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.

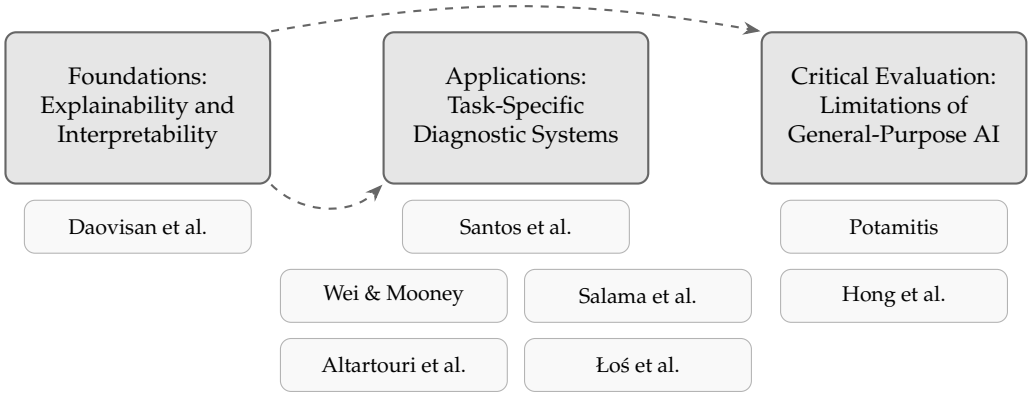

Figure 1. The organization of the three thematic categories for this Special Issue. The foundations pillar establishes theoretical principles that inform the task-specific applications, while critical evaluation provides necessary constraints on technology selection (Contributions 1–8).

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

Received: 3 April 2026
Revised: 14 April 2026
Accepted: 22 April 2026
Published: 24 April 2026

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.

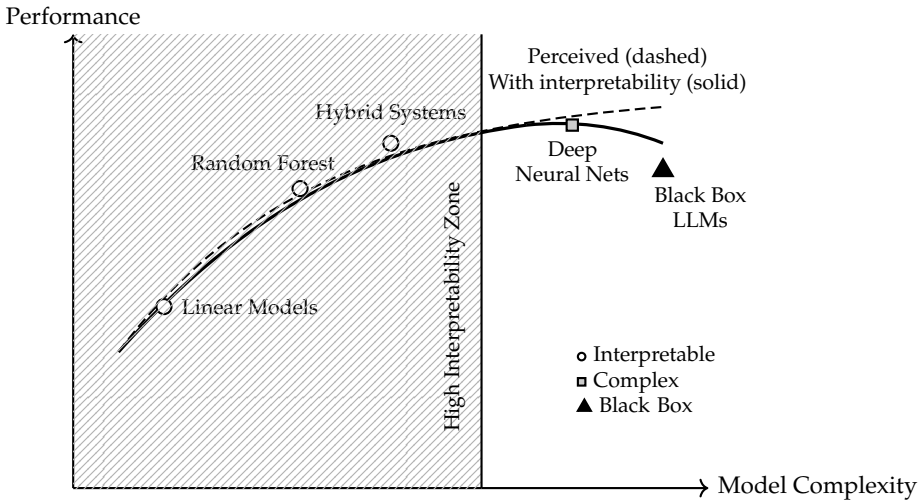

Figure 2. The relationship between model complexity, performance, and interpretability in clinical AI. The dashed curve shows the perceived unlimited performance gain with complexity, while the solid curve represents actual performance when interpretability constraints are considered. Excessive complexity 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%

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.

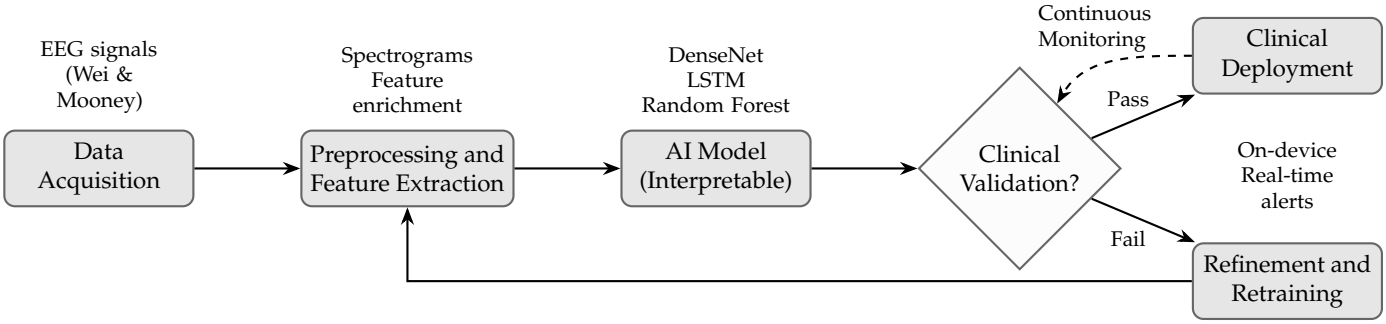

Figure 3. Common pipeline for task-specific diagnostic AI systems illustrated by the applications in this Special Issue. All systems emphasize interpretability and require rigorous clinical validation before deployment, with continuous monitoring post-deployment (Contribution 3).

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-

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).

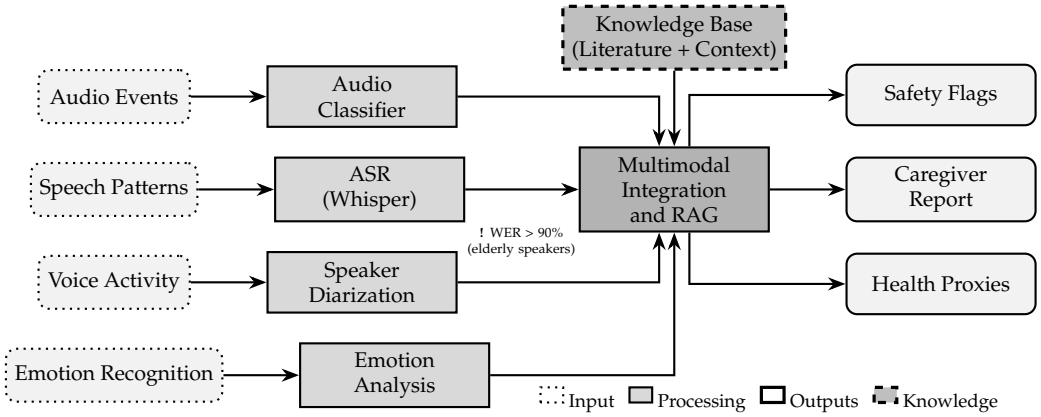

Figure 4. Multimodal integration architecture for dementia care monitoring (Potamitis). The system combines audio event detection, speech recognition, speaker diarization, and emotion analysis with retrieval-augmented generation to produce comprehensive caregiver reports. Node styles: dotted borders 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.

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.

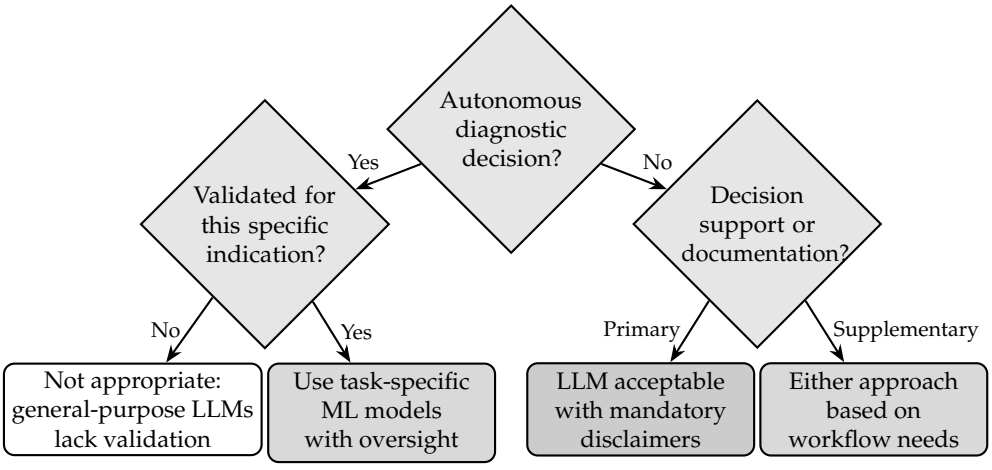

Figure 5. Decision framework for selecting between general-purpose LLMs and task-specific machine learning models in clinical settings. Autonomous diagnostic applications require cleared task-specific models, while LLMs may serve supporting roles with appropriate disclaimers.

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

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.

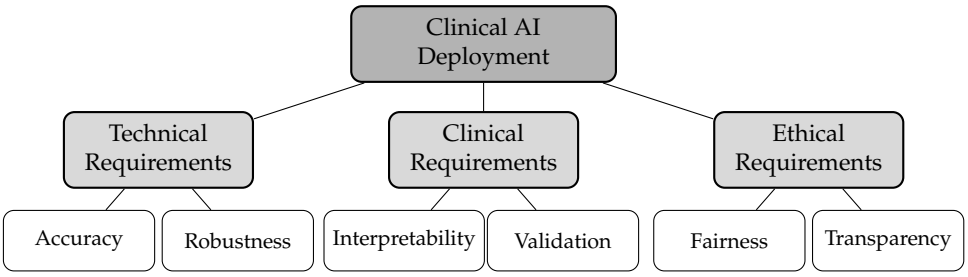

Figure 6. Hierarchical requirements for responsible clinical AI deployment. Technical performance alone is insufficient; clinical and ethical requirements are equally essential. Visual hierarchy: darker shading and thicker borders indicate higher-level concepts. All requirements must be satisfied simultaneously. 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.

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.