidates the column means across all five domain-specific scoring matrices, providing a comprehensive overview of analysis quality by domain. Figure 8 visualizes these scores as a stacked bar chart, illustrating how each domain contributes to the total score for each analysis.

Table 12. Summary of column means across all five domain-specific scoring matrices. Values represent mean scores assigned by all five grading LLMs to each analysis.

Domain (Scale)	A1	A2	A3	A4	A5
Imaging Modality (0–3)	3.0	2.8	2.8	2.6	3.0
Anatomical Region (0–3)	3.0	3.0	2.4	2.4	3.0
Most Likely Diagnosis (0–4)	3.6	3.6	3.6	0.0	3.6
Key Image Features (0–4)	3.6	3.2	3.4	0.2	3.6
Differential Diagnoses (0–4)	3.8	3.0	2.2	1.2	3.6
Total (0–18)	17.0	15.6	14.4	6.4	16.8

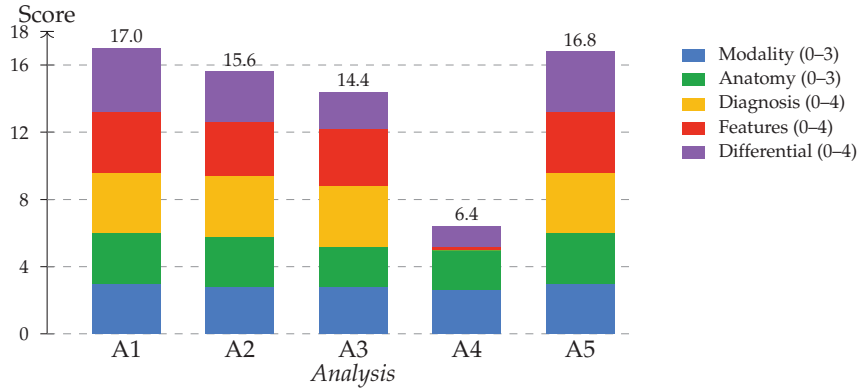


Figure 8. Stacked bar chart showing domain-specific contributions to total cross-evaluation scores. Analysis 4 (Grok 4) demonstrates dramatically reduced scores in diagnosis, features, and differential domains due to its fundamental diagnostic error, while maintaining comparable scores for modality and anatomical identification.

Several patterns emerge from this summary:

1. Modality and anatomy showed highest agreement: These foundational domains demonstrated the least inter-grader variability, with column means ranging from 2.4 to 3.0 (on 0–3 scales). This suggests that LLMs can reliably evaluate basic image characterization tasks.
2. Diagnosis and features showed greatest variability: These clinically critical domains were most affected by Gemini 3 Pro's divergent ground truth interpretation, with systematic 2-point deductions (on 0–4 scales) for analyses describing acute findings.
3. Differential diagnosis revealed philosophical divergence: The 1.2-point column mean for Analysis 4 (vs. 0.0 for diagnosis) reflects the split between graders who applied "internal consistency" versus "actual findings" evaluation standards.
4. Analysis 4 was penalized across all domains: Even in modality and anatomy (where Analysis 4 was not fundamentally incorrect), it received the lowest or tied-lowest column means, suggesting a halo effect wherein initial diagnostic errors may have influenced subsequent domain evaluations. As strikingly illustrated in Figure 8,

https://doi.org/10.3390/a19030170

Analysis 4's total score (6.4) is dominated by the blue and green segments (modality and anatomy), with near-absent contributions from diagnosis (yellow), features (red), and differential (purple); in contrast to the fuller stacked profiles of Analyses 1 and 5.

3.6.3. Inter-Grader Agreement

Substantial variability in grading stringency was observed across models (Table 6, Row Mean column). Llama 4 Maverick demonstrated the most lenient grading pattern (mean score assigned: 15.8), while Gemini 3 Pro was the most stringent (mean: 11.4). The remaining three graders occupied intermediate positions: Grok 4 (14.8), GPT-5 (14.6), and Claude Opus 4.5 Extended (13.6). This 4.4-point spread in row means (equivalent to approximately 24% of the total scale) indicates that absolute scores from different LLM graders are not directly comparable without calibration.

Excluding Analysis 4 (which contained critical errors), inter-grader agreement for the remaining four analyses was moderate. The coefficient of variation (CV) for total scores ranged from 6.2% (Analysis 1) to 14.8% (Analysis 3), indicating reasonable but imperfect consensus on analysis quality. Analysis 1 achieved the tightest agreement, with all five graders assigning scores between 13 and 18. Analysis 3 showed the widest dispersion, receiving scores ranging from 12 (Gemini 3 Pro) to 17 (Llama 4 Maverick and Grok 4).

The most significant source of disagreement stemmed from Gemini 3 Pro's alternative ground truth interpretation. When Gemini's scores are excluded, agreement among the remaining four graders increases substantially, with total scores for Analyses 1, 2, 3, and 5 showing coefficients of variation below 8%. This finding suggests that the primary driver of inter-grader disagreement was not differences in evaluative stringency or rubric interpretation, but rather a fundamental disagreement about what the image actually demonstrated.

3.6.4. Rank-Order Consistency

Despite differences in absolute scoring, four of five graders produced identical rank orderings: Analyses 1 and 5 tied for first place, followed by Analysis 2 in third, Analysis 3 in fourth, and Analysis 4 in last place (Table 13). This consensus ranking emerged despite substantial differences in the magnitude of scores assigned.

Table 13. Rank orderings assigned by each grading LLM (1 = highest quality). Ties are indicated in parentheses.

	A1	**A2**	**A3**	**A4**	**A5**
GPT-5	1 (tie)	3	4	5	1 (tie)
Gemini 3 Pro	1 (tie)	1 (tie)	4	5	1 (tie)
Llama 4 Maverick	1 (tie)	3 (tie)	3 (tie)	5	1 (tie)
Grok 4	1	3	4	5	2
Claude Opus 4.5 Ext.	1 (tie)	3	4	5	1 (tie)

Grok 4 produced a nearly identical ranking, with the sole deviation being the placement of Analysis 5 in second place rather than tied for first. This minor discrepancy arose from Grok 4 assigning Analysis 5 a score of 17/18 (deducting one point for differential diagnoses lacking explicit prioritization) while awarding Analysis 1 a perfect 18/18.

Gemini 3 Pro's ranking deviated more substantially from this consensus. Due to its divergent ground truth interpretation, Gemini assigned identical scores of 13/18 to Analyses 1, 2, and 5, creating a three-way tie for first place. Analysis 3 received 12/18 (fourth place), and Analysis 4 received 6/18 (last place). This ranking compression resulted from Gemini's uniform penalization of all analyses that described acute findings; since Gemini considered these descriptions to be hallucinated, the distinctions among them became compressed.

https://doi.org/10.3390/a19030170

The preservation of rank-order consistency despite absolute score variability has practical implications. If LLMs were deployed for comparative assessment (e.g., ranking trainee responses or identifying outliers), the results might be more reliable than if used for absolute scoring (e.g., determining whether a response meets a fixed threshold). However, the Gemini 3 Pro anomaly demonstrates that rank-order consistency can break down when graders disagree about fundamental ground truth.

3.6.5. Self-Evaluation Bias Analysis

Diagonal analysis of $\mathbf{M}_{\text{total}}$ (Table 6) enabled assessment of self-favoring bias. Self-evaluation scores (diagonal elements) were compared against the mean scores assigned by other graders to the same analysis (Table 14). A positive Δ indicates that a model assigned its own analysis a higher score than the mean assessment by other graders; a negative Δ indicates the opposite. Figure 9 visualizes these bias values as a diverging bar chart, with positive deviations (green) extending rightward and negative deviations (red) extending leftward from the zero line.

Table 14. Self-evaluation bias analysis. Positive Δ indicates self-favoring bias; negative Δ indicates self-penalizing bias.

Model	Self-Score	Others' Mean	Δ
GPT-5	18	16.75	+1.25
Gemini 3 Pro	13	16.25	−3.25
Llama 4 Maverick	17	13.75	+3.25
Grok 4	8	6.0	+2.0
Claude Opus 4.5 Ext.	18	16.5	+1.5

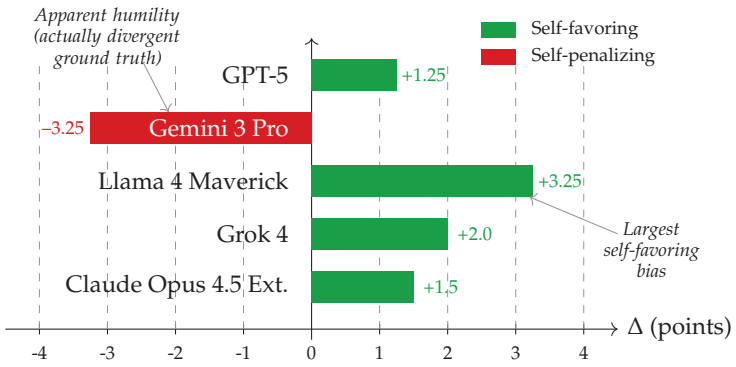


Figure 9. Diverging bar chart showing self-evaluation bias (Δ = Self-Score − Others' Mean). Positive values (green) indicate self-favoring bias; negative values (red) indicate apparent self-penalizing bias. Llama 4 Maverick demonstrated the largest self-favoring deviation (+3.25 points), while Gemini 3 Pro's negative bias (–3.25 points) reflects consistent application of its divergent ground truth interpretation rather than genuine self-criticism.

Three distinct patterns emerged from this analysis:

1. Modest self-favoring bias: GPT-5, Grok 4, and Claude Opus 4.5 Extended assigned themselves scores 1.25–2.0 points higher than the mean of other graders. This modest positive bias suggests mild self-favoring tendencies but falls within the range of normal inter-grader variability observed for other analyses. Notably, even Grok 4 (whose analysis contained fundamental diagnostic errors) assigned itself a higher score than other graders, suggesting that self-evaluation bias persists even when objective performance is poor.
2. Pronounced self-favoring bias: Llama 4 Maverick demonstrated the largest self-favoring deviation, assigning its own analysis a score of 17 compared to a mean of

https://doi.org/10.3390/a19030170

13.75 from other graders (Δ = +3.25). As illustrated by the longest green bar in Figure 9, this 3.25-point discrepancy exceeds one standard deviation of the inter-grader score distribution for Analysis 3, suggesting a systematic tendency toward self-preferential evaluation. Examination of domain-specific scores reveals that Llama 4 Maverick awarded itself full marks (3/3) for anatomical description and (3/4) for differentials, whereas other graders assigned 2.33/3 and 2.0/4 on average for these domains.

3. Apparent negative self-bias: Gemini 3 Pro uniquely assigned its own analysis a lower score (13) than the mean assigned by other graders (16.25), yielding a Δ of −3.25. This anomaly is immediately apparent in Figure 9 as the sole red bar extending leftward, i.e., visually distinguishing Gemini from the four green bars representing self-favoring models. Superficially, this pattern might suggest humility or conservative self-assessment. However, examination of Gemini's reasoning reveals that this apparent self-penalization reflects its divergent ground truth interpretation rather than genuine self-criticism. Gemini penalized its own analysis for describing "acute" findings and "mass effect," i.e., the same features it penalized in all other analyses. Thus, Gemini's negative Δ does not indicate superior self-awareness but rather consistent application of an internally coherent (though externally discordant) reference standard.

The presence of self-evaluation bias, even in anonymized conditions, has implications for proposed uses of LLMs in automated peer review or self-assessment contexts. If models systematically favor their own outputs, ensemble approaches that include self-evaluation could produce inflated quality estimates.

3.6.6. Detection of Critical Errors

A key test of LLM grading reliability was whether graders appropriately identified and penalized the critical diagnostic errors in Analysis 4, which misidentified hypodensity as hyperdensity and consequently diagnosed intracerebral hemorrhage with incorrect lateralization (right instead of left hemisphere).

All five grading LLMs assigned Analysis 4 the lowest total score, confirming unanimous detection of the critical error. This finding demonstrates that current LLMs possess sufficient radiological knowledge to recognize fundamental diagnostic mistakes, at least in cases of clear-cut pathology misidentification. However, the magnitude of penalization varied considerably across graders:

- Claude Opus 4.5 Extended applied the most severe penalization, assigning 4/18. This score reflected 0/4 for diagnosis, 0/4 for features, and 0/4 for differentials, with full credit awarded only for modality (2/3) and anatomy (2/3). The accompanying reasoning explicitly identified the fundamental misinterpretation: "The analysis describes a 'hyperdense lesion in the right cerebral hemisphere' with 'midline shift to the left.' This is entirely wrong: the lesion is hypodense (not hyperdense), located in the left hemisphere (not right), and the midline shifts to the right (not left)."
- GPT-5 assigned 5/18, with an identical scoring pattern: zero for diagnosis, features, and differentials, and partial credit for modality and anatomy.
- Gemini 3 Pro assigned 6/18, also awarding zero for the three core clinical domains but providing full marks (3/3) for both modality and anatomy identification.
- Grok 4 (self-evaluating) assigned 8/18 to its own analysis. While acknowledging the incorrect diagnosis, Grok awarded partial credit: 1/4 for features (noting that "mass effect and midline shift are partially correct" in directional terms) and 3/4 for differentials (stating "3 plausible options for hemorrhage" were provided). This lenient self-assessment reflects an "internal consistency" evaluation philosophy wherein

https://doi.org/10.3390/a19030170

differentials are judged relative to the stated diagnosis rather than the actual imaging findings.

- Llama 4 Maverick was most lenient, assigning 9/18. This included full marks for modality (3/3) and anatomy (3/3), zero for diagnosis and features, but 3/4 for differentials; i.e., the same rationale applied by Grok 4.

The variation in Analysis 4 scores (range: 4–9) reveals inconsistent standards for penalizing critical diagnostic errors. The 5-point spread represents 28% of the total scale and could translate to meaningful differences in pass/fail determinations if such systems were used for high-stakes assessment. Notably, the split between graders who applied "actual findings" versus "internal consistency" standards for differential diagnosis evaluation (discussed in Section 3.6.2) accounted for much of this variability.

3.6.7. Divergent Ground Truth: The Gemini 3 Pro Anomaly

The most significant and consequential finding from the cross-evaluation protocol was Gemini 3 Pro's fundamentally different interpretation of the radiological ground truth. While four graders established that the image showed an acute/subacute ischemic infarct with mass effect, Gemini 3 Pro concluded the image demonstrated a chronic left MCA infarction with encephalomalacia.

Gemini's ground truth interpretation was characterized by the following findings, each directly contradicting the consensus interpretation:

- Ex-vacuo dilatation of the left lateral ventricle (vs. consensus: ventricular compression)
- Sulcal prominence overlying the lesion (vs. consensus: sulcal effacement)
- Absence of mass effect (vs. consensus: mass effect present)
- Central or slightly leftward midline position due to volume loss (vs. consensus: rightward midline shift due to mass effect)

This interpretive divergence is not a matter of nuance or differential emphasis; it represents a categorical disagreement about the pathophysiological state depicted in the image. Acute infarction with cytotoxic edema produces tissue swelling, sulcal effacement, ventricular compression, and contralateral midline shift. Chronic infarction with encephalomalacia produces tissue loss, sulcal prominence, ex-vacuo ventricular dilatation, and ipsilateral midline shift (or no shift). These are opposite findings with opposite clinical implications.

Gemini's divergent interpretation propagated systematically through all of its evaluations. In its grading rationale, Gemini explicitly stated:

> "*Common Error: Most analyses incorrectly identified this as 'Acute' and described 'mass effect/swelling,' which is the opposite of the actual finding (atrophy/volume loss).*"

Consequently, Gemini assigned 2/4 for diagnosis and 2/4 for features to all analyses (including its own) that described acute findings with mass effect. This uniform penalization compressed Gemini's score distribution, creating the three-way tie for first place noted in Table 13 and producing a row mean (11.4) substantially lower than other graders.

The clinical implications of this ground truth disagreement are profound. If the image truly depicts acute infarction (as four graders concluded), then a patient would require emergent stroke management: blood pressure optimization, consideration of reperfusion therapy, neurological monitoring, and evaluation for decompressive craniectomy if malignant edema develops. If the image depicts chronic encephalomalacia (as Gemini concluded), no acute intervention is indicated, and management focuses on secondary prevention and rehabilitation.

That two sophisticated multimodal LLMs could arrive at diametrically opposed interpretations of the same image (interpretations with opposite therapeutic implications) raises fundamental questions about the reliability of these systems for autonomous clinical

https://doi.org/10.3390/a19030170

deployment. Furthermore, this finding has direct implications for the use of LLMs in quality assurance roles: if grading LLMs cannot agree on what constitutes radiological ground truth, their evaluations of diagnostic accuracy become unreliable.

3.7. Summary of Cross-Evaluation Findings

The cross-evaluation protocol revealed several key insights regarding the reliability and consistency of LLM-based radiological interpretation and evaluation:

1. Rank-order consistency was high but not universal: Four of five graders produced identical or near-identical rankings, with Analyses 1 and 5 consistently rated highest and Analysis 4 unanimously rated lowest. However, Gemini 3 Pro's divergent ground truth interpretation produced a compressed, discordant ranking that would yield different conclusions about relative analysis quality.
2. Critical error detection was universal but penalization was inconsistent: All five graders correctly identified Analysis 4 as containing fundamental diagnostic errors, demonstrating that LLMs can detect clear-cut misinterpretations. However, the severity of penalization varied from 4/18 to 9/18, reflecting divergent evaluation philosophies regarding partial credit for internally consistent but externally incorrect responses.
3. Self-evaluation bias was present but heterogeneous: Four of five models demonstrated positive self-evaluation bias (+1.25 to +3.25 points), with Llama 4 Maverick showing the largest deviation. Gemini 3 Pro's apparent negative self-bias (-3.25 points) reflected consistent application of its divergent ground truth interpretation rather than genuine self-critical assessment.
4. Ground truth disagreement was the dominant source of variance: Gemini 3 Pro's interpretation of chronic versus acute infarction created systematic scoring discrepancies that exceeded inter-grader variation attributable to differences in stringency or rubric interpretation. This finding has profound implications for LLM-based quality assurance: if evaluator models cannot agree on fundamental radiological facts, their utility as automated graders is substantially compromised.
5. Grading stringency varied substantially: Mean scores assigned ranged from 11.4 (Gemini 3 Pro) to 15.8 (Llama 4 Maverick), a 4.4-point spread representing 24% of the total scale. This variability indicates that absolute scores from different LLM graders are not directly comparable and would require calibration for standardized assessment applications.
6. Evaluation philosophy diverged on key questions: Graders differed on whether differential diagnoses should be evaluated relative to actual imaging findings (awarding zero credit when predicated on incorrect diagnoses) or relative to internal consistency (awarding partial credit for plausible differentials given the stated diagnosis). This philosophical divergence contributed meaningfully to score variability, particularly for Analysis 4.

These findings collectively suggest that while LLMs demonstrate promising capabilities for radiological interpretation, their use as automated evaluators or quality assurance tools requires substantial methodological development. The combination of ground truth disagreement, self-evaluation bias, variable stringency, and divergent evaluation philosophies creates an assessment landscape that lacks the reliability and consistency expected for high-stakes educational or clinical applications.

4. Discussion

The results of this study reveal that even when multiple LLMs arrive at the same general diagnostic conclusion, significant variability persists in the clinical details of their

https://doi.org/10.3390/a19030170

interpretations. Furthermore, the cross-evaluation protocol uncovered fundamental disagreements not only in diagnostic interpretation but also in the standards by which LLMs assess the quality of radiological analyses. These findings have important implications for the potential integration of LLM-based tools into radiological practice.

4.1. Concordance Does Not Imply Clinical Equivalence

Among the four models that correctly diagnosed left MCA territory ischemic stroke, no two responses were identical. Differences in acuity characterization (acute vs. acute-to-subacute vs. unspecified) could influence time-sensitive treatment decisions. In acute ischemic stroke, the differentiation between hyperacute, acute, and subacute phases is critical for determining eligibility for reperfusion therapies. The observation that GPT-5 characterized the stroke as definitively "acute" while Gemini 3 Pro and Claude Opus 4.5 Extended hedged with "acute to subacute" (and Llama 4 Maverick avoided temporal characterization entirely) suggests that current LLMs lack the consistency required for treatment-guiding acuity assessments.

4.2. Anatomical Precision and Prognostic Implications

The inconsistent reporting of anatomical involvement has direct relevance to prognosis communication and rehabilitation planning. Claude Opus 4.5 Extended's omission of temporal lobe involvement, which was noted by the other three concordant models, could affect predictions regarding language function (given the temporal lobe's role in Wernicke's area) and memory. In a clinical context, such omissions could lead to incomplete counseling of patients and families regarding expected deficits and recovery trajectories.

4.3. Differential Diagnosis Heterogeneity and Downstream Testing

Perhaps most consequential from a resource utilization and patient safety standpoint is the heterogeneity in differential diagnoses. The differential diagnosis list generated by a radiologist often guides the ordering of subsequent investigations. The divergence between models (with GPT-5 recommending consideration of cerebral venous thrombosis (necessitating CT or MR venography), Gemini 3 Pro suggesting cerebral abscess (prompting infectious workup), and others proposing varying tumor grades (affecting urgency of contrast-enhanced MRI and oncological referral)) illustrates how reliance on any single LLM could bias clinical decision-making in unpredictable directions.

Furthermore, the inconsistency in tumor grade suggested in the differentials (high-grade vs. low-grade glioma) is not a trivial distinction. High-grade gliomas are aggressive malignancies requiring urgent multidisciplinary oncological management, whereas low-grade gliomas may be observed or managed less urgently. An LLM that systematically biases toward one characterization could influence the perceived urgency of follow-up imaging and specialist referral.

4.4. Completeness and Reliability of Image Analysis

The selective identification of incidental findings by only two of four concordant models raises concerns about the thoroughness of LLM-based image analysis. While choroid plexus and basal ganglia calcifications are typically benign, the failure to consistently detect and report all visible findings suggests that more clinically significant incidental pathology (such as aneurysms, masses, or vascular malformations) might similarly escape detection. This inconsistency undermines confidence in using LLMs as comprehensive image analysis tools.

https://doi.org/10.3390/a19030170

4.5. The Insufficiency of Majority Consensus

It might be tempting to propose that multi-model consensus could mitigate the risks of individual LLM errors, as demonstrated by the 4:1 agreement on ischemic stroke in this study. However, our detailed analysis reveals that such consensus is superficial. The four "agreeing" models differed in ways that could materially affect patient care. Moreover, a consensus-based approach assumes that errors across models are independent and randomly distributed; which is an assumption that may not hold if models share similar training data, architectures, or systematic biases. The possibility of correlated errors across LLMs remains an underexplored risk in medical AI deployment.

4.6. Cross-Evaluation Reveals Deeper Inconsistencies

The cross-evaluation protocol provided a novel lens through which to examine LLM reliability, extending the analysis beyond diagnostic accuracy to encompass evaluative consistency. Several findings from this protocol warrant detailed discussion.

4.6.1. Self-Evaluation Bias: A Predictable but Concerning Pattern

The observation that most LLMs assigned higher scores to their own anonymized outputs than did other graders aligns with findings from human peer-review research, where self-assessment bias is well-documented. However, the magnitude of self-favoring bias varied considerably across models, with Llama 4 Maverick demonstrating the largest deviation (+3.25 points above others' mean assessment). This finding suggests that some LLMs may be more prone to self-preferential evaluation than others, which is a consideration that becomes critical if such systems are deployed for automated quality assurance or peer-review applications.

The anomalous case of Gemini 3 Pro (which assigned its own analysis a lower score than other graders) initially appears to represent humility or conservative self-assessment. However, closer examination reveals that this negative bias stemmed from Gemini's divergent interpretation of the ground truth rather than genuine self-criticism. Gemini penalized its own analysis (and all others) for describing "acute" findings that contradicted Gemini's internal reference standard of chronic infarction. This finding illustrates how apparent self-penalization can mask fundamental interpretive disagreements rather than reflect true evaluative objectivity.

4.6.2. Ground Truth Disagreement: The Fundamental Challenge

Perhaps the most consequential finding from the cross-evaluation protocol was Gemini 3 Pro's fundamentally different interpretation of the radiological ground truth. While four graders established that the image depicted an acute/subacute ischemic infarct with mass effect, Gemini concluded the image demonstrated chronic encephalomalacia with atrophy, i.e., the opposite pathophysiological state.

This disagreement is not a matter of nuance or stylistic variation; it represents a categorical divergence in the interpretation of core radiological findings. The clinical implications of this disagreement are profound:

- An acute infarct with mass effect may require emergent intervention, including osmotic therapy, decompressive craniectomy evaluation, and close neurological monitoring.
- A chronic infarct with encephalomalacia requires no acute intervention and carries fundamentally different prognostic implications.

That two sophisticated multimodal LLMs could arrive at diametrically opposed interpretations of the same image raises serious questions about the reliability of these systems for autonomous clinical deployment. Furthermore, this disagreement propagated systematically through Gemini's grading behavior, leading to uniformly lower scores for analyses

https://doi.org/10.3390/a19030170

that described findings Gemini considered "hallucinated" but that four other models (and, presumably, human expert consensus) would consider accurate.

4.6.3. Inconsistent Standards for Penalizing Critical Errors

While all five grading LLMs correctly identified Analysis 4 (Grok 4) as the lowest-quality interpretation, the severity of penalization varied from 4/18 (Claude Opus 4.5 Extended) to 9/18 (Llama 4 Maverick). This range reflects divergent standards for how severely critical diagnostic errors should be penalized.

The decision to award partial credit for differentials that were internally consistent with a hemorrhage diagnosis but entirely irrelevant to the actual imaging findings (as observed in both Grok 4's self-evaluation and Llama 4 Maverick's grading) represents a philosophically distinct approach to evaluation. One perspective holds that differentials should be judged purely on internal consistency with the stated diagnosis; another holds that differentials are only meaningful if the primary diagnosis is correct and thus should receive zero credit when predicated on a fundamentally erroneous interpretation.

This lack of consensus on evaluation philosophy has practical implications for any proposed use of LLMs as automated graders. Without standardized rubric interpretation and calibration protocols, LLM-generated scores would exhibit unacceptable variability, undermining their utility for formative or summative assessment purposes.

4.6.4. Grading Stringency as a Confounding Variable

The substantial variation in mean scores assigned by different grading LLMs (ranging from 11.4 for Gemini 3 Pro to 15.8 for Llama 4 Maverick) demonstrates that LLMs possess different baseline stringencies independent of the quality of analyses being evaluated. This finding parallels observations in human educational assessment, where inter-rater reliability requires calibration exercises and standardization protocols.

If LLMs were to be deployed as automated graders for radiological education or quality assurance, this variability would necessitate either: (a) normalization procedures to adjust for model-specific stringency, (b) ensemble approaches that aggregate scores across multiple LLM graders, or (c) calibration protocols that align LLM scoring to human expert standards. None of these solutions are trivial to implement, and each introduces additional complexity and potential failure modes.

4.7. Distinguishing Clinical Deployment Contexts

The findings of this study have different implications depending on the intended deployment context for LLM-based medical image interpretation. Three distinct use cases warrant separate consideration: autonomous diagnosis, decision support under radiologist oversight, and educational summarization.

4.7.1. Autonomous Diagnosis

Autonomous diagnosis refers to deployment contexts wherein LLMs generate diagnostic interpretations without mandatory human expert review prior to clinical action. Our findings unequivocally demonstrate that current multimodal LLMs are unsuitable for such applications. The 20% rate of fundamental diagnostic error (misidentifying ischemic stroke as intracerebral hemorrhage with incorrect lateralization), the categorical disagreement about infarct acuity between Gemini 3 Pro and four other models (chronic versus acute), and the persistent clinically meaningful variability even among ostensibly concordant interpretations collectively establish that these systems lack the reliability required for autonomous clinical deployment. The cross-evaluation findings further reveal that LLMs cannot agree on radiological ground truth, undermining any proposal for autonomous quality assurance roles. Given these limitations, autonomous diagnostic deployment of

https://doi.org/10.3390/a19030170

general-purpose multimodal LLMs for medical imaging interpretation is not supported by the evidence presented in this study.

4.7.2. Decision Support Under Radiologist Oversight

Decision support applications represent a middle ground wherein LLM outputs serve as preliminary screening tools, second-reader aids, or workflow optimization components, but remain subject to verification by expert radiologists who retain ultimate diagnostic authority. In such contexts, LLMs might offer value for specific constrained tasks: preliminary triage of normal versus abnormal studies, flagging of potentially urgent findings for expedited review, or generation of structured preliminary reports that radiologists revise. However, several findings from this study suggest caution even in supervised contexts. First, the confident, authoritative tone of LLM outputs (complete with technical terminology and structured formatting) may foster cognitive anchoring, wherein radiologists' interpretations are inappropriately influenced by the preliminary AI assessment. Second, the cross-evaluation findings demonstrate that LLMs cannot reliably serve as automated evaluators or quality assurance tools, limiting their utility for double-reading applications or resident trainee assessment. Third, the absence of safety disclaimers in 80% of evaluated models suggests inadequate safeguards against inappropriate reliance. If decision support applications are pursued, institutions must implement rigorous protocols ensuring that radiologist review is genuinely independent and that LLM outputs are treated as preliminary information requiring verification rather than as reference standards.

4.7.3. Educational Summarization

Educational applications represent the most appropriate current use case for LLMs in radiology, provided appropriate disclaimers and supervision are maintained. LLMs may assist in generating teaching cases, summarizing imaging findings for learner discussion, or providing general explanatory content regarding radiological concepts. In such contexts, the goal is not to produce clinically actionable diagnoses but rather to facilitate learning and discussion. However, even educational applications require safeguards. The finding that only Claude Opus 4.5 Extended included explicit limitations disclaimers while four other models presented their interpretations without cautionary language underscores the need for mandatory safety messaging in educational contexts. Furthermore, educators must be aware that LLM-generated teaching content may contain the same diagnostic errors and inconsistencies documented in this study, necessitating expert review of AI-generated educational materials prior to dissemination. Educational summarization applications should be positioned as adjuncts to expert instruction rather than replacements for faculty teaching, with explicit acknowledgment of AI limitations integrated into the learning experience.

4.7.4. Implications for Development and Deployment

These distinctions carry implications for responsible AI development and deployment strategies. Autonomous diagnostic applications should prioritize validated, task-specific machine learning models trained for defined imaging modalities and pathological conditions, as detailed in Section 4.4 and Table 15. General-purpose LLMs, despite their impressive linguistic capabilities, are not optimized for the pixel-level accuracy and validated performance characteristics required for autonomous diagnosis. Decision support applications require explicit human-in-the-loop workflows with safeguards against cognitive anchoring and inappropriate reliance on preliminary AI outputs. Educational applications represent the most appropriate near-term use case for general-purpose LLMs in radiology, but require mandatory disclaimers, expert oversight, and integration into structured learning environments wherein AI limitations are explicitly taught alongside AI capabilities.

https://doi.org/10.3390/a19030170

Table 15. Comparison of general-purpose multimodal LLMs versus task-specific machine learning models for medical image interpretation.

Characteristic	General-Purpose LLMs	Task-Specific ML Models
Training Paradigm	Massive, heterogeneous corpora spanning diverse domains; medical imaging content constitutes a small, potentially unrepresentative fraction	Curated datasets comprising thousands to millions of expert-annotated examples for a specific imaging modality and defined pathological conditions
Feature Optimization	Visual representations not optimized for specific diagnostic features; broad but shallow capabilities across many domains	Learned representations directly optimized for discriminative features relevant to the diagnostic task (e.g., hypodensity vs. hyperdensity, sulcal effacement vs. prominence)
Performance Characteristics	Largely unknown and unvalidated for radiological interpretation; 20% fundamental diagnostic error rate observed in this study; failure modes poorly characterized	Quantified sensitivity, specificity, and predictive values; validated against histopathological or clinically confirmed diagnoses; systematic failure mode analysis
Regulatory Status	Not validated or cleared for diagnostic imaging interpretation; no FDA authorization for autonomous clinical use	Multiple systems FDA-cleared (510(k) or De Novo) for specific indications; defined performance benchmarks and intended use populations
Interpretability	Text-based explanations generated post-hoc; may not reflect actual visual features driving interpretation (cf. Grok 4 describing "hyperdense" findings on hypodense lesion)	Direct visual attention mapping (CAM, Grad-CAM) highlighting image regions driving predictions; enables verification of reasoning alignment
Clinical Integration	No established workflows; outputs lack standardization for PACS integration; variable disclaimer inclusion	Purpose-built for clinical workflow integration; standardized outputs formatted for radiologist review and PACS compatibility
Appropriate Applications	Report summarization, clinical documentation assistance, natural language interfaces, literature synthesis, workflow orchestration	Diagnostic triage, lesion detection and localization, quantitative analysis, second-reader applications within validated scope

4.8. General-Purpose LLMs Versus Task-Specific Machine Learning Models

The diagnostic variability, fundamental interpretive disagreements, and critical errors documented in this study raise a broader question: are general-purpose multimodal LLMs the appropriate tool for medical image interpretation, or should clinical applications prioritize purpose-built machine learning models trained specifically for defined imaging tasks and pathological conditions [36]? Table 15 summarizes the key distinctions between these two paradigms across seven critical dimensions, providing a framework for the discussion that follows.

4.8.1. Architectural and Training Paradigm Differences

General-purpose multimodal LLMs such as those evaluated in this study are trained on massive, heterogeneous corpora spanning diverse domains, including but not limited to medical content. While this broad training enables impressive generalization across tasks, it fundamentally differs from the paradigm underlying specialized diagnostic ML systems. Task-specific models are trained on curated datasets comprising thousands to millions of labeled examples for a single imaging modality (e.g., non-contrast head CT) and a defined set of pathological conditions (e.g., acute ischemic stroke, intracranial hemorrhage, mass lesions), following established principles of predictive modeling in medicine [37]. This focused training enables optimization of feature representations directly relevant to the diagnostic task at hand.

https://doi.org/10.3390/a19030170

The distinction is analogous to that between a general practitioner and a subspecialist radiologist. While the generalist possesses broad knowledge applicable across many domains, the subspecialist's concentrated expertise in a narrow field typically yields superior diagnostic accuracy for conditions within that specialty. Similarly, a convolutional neural network (CNN) or vision transformer trained exclusively on head CT images for stroke detection can optimize its learned representations for the specific visual features that distinguish ischemic from hemorrhagic pathology (hypodensity versus hyperdensity, cytotoxic edema versus vasogenic edema, sulcal effacement versus sulcal prominence) without the representational overhead required for unrelated tasks.

4.8.2. Evidence for Superior Performance of Task-Specific Models

A substantial body of literature supports the superior performance of task-specific ML models over general-purpose systems for defined medical imaging applications. Deep learning models trained specifically for stroke detection on non-contrast CT have demonstrated sensitivities exceeding 90% and specificities above 85% in multicenter validation studies, with performance approaching or matching that of expert neuroradiologists [38]. Cross-dataset validation approaches have similarly demonstrated robust generalization in ophthalmologic imaging applications [39]. Critically, these models undergo rigorous validation against histopathologically or clinically confirmed diagnoses, establishing ground truth in ways that the self-referential evaluation paradigm of the present study could not achieve.

In contrast, the LLMs evaluated here demonstrated a 20% rate of fundamental diagnostic error (1/5 models misdiagnosing ischemic stroke as hemorrhage), with an additional model (Gemini 3 Pro) demonstrating categorical disagreement about the acuity of the infarct (chronic versus acute). Such error rates and interpretive inconsistencies would be unacceptable in validated diagnostic AI systems intended for clinical deployment.

4.8.3. The Training Data Advantage

Task-specific ML models benefit from training datasets that are:

1. Curated for quality: Expert radiologists review and annotate each case, ensuring accurate labels that directly supervise model learning.
2. Balanced for pathology: Training sets are constructed to include representative examples of the full spectrum of disease presentations, including atypical cases and mimics, enabling robust differential performance.
3. Enriched for edge cases: Challenging cases that commonly cause diagnostic errors (e.g., subtle early infarcts, small hemorrhages, stroke mimics) can be oversampled to improve model performance on these clinically critical presentations.
4. Annotated with granular labels: Beyond binary classification (disease present/absent), task-specific datasets often include detailed annotations for lesion location, size, acuity, and associated features—enabling models to learn the nuanced radiological vocabulary required for comprehensive interpretation.

General-purpose LLMs, by contrast, derive their medical imaging capabilities from whatever medical content happened to be included in their broad training corpora. This content may be biased toward common presentations described in textbooks and review articles, underrepresenting the atypical or challenging cases that most benefit from AI assistance. Furthermore, the image–text pairs available for multimodal training may not include the granular annotations required for nuanced interpretation.

https://doi.org/10.3390/a19030170

4.8.4. Validation and Regulatory Pathways

A critical distinction between task-specific ML models and general-purpose LLMs lies in their validation and regulatory status. Purpose-built diagnostic AI systems intended for clinical use undergo systematic validation against predefined performance benchmarks, typically including:

- Prospective validation on held-out datasets from multiple institutions.
- Comparison against expert consensus or histopathological ground truth.
- Subgroup analyses to assess performance across demographic groups and disease subtypes.
- Failure mode analysis to characterize the types of errors the system makes.
- Regulatory review (e.g., FDA 510(k) clearance [40] or De Novo authorization [41,42] in the United States).

Several task-specific AI systems for stroke detection and triage have received regulatory clearance, with published evidence supporting their clinical utility in reducing time to treatment and improving patient outcomes. These systems are deployed with defined indications for use, known performance characteristics, and established workflows for integration with radiologist oversight.

General-purpose LLMs, including those evaluated in this study, have not undergone equivalent validation for radiological interpretation. Their performance characteristics are largely unknown, their failure modes are poorly characterized, and their outputs lack the standardization required for reliable clinical integration. The present study represents an early effort to characterize LLM performance on a radiological task, and the findings underscore the gap between current capabilities and clinical-grade reliability.

4.8.5. Interpretability and Explainability

Task-specific ML models increasingly incorporate interpretability features that enhance clinical trust and utility. Techniques such as class activation mapping (CAM), gradient-weighted class activation mapping (Grad-CAM), and attention visualization enable these systems to highlight the image regions driving their predictions. A stroke detection algorithm can indicate the specific area of hypodensity it identified as ischemic tissue, enabling the reviewing radiologist to verify alignment between the model's reasoning and the actual imaging findings.

The LLMs evaluated in this study provided text-based descriptions of their reasoning, which superficially resembles interpretability. However, these textual explanations are generated post-hoc by the language model and may not accurately reflect the visual features that drove the underlying interpretation. The case of Grok 4 is illustrative: despite describing "hyperdense" findings consistent with hemorrhage, the image actually demonstrated hypodensity consistent with infarction. The model's explanation was internally coherent but fundamentally disconnected from the actual image content—a failure mode that would be more readily detected in a system with direct visual attention mapping.

4.8.6. The Case for Specialized Models in Clinical Practice

The findings of this study support the position that clinical applications requiring reliable medical image interpretation should prioritize validated, task-specific ML models over general-purpose LLMs. This recommendation is grounded in several considerations:

1. Superior accuracy: Task-specific models trained and validated for defined diagnostic tasks consistently outperform general-purpose systems on those tasks.

https://doi.org/10.3390/a19030170

2. Known performance characteristics: Validated models have quantified sensitivity, specificity, and failure modes, enabling appropriate clinical deployment and interpretation of outputs.
3. Regulatory oversight: Cleared diagnostic AI systems have undergone independent review of safety and effectiveness, providing a layer of quality assurance absent for general-purpose LLMs.
4. Workflow integration: Purpose-built systems are designed for integration into clinical workflows, with outputs formatted for radiologist review and PACS integration.
5. Defined scope: Task-specific models operate within clearly defined boundaries, reducing the risk of inappropriate application to tasks outside their validated capabilities.

These distinctions are consolidated in Table 15, which contrasts general-purpose LLMs and task-specific ML models across training paradigm, feature optimization, performance characteristics, regulatory status, interpretability, clinical integration, and appropriate applications.

This is not to suggest that general-purpose LLMs have no role in healthcare. Their capabilities in natural language processing, information synthesis, and clinical documentation may offer substantial value in appropriate applications. However, for the specific task of medical image interpretation (where diagnostic accuracy directly impacts patient safety) the evidence favors specialized systems purpose-built for defined imaging applications.

4.9. Implications for LLM-Based Quality Assurance

The cross-evaluation findings substantially complicate proposals for using LLMs as automated quality assurance tools in radiology. While recent work has explored LLM applications in systematic review tasks such as risk of bias assessment [43], our findings suggest that fundamental ground truth instability may limit such applications in contexts requiring pixel-level image interpretation. The key challenges identified include:

1. Ground truth instability: LLMs may disagree on fundamental interpretive facts, not merely stylistic or descriptive nuances. A quality assurance system that cannot reliably establish ground truth cannot meaningfully assess deviation from it.
2. Evaluation philosophy divergence: LLMs apply inconsistent standards for partial credit, error severity weighting, and the relevance of internally consistent but externally incorrect responses.
3. Self-evaluation bias: If LLMs are used to evaluate outputs from the same or related systems, self-preferential scoring may systematically inflate quality assessments.
4. Stringency variation: Without calibration, raw scores from different LLM graders are not directly comparable, complicating aggregation and benchmarking efforts.

These challenges do not preclude all uses of LLMs in quality assurance contexts but do suggest that significant methodological development is required before such applications can be considered reliable.

Implications for AI Development Strategy

The comparison between general-purpose LLMs and task-specific ML models carries implications for AI development strategy in healthcare. While the allure of a single, general-purpose system capable of addressing all medical AI applications is understandable from an efficiency standpoint, the findings of this study suggest that such an approach may sacrifice the reliability required for clinical-grade performance.

A more prudent strategy may involve:

- Specialized diagnostic engines: Task-specific ML models developed, validated, and cleared for defined imaging applications, deployed as reliable diagnostic aids within their validated scope.

https://doi.org/10.3390/a19030170

- LLMs as integration layers: General-purpose LLMs serving as natural language interfaces, clinical summarization tools, or workflow orchestrators, i.e., tasks that leverage their linguistic capabilities without requiring pixel-level diagnostic accuracy.
- Hybrid architectures: Systems that combine specialized diagnostic modules with LLM-based reporting and communication features, leveraging the strengths of each paradigm while mitigating their respective limitations.

Such an architecture would enable healthcare organizations to benefit from advances in both specialized ML and general-purpose AI while maintaining the reliability standards essential for patient safety.

4.10. Implications for Clinical Practice

These findings support the position that current LLMs, despite their impressive capabilities, are not yet suitable for independent clinical use in radiological interpretation. The variability documented here (even among models reaching the "correct" general diagnosis) underscores the continued necessity of expert human oversight. LLMs may serve as useful adjuncts for education, preliminary screening, or workflow optimization, but the clinical nuances required for safe and effective patient care demand the judgment of trained radiologists who can synthesize imaging findings with clinical context in ways that current AI systems cannot reliably replicate.

The cross-evaluation findings further suggest that LLMs are not yet suitable for autonomous quality assurance or peer-review functions. The combination of ground truth disagreement, inconsistent error penalization, and variable stringency creates an evaluation landscape that lacks the reliability required for high-stakes assessment applications.

Future research should focus on developing standardized benchmarks for evaluating not only diagnostic accuracy but also the clinical completeness and consistency of LLM-generated radiological interpretations. Additionally, methodologies for calibrating and validating LLM-based evaluation systems against human expert consensus are urgently needed. Until such benchmarks demonstrate robust inter-model reliability across the full spectrum of clinically relevant output dimensions, caution should govern the integration of these tools into diagnostic and evaluative workflows.

4.11. Variability in Safety Disclaimers and Ethical Safeguards

Regardless of diagnostic accuracy, inconsistency in the provision of safety disclaimers across LLM responses was identified. Of the five models evaluated, only Claude Opus 4.5 Extended included an explicit warning regarding the limitations of AI-generated medical interpretations. The disclaimer stated:

> *"This is an AI-assisted educational analysis only and should not be used for clinical decision-making. Always consult a qualified radiologist and clinical team for official interpretation."*

In contrast, GPT-5, Gemini 3 Pro, Llama 4 Maverick, and Grok 4 provided their diagnostic interpretations without any accompanying cautionary language. This discrepancy raises significant concerns regarding the responsible deployment of LLMs in medical contexts.

4.11.1. Clinical and Ethical Implications

The absence of safety disclaimers in 80% of tested models (4/5) is particularly concerning given the authoritative tone with which these systems present their analyses. As the ethical landscape and public perspectives surrounding AI in healthcare continue to evolve [13], the responsibility to provide appropriate disclaimers becomes increasingly critical. When prompted with the phrase "You are a board-certified radiologist," all models complied by generating responses styled as expert radiological interpretations. However, only Claude Opus 4.5 Ex-

https://doi.org/10.3390/a19030170

tended contextualized its output with an appropriate disclaimer, explicitly acknowledging the boundary between AI-assisted analysis and clinical decision-making.

This finding has several implications:

1. Risk of Overreliance: Users unfamiliar with the limitations of LLMs may interpret undisclaimed outputs as authoritative medical opinions. The confident, structured format of responses from GPT-5, Gemini 3 Pro, Llama 4 Maverick, and Grok 4 (complete with bullet points, differential diagnoses, and technical terminology) could foster unwarranted trust in the accuracy and completeness of these interpretations.
2. Medicolegal Considerations: In jurisdictions where AI-assisted diagnostic tools are subject to regulatory oversight, the absence of disclaimers could expose developers, healthcare institutions, and individual practitioners to liability. The failure to communicate the investigational or educational nature of AI outputs may constitute a breach of informed consent principles if patients are affected by decisions influenced by these tools.
3. Inconsistent Safety Culture Across Developers: The variability in disclaimer inclusion reflects divergent approaches to safety and responsible AI deployment among leading AI developers. While Anthropic's Claude model demonstrated alignment with principles of transparency and user protection, other models (including those from organizations with substantial resources and stated commitments to AI safety) did not incorporate similar safeguards in this use case.
4. Prompt Vulnerability: The observation that role-playing prompts (e.g., "You are a board-certified radiologist") can elicit authoritative medical responses without automatic safety guardrails in most models suggests a vulnerability in current LLM design. Responsible deployment in healthcare settings may require mandatory, non-overridable disclaimers when medical content is detected, regardless of prompt framing.

4.11.2. The Disclaimer as a Minimum Standard

It is important to note that the inclusion of a disclaimer does not validate an LLM for clinical use; rather, it represents a minimum ethical safeguard. Claude Opus 4.5 Extended's diagnostic interpretation, while accompanied by appropriate cautionary language, still contained the same types of variability observed in other models (e.g., omission of temporal lobe involvement, characterization as "acute-to-subacute"). The disclaimer appropriately signals that such outputs require expert verification; i.e., a message absent from the other four models.

4.12. The Correlated Error Problem

A fundamental assumption underlying multi-model consensus approaches is that errors across different LLMs are statistically independent. However, several factors suggest this assumption may be violated:

1. Shared Training Data: Many LLMs are trained on overlapping corpora, including publicly available medical literature, textbooks, and imaging databases. Systematic biases or errors present in these shared sources may propagate to multiple models.
2. Architectural Similarities: Despite differences in specific implementations, many leading LLMs share fundamental architectural principles (transformer-based architectures, attention mechanisms, similar tokenization strategies). These shared design choices may lead to correlated failure modes.
3. Benchmark Contamination: Models are increasingly optimized against publicly available benchmarks, which may lead to convergent behavior on tasks similar to benchmark items while masking divergent performance on novel or atypical presentations.

https://doi.org/10.3390/a19030170

4. Knowledge Cutoff Effects: LLMs with similar training cutoff dates may share the same knowledge gaps regarding recent developments, guideline updates, or emerging pathological entities.

The observation that four of five models converged on the same general diagnosis (ischemic stroke) while one model (Grok 4) made a fundamental error (hemorrhage) might initially suggest that majority consensus is protective. However, the cross-evaluation findings (particularly Gemini 3 Pro's divergent ground truth interpretation) demonstrate that even among "agreeing" models, fundamental interpretive differences may exist. Furthermore, a scenario in which correlated errors lead multiple models to converge on the same incorrect diagnosis cannot be excluded based on the present study and would represent a particularly dangerous failure mode for consensus-based systems.

4.13. Recommendations

Based on these findings, we recommend that:

- For clinical diagnostic imaging applications requiring reliable automated interpretation, healthcare institutions prioritize validated, task-specific machine learning models over general-purpose multimodal LLMs. Purpose-built diagnostic AI systems trained and cleared for defined imaging modalities and pathological conditions offer superior accuracy, known performance characteristics, regulatory oversight, and established clinical workflows compared to general-purpose LLMs, which demonstrate unacceptable variability and error rates for autonomous diagnostic use.
- General-purpose LLMs, if deployed in healthcare settings, should be restricted to applications that leverage their linguistic capabilities (such as report summarization, clinical documentation assistance, or natural language interfaces) rather than pixel-level diagnostic interpretation, for which they are not optimized.
- AI developers implement mandatory, context-aware disclaimers for medical content that cannot be suppressed by user prompts.
- Healthcare institutions establishing policies for LLM use require disclaimer verification as a prerequisite for any permitted applications.
- Regulatory bodies consider disclaimer requirements as part of pre-market approval processes for AI tools intended for healthcare-adjacent applications.
- Users of LLMs for medical education or preliminary analysis be trained to recognize the limitations of AI-generated content, regardless of whether disclaimers are present.
- Research into LLM-based quality assurance systems prioritize the development of calibration protocols, ground truth validation methods, and standardized rubric interpretation guidelines before deployment in educational or clinical settings.
- Multi-model consensus approaches, if pursued, be designed with explicit consideration of correlated error risks and validated against diverse, externally curated reference standards.
- Healthcare AI development strategies consider hybrid architectures that combine specialized diagnostic ML modules (for accurate image interpretation) with LLM-based components (for reporting, communication, and workflow integration), thereby leveraging the respective strengths of each paradigm while mitigating their limitations.

4.14. Limitations

The five models evaluated were selected to represent major Western AI developers and their flagship multimodal products. Notable models from other developers were not included due to practical constraints on study scope. Future research should examine whether the patterns of diagnostic variability and cross-evaluation inconsistency observed here extend to models developed outside the organizational contexts represented in this study.

https://doi.org/10.3390/a19030170

This study utilized a single non-contrast head CT case depicting unambiguous intracranial pathology. This design choice was intentional: by selecting a straightforward case, we aimed to establish a minimum competency threshold for LLM diagnostic performance. However, we acknowledge that a single case cannot capture the full spectrum of stroke presentations, imaging artifacts, anatomical variants, or diagnostically challenging edge cases that these systems would encounter in clinical practice. Our study design prioritized depth of analysis over breadth of cases, but future research employing multi-case datasets with varying levels of diagnostic difficulty would provide valuable complementary evidence regarding the generalizability of these findings across diverse clinical presentations.

This study did not include human radiologist ground truth labeling or grading of LLM responses. This omission was intentional: the authors, as researchers at a medical college with no financial or advisory relationships with any AI developer, sought to avoid producing comparative accuracy data that could be selectively cited to market specific LLM products. The selected image depicts pathology sufficiently obvious that radiologist confirmation would add little beyond affirming what is clinically apparent from the image itself, while potentially enabling claims such as "Model X achieved Y% agreement with expert radiologists." Our objective was to demonstrate that even on straightforward cases, leading LLMs exhibit fundamental diagnostic errors and irreconcilable interpretive disagreements, i.e., findings that argue against the entire class of tools rather than favoring any particular product. Future research incorporating blinded radiologist evaluation may provide complementary evidence, but such studies should be designed with explicit safeguards against selective commercial citation of model-specific accuracy figures.

Finally, the inherent stochasticity of LLM inference limits the strict reproducibility of the findings reported in this study. Commercial LLM providers do not disclose system prompts, safety filter configurations, or image preprocessing pipelines, and these parameters may change without notice. The model versions and access dates are documented in Table 3, but researchers attempting replication should anticipate output variability. This irreproducibility is itself clinically relevant: medical applications typically require auditable, consistent outputs that current LLM architectures cannot guarantee.

5. Conclusions

This study demonstrates that current multimodal large language models are not reliable tools for autonomous medical image interpretation. By deliberately selecting a straightforward case depicting unambiguous intracranial pathology, we established a minimum competency threshold: if state-of-the-art LLMs cannot consistently interpret obvious pathology, their deployment on diagnostically challenging cases cannot be justified. A 20% rate of fundamental diagnostic error (misidentifying ischemic stroke as intracerebral hemorrhage with incorrect lateralization), persistent clinically meaningful variability among ostensibly concordant models, and irreconcilable ground truth disagreement between grading LLMs collectively establish that these systems lack the consistency and accuracy required for autonomous clinical deployment. The cross-evaluation protocol further revealed that LLMs cannot serve as reliable automated evaluators: self-evaluation bias, divergent grading philosophies, and categorical disagreement about radiological findings undermine any proposed use in quality assurance or educational assessment contexts.

The implications of these findings vary by deployment context. For autonomous diagnostic applications, healthcare institutions should recognize that general-purpose LLMs (despite their linguistic sophistication) are architecturally unsuited for tasks demanding pixel-level accuracy and validated performance characteristics. For applications requiring reliable automated image interpretation, validated task-specific machine learning models trained and cleared for defined imaging modalities and pathological conditions remain the

https://doi.org/10.3390/a19030170

appropriate technological choice. Decision support applications, wherein LLMs serve as preliminary screening tools subject to expert verification, require explicit human-in-the-loop workflows with safeguards against cognitive anchoring and inappropriate reliance. Educational summarization represents the most appropriate current use case for LLMs in radiology, provided that mandatory disclaimers and expert oversight are maintained, and that AI limitations are explicitly integrated into the learning experience.

LLMs may offer value in healthcare applications that leverage their linguistic capabilities, such as clinical documentation, report summarization, or patient communication, but expert radiologist oversight remains non-negotiable for all diagnostic imaging interpretation. This study was designed to demonstrate class-wide limitations of multimodal LLMs for diagnostic imaging rather than to rank individual products, reflecting the authors' position that comparative performance assessments risk serving as inadvertent product endorsements. The intentional exclusion of human radiologist ground truth avoided generating accuracy metrics that could be selectively cited for commercial purposes while focusing attention on the fundamental reliability concerns that affect the entire category of general-purpose multimodal language models. Until LLMs demonstrate reproducible accuracy validated against external reference standards across diverse clinical presentations, their outputs should be treated as unverified preliminary information rather than actionable clinical guidance. The inherent stochasticity of LLM inference and the opacity of commercial model updates further limit the reproducibility and auditability that medical applications demand, reinforcing the case for task-specific models in contexts requiring consistent, validated performance.

Author Contributions: Conceptualization, S.H., M.M. and M.T.; methodology, S.H., M.M. and M.T.; software, S.H., M.M. and M.T.; validation, S.H., M.M. and M.T.; formal analysis, S.H., M.M. and M.T.; investigation, S.H., M.M. and M.T.; resources, S.H., M.M. and M.T.; data curation, S.H., M.M. and M.T.; writing—original draft preparation, S.H., M.M. and M.T.; writing—review and editing, S.H., M.M. and M.T.; visualization, S.H., M.M. and M.T.; supervision, S.H., M.M. and M.T.; project administration, S.H., M.M. and M.T.; funding acquisition, S.H., M.M. and M.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All data generated for this study are contained within the article. The complete set of LLM-generated radiological interpretations, cross-evaluation scoring matrices, and analytical results are presented in the manuscript. Further inquiries can be directed to the corresponding author.

Acknowledgments: During the preparation of this manuscript, the author(s) used AI-assisted tools for the purposes of grammar review and language editing. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest. The authors have no financial, advisory, employment, or other relationships with OpenAI, Google DeepMind, Meta AI, xAI, Anthropic, or any other AI developer, whether or not their products were evaluated in this study. No AI developer was consulted during study design, provided funding, or had any input into data analysis or manuscript preparation. The selection of models for evaluation was based solely on technical criteria (multimodal capabilities, availability, and market prominence) without consideration of commercial interests. The study was conducted independently with the sole objective of evaluating the reliability of multimodal LLMs for medical image interpretation, not to rank or endorse any particular product.

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial Intelligence
API	Application Programming Interface
AVM	Arteriovenous Malformation
CT	Computed Tomography
CV	Coefficient of Variation
HSV	Herpes Simplex Virus
ICC	Intraclass Correlation Coefficient
LLM	Large Language Model
MCA	Middle Cerebral Artery
MRI	Magnetic Resonance Imaging

Appendix A. Complete Prompts Used in This Study

This appendix provides the complete, verbatim prompts used for both the diagnostic analysis phase and the cross-evaluation grading phase of the study. These prompts were submitted identically to all five LLMs to ensure consistency in evaluation parameters.

Appendix A.1. Diagnostic Analysis Prompt

The following prompt was submitted to all five LLMs along with the test image (Figure 3):

> *You are a board-certified radiologist. Please analyze the attached image and answer the following in a concise, bullet-point format:*
>
> 1. *What type of imaging modality is shown? (e.g., X-ray, CT, MRI, ultrasound)*
> 2. *Which anatomical region or organ is depicted?*
> 3. *What is the most likely diagnosis based on the image findings?*
> 4. *Briefly describe the key image features that support your diagnosis.*
> 5. *List up to three possible differential diagnoses.*

No additional context, patient history, or clinical information was provided beyond what is contained in the prompt above. The image was uploaded directly through each platform's native image upload functionality without preprocessing or modification.

Appendix A.2. Cross-Evaluation Grading Prompt

The following prompt was submitted to each LLM for the cross-evaluation phase, along with the original test image and all five anonymized analyses labeled as Analysis 1 through Analysis 5:

> *You are an expert medical educator and radiologist tasked with systematically evaluating and grading radiological image analyses. You will assess 5 medical image interpretations, focusing on accuracy, completeness, and clinical appropriateness.*
>
> *Evaluation Context*
>
> *Image: [Medical image provided]*
>
> *Original Prompt: [Original prompt provided]*
>
> *Task: Grade 5 analyses (provided) for accuracy across 5 domains detailed in the original prompt.*

The grading prompt was followed by the complete grading rubric (Appendix B), the stepwise evaluation instructions (Appendix C), and the quality assurance guidelines

https://doi.org/10.3390/a19030170

(Appendix C.1). The five anonymized analyses were appended sequentially, with no identification of the originating model.

Appendix B. Detailed Grading Rubric

This appendix provides the complete grading rubric used in the cross-evaluation protocol. Each of the five LLMs used these criteria to systematically evaluate all five anonymized analyses across five domains.

Table A1. Complete grading rubric for cross-model evaluation of radiological image analyses.

Domain	Scale	Detailed Scoring Criteria
Imaging Modality	0–3	0 = Incorrect identification of modality; 1 = Vague or partially correct (e.g., "scan" without specifying type); 2 = Correct modality but lacks specificity (e.g., "CT" without mentioning contrast status or plane); 3 = Fully correct and specific (e.g., "non-contrast axial CT")
Anatomical Region and Organ	0–3	0 = Incorrect anatomical localization; 1 = Vague or partially correct (e.g., "body" or "head" without further specification); 2 = Correct general region but lacks anatomical specificity (e.g., "brain" without specifying structures or slice level); 3 = Fully correct and specific with appropriate anatomical detail (e.g., "brain at the level of the lateral ventricles and basal ganglia")
Most Likely Diagnosis	0–4	0 = Incorrect or irrelevant diagnosis; 1 = Vague, overly broad, or tangentially relevant diagnosis; 2 = Plausible diagnosis but not the most likely given the findings; 3 = Correct diagnosis but lacks clinical detail or precision; 4 = Fully correct, specific, and clinically appropriate diagnosis with relevant qualifiers (e.g., acuity, lateralization, vascular territory)
Key Image Features	0–4	0 = Absent, incorrect, or contradictory feature descriptions; 1 = Incomplete or vague descriptions that lack clinical utility; 2 = Some correct features identified but missing key findings; 3 = Mostly complete and accurate feature descriptions with minor omissions; 4 = Complete, accurate, and well-described key features that comprehensively support the diagnosis
Differential Diagnoses	0–4	0 = Absent, irrelevant, or implausible differential diagnoses; 1 = Only one plausible alternative provided; 2 = Two plausible alternatives provided; 3 = Three plausible alternatives provided but not prioritized or justified; 4 = Three plausible alternatives that are appropriately prioritized and clinically justified
Maximum Total Score		18 points

Each domain was scored independently, and total scores were calculated by summing across all five domains. The rubric was designed to balance objectivity (through clear anchors for each score level) with flexibility to accommodate different but clinically equivalent descriptions. Grading LLMs were instructed to apply these criteria systematically to all five analyses without knowledge of which model generated each response.

https://doi.org/10.3390/a19030170

Appendix C. Grading Protocol: Stepwise Instructions and Quality Assurance Guidelines

To promote systematic reasoning and reduce variability, the grading prompt instructed each LLM to follow a stepwise chain-of-thought evaluation process. Notably, rather than providing an external reference standard, each grading LLM was instructed to use its own expert knowledge to establish the expected correct answer for each domain, i.e., an approach that enables assessment of whether LLMs hold consistent internal standards for radiological accuracy.

The stepwise instructions were as follows:

1. For each of the five analyses, follow these steps:
    (a) For each domain (modality, anatomy, diagnosis, features, differential):
        i. Use your expert knowledge to determine the most accurate and specific answer based on the image and context.
        ii. Compare the analysis to your expert expectation for that domain.
        iii. Assign a score based on the rubric.
        iv. Provide a brief justification for the score, referencing your clinical reasoning.
    (b) After all domains are graded:
        i. Calculate and state the total score for the analysis.
        ii. Provide a concise summary of the analysis's strengths and weaknesses.
2. Repeat this process for all five analyses.
3. At the end, rank the analyses from highest to lowest total score.

This self-referential grading paradigm (wherein each LLM serves as both participant and evaluator using its own internal knowledge as the reference standard) introduces an additional dimension of analysis. Discrepancies in scores assigned by different grading LLMs to the same analysis may arise not only from differences in grading stringency but also from fundamental disagreements about what constitutes the correct radiological interpretation. This design choice enables examination of whether LLMs possess internally consistent and externally concordant standards for radiological accuracy, a prerequisite for their use in automated quality assurance or educational assessment applications.

Appendix C.1. Quality Assurance Instructions

Additionally, the grading prompt included explicit quality assurance instructions to promote consistency and objectivity:

- Base all judgments strictly on expert clinical knowledge and the information provided in the analysis.
- Avoid speculation or assumptions beyond the evidence presented in the analysis.
- If information is missing or ambiguous, note this explicitly in the reasoning.
- Be concise, objective, and systematic in all grading.

These instructions were designed to standardize the evaluative approach across all grading LLMs, reducing variability attributable to differences in interpretive style or inferential tendencies. By requiring graders to flag missing or ambiguous information rather than making assumptions, the protocol aimed to distinguish between analyses that were incomplete versus those that were incorrect, which is a clinically important distinction.

https://doi.org/10.3390/a19030170

References

1. Turing, A.M. I.—Computing Machinery and Intelligence. *Mind* **1950**, *LIX*, 433–460. [CrossRef]
2. McCarthy, J.; Minsky, M.L.; Rochester, N.; Shannon, C.E. A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence, August 31, 1955. *AI Mag.* **2006**, *27*, 12. [CrossRef]
3. Newell, A.; Simon, H.A. Computer science as empirical inquiry: Symbols and search. *Commun. ACM* **1976**, *19*, 113–126. [CrossRef]
4. Rumelhart, D.E.; Hinton, G.E.; Williams, R.J. Learning representations by back-propagating errors. *Nature* **1986**, *323*, 533–536. [CrossRef]
5. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [CrossRef]
6. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet classification with deep convolutional neural networks. *Commun. ACM* **2017**, *60*, 84–90. [CrossRef]
7. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. In Proceedings of the 31st International Conference on Neural Information Processing Systems, Red Hook, NY, USA, 4–9 December 2017; pp. 6000–6010. Available online: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf (accessed on 14 December 2025).
8. Tay, Y.; Dehghani, M.; Bahri, D.; Metzler, D. Efficient Transformers: A Survey. *ACM Comput. Surv.* **2022**, *55*, 1–28. [CrossRef]
9. Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language models are few-shot learners. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 1877–1901. Available online: https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf (accessed on 14 December 2025).
10. Chen, K.; Shao, A.; Burapacheep, J.; Li, Y. Conversational AI and equity through assessing GPT-3's communication with diverse social groups on contentious topics. *Sci. Rep.* **2024**, *14*, 1561. [CrossRef]
11. Denecke, K.; May, R.; Rivera-Romero, O. Transformer Models in Healthcare: A Survey and Thematic Analysis of Potentials, Shortcomings and Risks. *J. Med Syst.* **2024**, *48*, 23. [CrossRef]
12. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2019; Volume 1 (Long and Short Papers), pp. 4171–4186. [CrossRef]
13. Bekbolatova, M.; Mayer, J.; Ong, C.W.; Toma, M. Transformative Potential of AI in Healthcare: Definitions, Applications, and Navigating the Ethical Landscape and Public Perspectives. *Healthcare* **2024**, *12*, 125. [CrossRef] [PubMed]
14. Arısan, A.; Genç, C.; Duran, G.S. Orthodontic Biomechanical Reasoning with Multimodal Language Models: Performance and Clinical Utility. *Bioengineering* **2025**, *12*, 1165. [CrossRef] [PubMed]
15. Goh, E.; Gallo, R.; Hom, J.; Strong, E.; Weng, Y.; Kerman, H.; Cool, J.A.; Kanjee, Z.; Parsons, A.S.; Ahuja, N.; et al. Large Language Model Influence on Diagnostic Reasoning: A Randomized Clinical Trial. *JAMA Netw. Open* **2024**, *7*, e2440969. [CrossRef] [PubMed]
16. Urda-Cimpean, A.E.; Leucuța, D.C.; Drugan, C.; Dutu, A.G.; Calinici, T.; Drugan, T. Assessing the Accuracy of Diagnostic Capabilities of Large Language Models. *Diagnostics* **2025**, *15*, 1657. [CrossRef]
17. Tan, J.C.K. Coherent Interpretation of Entire Visual Field Test Reports Using a Multimodal Large Language Model (ChatGPT). *Vision* **2025**, *9*, 33. [CrossRef]
18. Lodwick, G.S.; Keats, T.E.; Dorst, J.P. The Coding of Roentgen Images for Computer Analysis as Applied to Lung Cancer. *Radiology* **1963**, *81*, 185–200. [CrossRef]
19. Lodwick, G.S. Computer-aided Diagnosis in Radiology: A Research Plan. *Investig. Radiol.* **1966**, *1*, 72–80. [CrossRef]
20. Kruger, R.P.; Townes, J.R.; Hall, D.L.; Dwyer, S.J.; Lodwick, G.S. Automated Radiographic Diagnosis via Feature Extraction and Classification of Cardiac Size and Shape Descriptors. *IEEE Trans. Biomed. Eng.* **1972**, *BME-19*, 174–186. [CrossRef]
21. Toriwaki, J.I.; Suenaga, Y.; Negoro, T.; Fukumura, T. Pattern recognition of chest X-ray images. *Comput. Graph. Image Process.* **1973**, *2*, 252–271. [CrossRef]
22. Kruger, R.P.; Thompson, W.B.; Turner, A.F. Computer Diagnosis of Pneumoconiosis. *IEEE Trans. Syst. Man Cybern.* **1974**, *SMC-4*, 40–49. [CrossRef]
23. Turner, A.F.; Kruger, R.P.; Thompson, W.B. Automated Computer Screening of Chest Radiographs for Pneumoconiosis. *Investig. Radiol.* **1976**, *11*, 258–266. [CrossRef] [PubMed]
24. Fenton, J.J.; Taplin, S.H.; Carney, P.A.; Abraham, L.; Sickles, E.A.; D'Orsi, C.; Berns, E.A.; Cutter, G.; Hendrick, R.E.; Barlow, W.E.; et al. Influence of Computer-Aided Detection on Performance of Screening Mammography. *N. Engl. J. Med.* **2007**, *356*, 1399–1409. [CrossRef] [PubMed]
25. Esteva, A.; Kuprel, B.; Novoa, R.A.; Ko, J.; Swetter, S.M.; Blau, H.M.; Thrun, S. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **2017**, *542*, 115–118. [CrossRef] [PubMed]

26. Rajpurkar, P.; Irvin, J.; Ball, R.L.; Zhu, K.; Yang, B.; Mehta, H.; Duan, T.; Ding, D.; Bagul, A.; Langlotz, C.P.; et al. Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists. *PLoS Med.* **2018**, *15*, e1002686. [CrossRef]
27. Gulshan, V.; Peng, L.; Coram, M.; Stumpe, M.C.; Wu, D.; Narayanaswamy, A.; Venugopalan, S.; Widner, K.; Madams, T.; Cuadros, J.; et al. Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. *JAMA* **2016**, *316*, 2402. [CrossRef]
28. Husain, G.; Mayer, J.; Bekbolatova, M.; Vathappallil, P.; Matalia, M.; Toma, M. Machine learning for medical image classification. *Acad. Med.* **2024**, *1*. [CrossRef]
29. Toma, M.; Husain, G. Algorithm Selection and Data Utilization in Machine Learning for Medical Imaging Classification. In *2024 IEEE Long Island Systems, Applications and Technology Conference (LISAT)*; IEEE: Piscataway, NJ, USA, 2024; pp. 1–6. [CrossRef]
30. Litjens, G.; Kooi, T.; Bejnordi, B.E.; Setio, A.A.A.; Ciompi, F.; Ghafoorian, M.; van der Laak, J.A.; van Ginneken, B.; Sánchez, C.I. A survey on deep learning in medical image analysis. *Med. Image Anal.* **2017**, *42*, 60–88. [CrossRef]
31. Franco, L.; Toma, M. Validation strategies for automated MRI-based classification of Alzheimer's disease using deep feature extraction and machine learning. *Artif. Intell. Health* **2025**, 025360073. [CrossRef]
32. Singhal, K.; Azizi, S.; Tu, T.; Mahdavi, S.S.; Wei, J.; Chung, H.W.; Scales, N.; Tanwani, A.; Cole-Lewis, H.; Pfohl, S.; et al. Large language models encode clinical knowledge. *Nature* **2023**, *620*, 172–180. [CrossRef]
33. Azizian, P.; Honarmand, M.; Jaiswal, A.; Kline, A.; Dunlap, K.; Washington, P.; Wall, D.P. Multimodal LLM vs. Human-Measured Features for AI Predictions of Autism in Home Videos. *Algorithms* **2025**, *18*, 687. [CrossRef]
34. Bender, E.M.; Gebru, T.; McMillan-Major, A.; Shmitchell, S. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *FAccT '21: 2021 ACM Conference on Fairness, Accountability, and Transparency*; ACM: New York, NY, USA, 2021; pp. 610–623. [CrossRef]
35. Nasef, D.; Nasef, D.; Sher, M.; Toma, M. A Standardized Validation Framework for Clinically Actionable Healthcare Machine Learning with Knee Osteoarthritis Grading as a Case Study. *Algorithms* **2025**, *18*, 343. [CrossRef]
36. Toma, M. *AI-Assisted Medical Diagnostics: A Clinical Guide to Next-Generation Diagnostics*; Dawning Research Press: New York, NY, USA, 2025. Available online: https://openlibrary.org/works/OL44048041W/ (accessed on 14 December 2025).
37. Toma, M.; Wei, O.C. Predictive Modeling in Medicine. *Encyclopedia* **2023**, *3*, 590–601. [CrossRef]
38. Chilamkurthy, S.; Ghosh, R.; Tanamala, S.; Biviji, M.; Campeau, N.G.; Venugopal, V.K.; Mahajan, V.; Rao, P.; Warier, P. Deep learning algorithms for detection of critical findings in head CT scans: A retrospective study. *Lancet* **2018**, *392*, 2388–2396. [CrossRef] [PubMed]
39. Remyes, D.; Nasef, D.; Remyes, S.; Tawfellos, J.; Sher, M.; Nasef, D.; Toma, M. Clinical Applicability and Cross-Dataset Validation of Machine Learning Models for Binary Glaucoma Detection. *Information* **2025**, *16*, 432. [CrossRef]
40. Muehlematter, U.J.; Bluethgen, C.; Vokinger, K.N. FDA-cleared artificial intelligence and machine learning-based medical devices and their 510(k) predicate networks. *Lancet Digit. Health* **2023**, *5*, e618–e626. [CrossRef]
41. Merk, D.; Grisoni, F.; Friedrich, L.; Schneider, G. Tuning artificial intelligence on the de novo design of natural-product-inspired retinoid X receptor modulators. *Commun. Chem.* **2018**, *1*, 68. [CrossRef]
42. Aboy, M.; Crespo, C.; Stern, A. Beyond the 510(k): The regulation of novel moderate-risk medical devices, intellectual property considerations, and innovation incentives in the FDA's De Novo pathway. *npj Digit. Med.* **2024**, *7*, 29. [CrossRef]
43. Leucuța, D.C.; Urda-Cîmpean, A.E.; Istrate, D.; Drugan, T. Risk of Bias Assessment of Diagnostic Accuracy Studies Using QUADAS 2 by Large Language Models. *Diagnostics* **2025**, *15*, 1451. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

MDPI AG
Grosspeteranlage 5
4052 Basel
Switzerland
Tel.: +41 61 683 77 34

Algorithms Editorial Office
E-mail: algorithms@mdpi.com
www.mdpi.com/journal/algorithms



Disclaimer/Publisher's Note: The title and front matter of this reprint are at the discretion of the Guest Editor. The publisher is not responsible for their content or any associated concerns. The statements, opinions and data contained in all individual articles are solely those of the individual Editor and contributors and not of MDPI. MDPI disclaims responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.





Academic Open
Access Publishing

mdpi.com

ISBN 978-3-7258-9058-3