INTERNATIONAL JOURNAL OF ADVANCED ARTIFICIAL INTELLIGENCE RESEARCH (IJAAIR)

eISSN: 3087-4270

Volume. 02, Issue. 10, pp. 52-63, October 2025

ENHANCING TRUST AND CLINICAL ADOPTION: A SYSTEMATIC LITERATURE REVIEW OF EXPLAINABLE ARTIFICIAL INTELLIGENCE (XAI) APPLICATIONS IN HEALTHCARE

Dr. Elias T. Vance
Department of Health Informatics, Biomedical Technology Research Center, London, United Kingdom

Prof. Camille A. Lefevre
Department of Health Informatics, Biomedical Technology Research Center, London, United Kingdom

Article received: 13/08/2024, Article Accepted: 17/09/2025, Article Published: 31/10/2025

ABSTRACT

Background: The transformative potential of Artificial Intelligence (AI) in healthcare is hampered by the "black box" problem, where a lack of transparency in decision-making fundamentally undermines clinician trust and creates barriers to clinical adoption. Explainable Artificial Intelligence (XAI) is proposed as a necessary solution to bridge the gap between high-performance AI models and the critical need for justification and accountability in patient care.
Methods: This systematic literature review was conducted in adherence to PRISMA guidelines, analyzing literature published between January 2020 and early 2024. A rigorous search across major databases identified 50 relevant primary studies on XAI applications in clinical and biomedical contexts. Data extracted included the medical domain, AI model, XAI technique, and reported impact on trust and accuracy.
Results: Analysis of the 50 studies demonstrated a wide application of XAI across diverse medical fields, including diagnostics, medical imaging, and disease prediction. XAI—especially methods like SHAP, LIME, and GRAD-CAM—was found to significantly enhance interpretability, transparency, and diagnostic accuracy in these applications, successfully building clinician confidence in AI systems. The primary applications were observed in areas like chronic wound classification, cancer diagnosis, and cardiovascular risk prediction.
Conclusion: XAI is paramount for the safe and effective integration of AI into clinical practice. However, real-world integration is associated with persistent technical and data-quality challenges, including inconsistent validation and biased datasets. Future efforts must prioritize the development of standardized frameworks and regulatory compliance to ensure safe, ethical, and fully explainable AI use in healthcare.

KEYWORDS

Explainable Artificial Intelligence (XAI), Healthcare AI, Systematic Review, Clinical Adoption, Interpretability, Machine Learning, Trust.

INTRODUCTION

1.1. The AI Revolution in Healthcare: Promise and Peril

The integration of Artificial Intelligence (AI) and Machine Learning (ML) into healthcare has ushered in an era of unprecedented potential, promising to redefine clinical workflows, improve diagnostic speed, and enable truly personalized medicine . Across the globe, AI models are demonstrating high-level efficacy in tasks ranging from classifying medical images to predicting patient outcomes and assisting in drug discovery . The core promise lies in AI's ability to process massive, complex datasets—such as genomic profiles, electronic health records (EHRs), and high-resolution imaging—far faster and sometimes with greater precision than human practitioners alone . Applications span virtually every medical sub-specialty, including radiology, pathology, cardiology, and oncology, positioning AI as a crucial tool for future healthcare systems .

However, this technological leap is not without its significant challenges, particularly concerning the deployment of sophisticated deep learning models. These powerful models often operate as "black boxes," meaning their internal decision-making processes are opaque and unintuitive to human observers . While the model may deliver a correct diagnosis, the reasoning pathway that is associated with that decision remains obscured. This opacity presents a critical problem for clinical adoption, as it directly conflicts with the fundamental principles of medical practice: accountability, justification, and patient safety . Clinicians need to understand why a model made a specific prediction before they can confidently incorporate it into a treatment plan or diagnosis, especially when dealing with high-stakes health decisions. The inability to inspect and validate an AI’s internal logic creates a barrier to trust and poses ethical and legal quandaries that must be addressed before widespread clinical integration can occur .

1.2. The Imperative for Explainable Artificial Intelligence (XAI)

The necessity for transparency has led to the emergence of Explainable Artificial Intelligence (XAI). XAI is an umbrella term for a suite of techniques designed to make the predictions of complex ML models comprehensible to humans. Its aim is not merely to improve the AI's performance but, more critically, to build a system of mutual understanding and trust between the technology and its human users .

The core argument driving this review is that the lack of transparency in AI decision-making fundamentally undermines trust in healthcare applications, creating barriers to clinical adoption . A clinician who cannot explain an AI’s output to a patient or justify it to a peer is unlikely to rely on it. XAI provides the necessary tools for generating insights, often in the form of feature importance scores or visual heatmaps, that clarify which input data points—be they specific genes, image regions, or vital signs—were most influential in the model's final output .

Beyond clinical confidence, XAI addresses critical ethical and regulatory requirements. From an ethical standpoint, explanations help identify and mitigate potential biases embedded in the training data, ensuring the model's decisions are fair and equitable across different demographic groups . Legally, as regulatory bodies like the FDA and organizations publishing guidelines like DECIDE-AI begin to grapple with the complexities of AI-driven medicine, the ability to audit and explain an AI's decision-making process is becoming a non-negotiable requirement for clinical approval and implementation . Therefore, XAI is not a peripheral feature but an imperative for transforming AI from a laboratory tool into a dependable clinical assistant.

1.3. Review Objectives and Structure

Given the rapidly evolving landscape and the critical need for XAI integration, a comprehensive synthesis of current research is essential. This systematic literature review aims to provide a structured overview of the application of XAI in the healthcare domain over the past four years.

Our primary objectives are:

1. To map the diverse medical domains where XAI has been actively applied.

2. To identify and categorize the dominant XAI techniques (e.g., LIME, SHAP) being utilized with various ML models.

3. To synthesize evidence regarding the reported association of XAI with clinician trust, transparency, and model accuracy.

4. To discuss the inherent technical and practical challenges associated with limiting the widespread real-world adoption of XAI in clinical environments.

The remainder of this article is structured as follows: Section 2 details the methodology employed for identifying and synthesizing the literature. Section 3 presents a structured analysis of the results, categorized by medical application and XAI technique. Section 4 offers a comprehensive discussion of the findings, implications, challenges, and future research directions.

2. Methods

2.1. Protocol and Registration

This systematic literature review was conducted in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) statement . While this review involves a qualitative synthesis of literature rather than a meta-analysis of quantitative data, adhering to PRISMA ensures maximum transparency and rigor in the reporting of the search and selection process.

2.2. Search Strategy and Data Sources

A comprehensive search strategy was designed to capture primary research articles on XAI applied to the healthcare sector. The search was systematically executed across major bibliographic databases, including PubMed, Scopus, Web of Science, and IEEE Xplore.

The search period was strictly defined as January 1, 2020, to early 2024, to capture the most contemporary research reflecting the recent surge in XAI methodologies.

A structured set of keywords was employed, combining terms related to explainability with terms related to the application domain:

- (Explainable AI OR XAI OR Interpretability OR LIME OR SHAP OR GRAD-CAM) AND (Healthcare OR Medicine OR Clinical OR Medical OR Diagnosis OR Patient OR Health Records)

2.3. Eligibility and Selection Criteria (Inclusion/Exclusion)

Inclusion Criteria	Exclusion Criteria
Peer-reviewed journal articles and archival conference papers.	Conference abstracts, presentations, editorials, and opinions.
Articles published in **English**.	Non-English language articles.
Focus on the **application of XAI or intrinsically interpretable models** within a medical, clinical, or biomedical research context.	Theoretical XAI papers with no specific healthcare application or general AI in healthcare reviews without an XAI focus.
Original research detailing the **methodology and results** of a specific XAI model.	Articles where the specific XAI technique was not clearly identifiable or described.
Publication date between **January 2020 and early 2024**.	Publications outside the defined search period.

2.4. Study Selection Process

The selection process proceeded through four stages as recommended by PRISMA: identification, screening, eligibility, and inclusion.

1. Identification: Initial searches were conducted across the defined databases, and duplicate records were removed.

2. Screening: Titles and abstracts of the retrieved articles were independently screened against the inclusion/exclusion criteria.

3. Eligibility: The full texts of potentially relevant articles were retrieved and assessed in detail. Any article that failed to meet the specific criteria (e.g., lack of clear XAI method description, non-clinical focus) was excluded.

4. Inclusion: The final set of papers was determined for data extraction and synthesis.

2.5. Data Extraction and Synthesis

For each included study, the following key variables were systematically extracted:

- Publication year and primary author.

- Specific Medical Domain (e.g., Cardiology, Radiology).

- AI Model Type (e.g., Convolutional Neural Network (CNN), Support Vector Machine (SVM)).

- XAI Method utilized (e.g., SHAP, LIME, Grad-CAM, intrinsically interpretable).

- Key Findings related to model performance, interpretability, and trust.

The final synthesis involved a thematic analysis approach. Studies were grouped by their application domain and the type of XAI technique employed. This thematic grouping allowed for a comprehensive qualitative synthesis of findings, highlighting trends in XAI usage, identifying the most effective explanation modalities, and synthesizing the reported challenges to implementation.

3. Results

3.1. Study Selection and Characteristics

Following the systematic search and rigorous application of the eligibility criteria, a total of 50 studies published between 2020 and 2024 were included for qualitative synthesis. This high volume of recent publications underscores the explosive growth and recognized importance of XAI within the medical research community since the beginning of the decade.

The studies covered an extensive range of clinical applications, reflecting the universal need for transparency wherever AI is applied in healthcare. The distribution of studies by year shows a clear upward trend, with the highest concentration of research appearing in 2021 and 2022, signaling a rapid maturation of the field.

3.2. Mapping of XAI Applications Across Clinical Domains

The applications of XAI can be broadly grouped into four major clinical domains, with the majority focusing on visual diagnostics and critical risk prediction:

3.2.1. Medical Imaging and Diagnostics

Imaging-based diagnostics constitutes a highly active area for XAI research, primarily because XAI techniques can generate visual explanations that resonate intuitively with radiologists and pathologists.

- ● Cancer Detection: XAI is extensively used in oncology to justify the results of deep learning models. Studies covered detailed analysis of mammograms , interpretation of pulmonary diseases from chest radiographs , and even the identification and prediction of brain tumors . In these cases, heatmaps generated by methods like Grad-CAM are crucial, highlighting the specific tumor or lesion regions that were associated with the diagnostic prediction, thus increasing clinician confidence in the model's focus .
- ● Infectious Diseases and Wound Care: XAI has played a vital role in rapidly developing models for COVID-19 diagnosis using CT scans and X-rays . Similarly, XAI-CWC, a highly transparent tool, was developed for chronic wound classification , where transparency is essential for treatment planning. XAI has also been applied to diagnose fungal keratitis using in vivo confocal microscopy images and to detect tuberculosis from chest radiographs .
- ● Neurological Imaging and Ophthalmology: XAI models have been employed for retinoblastoma diagnosis, using LIME and SHAP to interpret deep learning model decisions on ocular images . The use of XAI in analyzing cerebrospinal fluid for diagnostics further demonstrates its utility in complex neurological conditions .

3.2.2. Disease Prediction and Risk Assessment

This domain leverages XAI to explain complex, multivariate risk models, often utilizing electronic health records (EHRs) or molecular data.

● Cardiovascular and Metabolic Risk: Studies focusing on cardiovascular event risk prediction use XAI to identify influential molecular data or analyze ECG signals . XAI was also applied to enhance a prediction model for heart failure survival and to predict cardiovascular outcomes using EHRs .

● Neurodegenerative Diseases: For conditions like Alzheimer's disease (AD), XAI is used to interpret multi-modal detection and prediction models . By explaining which features (e.g., MRI data, cognitive scores, genetic markers) were associated with an AD prediction, clinicians gain insight into the disease's progression factors.

● Acute Care and Risk Factors: XAI has been used for predicting the need for ventilator support and mortality in COVID-19 patients , predicting deterioration risk in hepatitis patients , and predicting readmission risk among frail patients . The interpretability provided is crucial for early intervention strategies.

3.2.3. Genetics, Drug Discovery, and Personalized Medicine

In highly complex, high-dimensional data environments like genomics and proteomics, XAI helps sift through millions of data points to identify causal or predictive features.

● Biomarker Identification: XAI has been instrumental in identifying biologically relevant gene expression patterns in longitudinal human studies , particularly in obesity research. It has also been applied to precision medicine in acute myeloid leukemia , where understanding the model's reliance on specific biomarkers is essential for tailoring treatment.

● Drug Development: XAI plays a key role in drug discovery, helping researchers understand why a model predicts a compound will be active or toxic . This can accelerate the discovery pipeline by focusing laboratory efforts on compounds with justifiable predicted efficacy.

3.2.4. Function and Behavioral Health

XAI is increasingly applied to monitor human function and behavioral health, where the context of the explanation is as important as the prediction itself.

● Mental and Neurological Health: XAI is being used to model biomedical mental disorder diagnoses and to tackle the complexity of mental health research by interpreting models of pediatric psychiatric conditions . In neurology, XAI models have interpreted stroke-impaired electromyography patterns and predicted stroke based on EEG signals .

- Functional Monitoring: XAI has been applied to interpret wearable sensor data for gait analysis, helping to identify patients with osteopenia and sarcopenia in daily life , and for falls prediction .

3.3. Analysis of Dominant XAI Techniques

The review confirms that the overwhelming majority of contemporary XAI applications in healthcare rely on model-agnostic, post-hoc explanation techniques . This preference is pragmatic: it allows researchers to deploy the most powerful, often opaque, deep learning architectures while still providing a subsequent, understandable rationale for a given decision. This synthesis of the 50 analyzed studies demonstrates unequivocally that Explainable Artificial Intelligence (XAI)—especially methods like SHAP, LIME, and GRAD-CAM—is associated with enhanced interpretability and builds clinician confidence in AI systems by translating complex algorithmic outputs into clinically actionable insights.

The most frequently employed techniques identified across the literature are SHAP, LIME, and GRAD-CAM. While a small number of studies utilized intrinsically interpretable models, particularly for low-dimensional or simplified risk models , their scope is limited by the exponential rise of deep learning in domains like medical imaging. Therefore, the focus of this analysis is on the mechanisms and limitations of the three dominant post-hoc approaches, which form the technical bedrock of XAI in modern clinical research.

3.3.1. Detailed Mechanisms of Dominant XAI Techniques

Understanding the internal workings of these XAI methods is crucial, as their underlying mathematical assumptions directly impact the faithfulness and utility of the explanation provided to the clinician. Each method approaches the problem of interpretability from a distinct theoretical foundation, leading to different strengths and weaknesses in clinical contexts.

3.3.1.1. SHAP (SHapley Additive exPlanations)

SHAP represents a conceptual leap in model explainability by unifying several existing methods (such as LIME, DeepLIFT, and feature importance methods) under a single theoretical framework rooted in cooperative game theory . The fundamental goal of SHAP is to assign a unique, justifiable importance value—a Shapley value—to every feature for a specific prediction.

Mechanism and Theoretical Foundation:

In cooperative game theory, the Shapley value, developed by Lloyd Shapley, is the only method that satisfies a set of desirable properties: local accuracy (the explanation must match the model output for the instance being explained), missingness (features whose value is zero should have no impact), and consistency (if a feature's contribution increases, its importance should not decrease).

In the context of a machine learning model, the "game" is the prediction task, the "players" are the input features, and the "payout" is the difference between the actual prediction and the average expected prediction (the baseline). The SHAP value for a feature is its weighted average marginal contribution across all possible feature coalitions (all permutations in which the feature could be introduced).

The SHAP explanation model is an additive feature attribution method:

where is the model prediction, is the baseline expectation (average prediction), is a simplified input feature (e.g.,), and is the SHAP value of feature . The Shapley value () for feature is calculated as:

where is the set of all features, is a subset of features without feature , and is the prediction function with only features in present.

Clinical Utility and Computational Challenges:

SHAP's key clinical advantage is its global consistency. Because the calculation is derived from a robust theoretical foundation, a SHAP value for a feature can be interpreted consistently across different predictions. For instance, in predicting lung and bronchus cancer mortality rates , SHAP can consistently quantify how socioeconomic factors or environmental exposures are associated with the risk compared to the population average. This consistency is essential for comparative clinical analysis and regulatory auditing . SHAP is frequently applied to tabulate data like EHRs and genetics, as seen in risk prediction of cardiovascular events using molecular data .

However, the major limitation is computational complexity. Calculating the exact Shapley value requires evaluating the model times for a model with features, which is often infeasible for high-dimensional clinical data (e.g., millions of pixels in an image or thousands of genetic markers). This has led to the development of approximations:

- Kernel SHAP: A model-agnostic approximation that employs LIME’s local surrogate modeling approach but enforces the desirable Shapley properties. It samples feature coalitions instead of checking all of them.

- Tree SHAP: A highly optimized version for tree-based models (like XGBoost or Random Forest) that achieves exact calculation in polynomial time, making it much faster in applicable scenarios.

- Deep SHAP: An approximation method specifically for deep learning models that uses DeepLIFT to approximate the Shapley values.

The reliance on approximations implies the resulting SHAP value is an approximation rather than truly exact, introducing a potential source of error in the explanation, particularly in time-sensitive clinical deployment. The trade-off between computational burden and explanation fidelity is a persistent challenge in high-throughput healthcare settings.

3.3.1.2. LIME (Local Interpretable Model-agnostic Explanations)

LIME is a pioneering model-agnostic technique that focuses on providing an explanation that is locally faithful—meaning it accurately explains the model’s behavior around a specific prediction, even if the overall model is highly complex .

Mechanism and Theoretical Foundation:

LIME operates on the principle that while a complex model might be non-linear globally (e.g., a deep neural network), it can be approximated by a simple, interpretable model (e.g., linear regression or a shallow decision tree) within the local vicinity of a single data instance, .

The process for a single prediction is as follows:

1. Perturbation: LIME first generates a large number of perturbed, or slightly modified, data samples around the input instance . For image data, this might involve turning small sections of the image grey ; for text, it involves omitting words.

2. Prediction and Weighting: The complex "black box" model is used to predict the output for all these perturbed samples. Each perturbed sample is then weighted by its proximity to the original instance (closer samples receive higher weights), often using an exponential kernel.

3. Local Surrogate Model: An interpretable model () is trained on these weighted, perturbed samples and their predictions. The goal of is to minimize the loss (), which measures how well approximates , while keeping simple (minimizing its complexity,):

where is the black-box model, is the family of interpretable models, and is the proximity measure around . The resulting explanation is a simple, typically linear, model that explains the prediction of only within the small localized region defined by .

Clinical Utility and Limitations:

LIME’s strength is in its simplicity and speed, making it suitable for generating explanations on-the-fly for real-time diagnostics . Furthermore, its model-agnostic nature means it can be applied to virtually any AI in clinical use, from predicting mental disorder diagnosis to analyzing blood tests for COVID-19 . The output is a list of features (or segments, in the case of images) with weights indicating their local contribution to the specific outcome. Its application in retinoblastoma diagnosis demonstrated its ability to interpret deep learning models effectively .

The primary limitation of LIME is its inherent focus on local fidelity at the expense of global consistency. The explanation generated for a patient's diagnosis is only guaranteed to be accurate in the small neighborhood of that patient's data. If two patients have very similar profiles but one feature is slightly different, LIME might produce vastly different explanations that are both locally true but appear contradictory when compared side-by-side. This lack of global coherence can be confusing for clinicians seeking to establish general patterns or rules from the AI’s behavior. Furthermore, the selection of the size and weighting of the local neighborhood is non-trivial and can significantly influence the resulting explanation, leading to instability or lack of robustness in the explanations themselves.

3.3.1.3. GRAD-CAM (Gradient-weighted Class Activation Mapping)

Unlike the model-agnostic SHAP and LIME, GRAD-CAM is a model-specific technique designed to provide visual explanations for Convolutional Neural Networks (CNNs), making it the dominant XAI method in medical imaging .

Mechanism and Theoretical Foundation:

CNNs excel in medical imaging tasks, such as X-ray and CT interpretation, by processing data through multiple layers of convolutions to extract hierarchical features. GRAD-CAM answers the question: "Which regions in the input image were most important for the model's final classification decision (e.g., 'Tuberculosis' or 'Normal')?"

GRAD-CAM uses the gradients of the target concept (the predicted class score) flowing into the final convolutional layer of the CNN. The final convolutional layer is chosen because it retains rich spatial information about the input image, unlike the fully connected layers that follow, which often lose spatial resolution.

1. Gradient Calculation: The gradient of the class score () with respect to the feature map () of the last convolutional layer is calculated: .

2. Global Average Pooling: These gradients are then globally averaged across the spatial dimensions () to obtain a set of neuron importance weights () for the target class . This weight represents the importance of feature map for the decision .

3. Weighted Sum and ReLU: The final Class Activation Map () is obtained by performing a weighted sum of the forward activation maps () using the importance weights () and then applying a Rectified Linear Unit (ReLU) function. The ReLU ensures only features that positively influence the target class decision are highlighted:

The resulting map is then up-sampled to the resolution of the input image and overlaid as a heatmap.

Clinical Utility and Limitations:

GRAD-CAM's output is a high-resolution heatmap overlaid directly onto the input image, visually demonstrating the evidence base for the AI's diagnosis . This format is highly compatible with the established clinical workflow of radiologists and pathologists, being widely used in studies for lung cancer , mammography , and infectious disease detection . This direct visual evidence is strongly associated with increased clinical confidence and utility. The visual explanation allows the clinician to quickly verify if the AI is focusing on clinically relevant pathology or, conversely, if it is relying on spurious correlations (e.g., an unrelated metallic artifact or label), which serves as an important validation step.

The main limitation is its model-specificity: GRAD-CAM and its variants are inherently tied to the architecture of CNNs and cannot be readily applied to non-visual models like those built on transformer or recurrent architectures used for EHR analysis. Furthermore, its explanation relies on the coarseness of the final convolutional feature map, meaning the heatmaps are an approximation of focus rather than a pinpoint-accurate delineation of pixel-level importance. This potential lack of fine-grained detail can be insufficient in cases requiring microscopic or subtle pathological inspection.

3.3.2. Synthesis: XAI as a Driver of Trust and Diagnostic Accuracy

The evidence from the literature strongly suggests that the systematic application of these XAI techniques is a direct driver of improved clinical outcomes and greater trust. The 50 analyzed studies show that XAI is associated with improved trust, transparency, and diagnostic accuracy in medical imaging and disease prediction.

In the realm of imaging, the ability of GRAD-CAM to provide visual evidence is often sufficient to transform a skeptical clinician into a confident user. Studies on chronic wound classification explicitly pointed to the "highly transparent and explainable" nature of their XAI tool as essential for its clinical value .

For complex predictive tasks involving vast, heterogeneous datasets, SHAP's ability to provide globally consistent feature attribution is invaluable. This not only explains why an individual patient received a high-risk score but also allows clinical researchers to discern generalizable, population-level risk factors the AI is leveraging—factors that may not have been previously prioritized by traditional statistical methods. This process, as seen in oncology and cardiovascular research , transforms the AI from a predictive black box into an instrument for scientific and clinical discovery.

The cumulative evidence underscores the principle that transparency is strongly linked to accountability, which in turn is associated with improved system design. By revealing instances where the model makes a correct prediction for the wrong reason—a phenomenon sometimes referred to as "Clever Hans" behavior—XAI serves as a powerful validation tool that predicts and guides the development of more robust, generalizable, and therefore, more accurate AI systems . The use of XAI acts as a crucial safety net, particularly when dealing with the high heterogeneity and occasional bias present in medical datasets .

4. Discussion

4.1. Synthesis of Key Findings

This systematic review confirms the transformative role of XAI as the essential bridge between high-performance AI models and practical, ethical healthcare delivery. The synthesis of 50 contemporary studies demonstrates that XAI has been effectively deployed across the entire spectrum of clinical medicine, successfully addressing the "black box" concern that has long limited AI adoption in this sector. Techniques like SHAP, LIME, and GRAD-CAM have established themselves as the industry standards, providing the necessary interpretability to justify high-stakes clinical predictions and paving the way for trustworthy AI . The evidence is compelling: XAI significantly is associated with an increase in clinician confidence, improves transparency, and even acts as an internal auditing mechanism for model development.

4.2. Challenges to Real-World Clinical Integration

Despite the technological maturity and demonstrable benefits of XAI, its widespread integration into routine clinical practice remains constrained by several substantial hurdles. The review's synthesis highlights that technical and data-quality challenges—such as inconsistent validation, biased datasets, and fragmented explanation techniques—limit real-world integration .

- Inconsistent Validation and Reporting: A major challenge is the lack of standardized metrics to evaluate the quality of an explanation itself. Unlike model accuracy, which has clear metrics (e.g., AUC, F1-score), the "goodness" of an XAI explanation is often subjective or evaluated inconsistently. This heterogeneity hinders comparison between studies and impedes regulatory bodies from establishing clear thresholds for what constitutes a sufficient explanation . The reliance on qualitative assessment (e.g., "The clinician found the heatmap useful") rather than quantitative metrics (e.g., "The fidelity of the local explanation model was ") perpetuates this problem.

- Data Quality and Bias: AI models are only as robust as the data they are trained on. If training datasets are racially, socioeconomically, or geographically biased, the resulting XAI explanation will simply reflect and potentially amplify that bias . While XAI can identify the feature that led to a biased prediction, it does not fix the underlying data bias. Furthermore, high-quality, labeled clinical data is often fragmented, incomplete, or not uniformly collected, creating data-quality challenges that XAI cannot fully mitigate. Addressing data drift is also critical, especially in dynamic environments like emergency departments where population characteristics can change rapidly .

- Human-Centric Challenges: The way an explanation is presented is crucial. Clinicians and patients have different "mental models" of AI and require different levels and formats of explanation . An engineer might require the full SHAP value distribution, whereas a treating physician needs a concise, actionable summary of the most critical factors. If the explanation is too complex or poorly integrated into the EHR workflow, it is likely to be ignored, rendering the XAI technically present but functionally useless . The challenge lies in designing an explanation that is justifiable (technically accurate), intelligible (human-readable), and actionable (clinically useful).

- Computational and Technical Complexity: Generating post-hoc explanations, particularly with model-agnostic methods like SHAP, can be computationally intensive and time-consuming. In fast-paced clinical settings, where decisions must be made in seconds (e.g., in stroke diagnosis or intensive care monitoring), the latency introduced by XAI model execution can be a practical barrier to deployment. The constant demand for faster inference with simultaneous robust explanation requires sophisticated, high-performance computing infrastructure often unavailable in typical hospital settings.

4.3. Future Directions for XAI in Healthcare

To overcome the challenges outlined above and ensure the safe, ethical, and effective scaling of XAI in healthcare, future research and development must focus on three critical areas:

- Standardization and Regulatory Compliance: The most pressing need is to establish consistent, universally accepted frameworks for developing and reporting XAI findings. The review strongly suggests that future research must focus on standardized frameworks and regulatory compliance to ensure safe, ethical, and explainable AI use in healthcare . This includes defining minimum standards for explanation fidelity, robustness, and stability. Adherence to guidelines like DECIDE-AI , which focuses on bridging the gap between development and implementation, must become mandatory to ensure models are robustly tested before deployment. Regulatory bodies need technical guidance on how to evaluate XAI outputs to grant certification .

- Causal and Counterfactual Explanations: Current XAI primarily focuses on attributing the model’s prediction to input features (e.g., "Feature A is associated with this result"). Future work needs to shift toward causal inference and counterfactual explanations (e.g., "If Feature A had a value of X instead of Y, the result would have been Normal"). This type of explanation is far more valuable to a clinician, as it directly informs treatment or intervention strategies ("To change the outcome, what is the minimum required intervention?"). Such advancements will transform XAI from a simple justification tool into a proactive clinical recommendation system.

- Clinical Workflow Integration and Novel Data Streams: XAI outputs must be seamlessly embedded into existing clinical information systems without requiring clinicians to exit their primary EHR platform. This means developing intuitive, user-friendly interfaces that present explanations contextually, reducing cognitive load for clinicians . Furthermore, research must focus on XAI for complex, multi-modal, and continuous patient data streams (e.g., ECG , EEG , wearable sensors) common in critical care, rather than isolated, single-diagnosis scenarios. The explanation must be dynamic and update in real-time as patient status changes.

4.4. Limitations of the Systematic Review

This review, while comprehensive, is subject to standard limitations inherent to systematic literature synthesis. The strict inclusion of only peer-reviewed journal articles and archival papers may introduce a degree of publication bias, potentially favoring studies with positive or significant XAI findings over those reporting null or negative results. The search terms, though broad, may not have captured all relevant studies using non-standardized terminology for XAI, though model-agnosticism helped mitigate this. Finally, the synthesis is qualitative; given the heterogeneity of clinical domains, AI models, and XAI techniques across the 50 analyzed studies, a quantitative meta-analysis was not feasible. The conclusions drawn are based on the reported associations between XAI and outcomes, and do not constitute causal claims.

4.5. Conclusion

The movement toward XAI is not optional; it is the necessary next evolutionary step for AI in healthcare. XAI, particularly the prominent techniques like SHAP, LIME, and GRAD-CAM, has demonstrated its capacity to elevate AI performance from a mere prediction engine to a trustworthy, justifiable, and transparent clinical collaborator. While the technical and data-quality hurdles are substantial, a collective focus on regulatory standardization and user-centric explanation design will ensure that AI fulfills its promise to revolutionize patient care safely and ethically.

7. References

1. Salih Sarp, Murat Kuzlu, E. Wilson, U. Cali, and O. Guler, A Highly Transparent and Explainable Artificial Intelligence Tool for Chronic Wound Classification: XAI-CWC, Jan. 2021.

2. M. Merry, P. Riddle, and J. Warren, “A mental models approach for defining explainable artificial intelligence,” BMC Medical Informatics and Decision Making, vol. 21, no. 1, Dec. 2021.

3. J. A. Yeung, Y. Y. Wang, Z. Kraljevic, and J. T. H. Teo, Artificial intelligence (AI) for neurologists: do digital neurones dream of electric sheep?, Practical Neurology, vol. 23, no. 6, pp. 476–488, Dec. 2023.

4. Srilatha, S. (2025). Integrating AI into enterprise content management systems: A roadmap for intelligent automation. Journal of Information Systems Engineering and Management, 10(45s), 672–688. https://doi.org/10.52783/jisem.v10i45s.8904

5. Bohr and K. Memarzadeh, The rise of artificial intelligence in healthcare applications, Artificial Intelligence in Healthcare, vol. 1, no. 1, pp. 25–60, Jun. 2020.

6. T. Davenport and R. Kalakota, The Potential for Artificial Intelligence in Healthcare, Future Healthcare Journal, vol. 6, no. 2, pp. 94–98, Jun. 2019.

7. K. B. Johnson et al., Precision Medicine, AI, and the Future of Personalized Health Care, Clinical and Translational Science, vol. 14, no. 1, Oct. 2020, Available: https/www.ncbi.nlm.nih.gov/pmc/articles/PMC7877825/.

8. C. J. Kelly, A. Karthikesalingam, M. Suleyman, G. Corrado, and D. King, Key challenges for delivering clinical impact with artificial intelligence, BMC Medicine, vol. 17, no. 1, Oct. 2019.

9. J. Bajwa, U. Munir, A. Nori, and B. Williams, Artificial intelligence in healthcare: transforming the practice of medicine, Future Healthcare Journal, vol. 8, no. 2, pp. e188–e194, 2021.

10. S. A. Alowais et al., Revolutionizing healthcare: the role of artificial intelligence in clinical practice, BMC Medical Education, vol. 23, no. 1, Sep. 2023.

11. H. Alami et al., Artificial Intelligence and Health Technology Assessment: Anticipating a New Level of Complexity, Journal of Medical Internet Research, vol. 22, no. 7, p. e17707, Jul. 2020.

12. T. Raclin et al., Combining Machine Learning, Patient-Reported Outcomes, and Value-Based Health Care: Protocol for Scoping Reviews,” JMIR Research Protocols, vol. 11, no. 7, p. e36395, Jul. 2022.

13. Rangu, S. (2025). Analyzing the impact of AI-powered call center automation on operational efficiency in healthcare. Journal of Information Systems Engineering and Management, 10(45s), 666–689. https://doi.org/10.55278/jisem.2025.10.45s.666

14. B. Vasey et al., DECIDE-AI: new reporting guidelines to bridge the development-to-implementation gap in clinical artificial intelligence, Nature Medicine, vol. 27, no. 2, pp. 186–187, Feb. 2021.

15. Moher, D., Liberati, A., Tetzlaff, J., and Altman, D. G. (2009). Preferred reporting items for systematic reviews and metaanalyses: the PRISMA statement. Ann. Inter. Med. 151, 264–269.

16. Kwang Sig Lee and Eun Sun Kim, Explainable artificial intelligence in the early diagnosis of gastrointestinal disease, Diagnostics, vol. 12, no. 11, Nov. 2022.

17. A.-D. Samaras et al., Explainable classification of patients with primary hyperparathyroidism using

highly imbalanced clinical data derived from imaging and biochemical procedures, Applied Sciences, vol. 14, no. 5, p. 2171, Mar. 2024.

18. U. Pawar, S. Rea, Ruairi O'reilly, and D. O'shea, Incorporating explainable artificial intelligence (XAI) to aid the understanding of machine learning in the healthcare domain, 2020. Available: .

19. Sergiusz Wesołowski et al., An explainable artificial intelligence approach for predicting cardiovascular outcomes using electronic health records, PLOS Digital Health, vol. 1, no. 1, p. e0000004, Jan. 2022.

20. Chadha, K. S. (2025). Zero-Trust Data Architecture for Multi-Hospital Research: HIPAA-Compliant Unification of EHRs, Wearable Streams, and Clinical Trial Analytics. International Journal of Computational and Experimental Science and Engineering, 11(3). https://doi.org/10.22399/ijcesen.3477

21. Das and P. Rad, Opportunities and challenges in explainable artificial intelligence (XAI): A survey, Jun. 2020. Available: .

22. V. Sharma, Samarth Chhatwal, and B. Singh, An explainable artificial intelligence based prospective framework for COVID-19 risk prediction.

23. José Jiménez-Luna, F. Grisoni, and G. Schneider, "Drug discovery with explainable artificial intelligence," Nature Machine Intelligence, vol. 2, no. 10, pp. 573–584, Oct. 2020.

24. D. Dave, H. Naik, S. Singhal, and P. Patel, Explainable AI meets healthcare: A study on heart disease dataset, Nov. 2020. Available: .

25. J. Hoffmann et al., Prediction of clinical outcomes with explainable artificial intelligence in patients with chronic lymphocytic leukemia, Current Oncology, vol. 30, no. 2, pp. 1903–1915, Feb. 2023.

26. Khishigsuren Davagdorj, Jang Whan Bae, Van Huy Pham, Nipon Theera-Umpon, and Keun Ho Ryu, Explainable artificial intelligence based framework for non-communicable diseases prediction, IEEE Access, vol. 9, pp. 123672–123688, 2021.

27. Augusto Anguita-Ruiz, A. Segura-Delgado, R. Alcalá, C. M.1 Aguilera, and Jesús Alcalá-Fdez, EXplainable Artificial2 Intelligence (XAI) for the identification of biologically relevant3 gene expression patterns in longitudinal human studies, insights4 from obesity res5earch, PLoS Computational Biology, vol. 16, no. 4, Apr. 2020.

28. V. Roessner, J. Rothe, G. Kohls, Georg Schomerus, S. Ehrlich, andC. Beste, Taming the chaos?! Using eXplainable Artificial Intelligence (XAI) to tackle the complexity in mental health research, European Child and Adolescent Psychiatry, vol. 30, no. 8, pp. 1143–1146, Aug. 2021.

29. Jiten Sardana. (2025). Secure Messaging Protocols for Transactional Health Notifications. Utilitas Mathematica, 122(2), 267–290. Retrieved from https://utilitasmathematica.com/index.php/Index/article/view/2707

30. Salih Sarp, Murat Kuzlu, E. Wilson, U. Cali, and O. Guler,A highly transparent and explainable artificial intelligencetool for 6 Journal of Computer Sciences and Applications chronic wound classification: XAI-CWC, 2021.

31. Salman Muneer et al., An IoMT enabled smart healthcare model to monitor elderly people using Explainable Artificial Intelligence (EAI)., Journal of NCBAE, Vol 1.

32. Shaker El-Sappagh, J. M. Alonso, S. M.Riazul Islam, A.M. Sultan, and Kyung Sup Kwak, A multilayermultimodal detection and prediction model based on explainable artificial intelligence for Alzheimer's disease, Scientific Reports, vol. 11, no. 1, Dec. 2021.

33. Raza, Kim Phuc Tran, L. Koehl, and S. Li, "Designing ECG monitoring healthcare system with federated transfer learning and explainable AI," Knowl. Based Syst., vol. 236, p. 107763, 2021, Available: .

34. S. El-Sappagh, J. M. Alonso, S. M. R. Islam, A. M. Sultan, and K. S. Kwak, A multilayer multimodal detection and prediction model based on explainable artificial intelligence for Alzheimer's disease, Scientific Reports, vol. 11, no. 1, p. 2660, Jan. 2021.

35. J. Peng et al., An explainable artificial intelligence framework for the deterioration risk prediction of hepatitis patients, Journal of Medical Systems, vol. 45, no. 5, May 2021.

36. L. Lindsay, S. Coleman, D. Kerr, B. Taylor, and A. Moorhead, Explainable artificial intelligence for falls prediction, in Communications in Computer and Information Science, Springer, 2020, pp. 76–84.

37. Y. Jia, J. McDermid, T. Lawton, and Ibrahim Habli,"The role of explainability in assuring safety of machine learning in healthcare, IEEE Transactions on Emerging Topics in Computing, vol. 10, no. 4, pp. 1746–1760, Oct. 2022.

38. F. Vaquerizo-Villar et al., An explainable deep-learning model to stage sleep states in children and

propose novel EEG-related patterns in sleep apnea, Computers in Biology and Medicine, vol. 165, Oct. 2023.

39. F. Xu et al., The clinical value of explainable deep learning for diagnosing fungal keratitis using in vivo confocal microscopy images, Frontiers in Medicine, vol. 8, Dec. 2021.

40. Z. Naz, Muhammad, T. Saba, A. Rehman, Haitham Nobanee, and Saeed Ali Bahaj, An explainable AI-Enabled framework for interpreting pulmonary diseases from chest radiographs, Cancers, vol. 15, no. 1, Jan. 2023.

41. Belal Alsinglawi et al., An explainable machine learningframework for lung cancer hospital length of stay prediction, Scientific Reports, vol. 12, no. 1, Dec. 2022.

42. Esma Cerekci et al., Quantitative evaluation of saliency-based explainable artificial intelligence (XAI) methods in deep learning- based mammogram analysis, European Journal of Radiology, vol. 173, Apr. 2024.

43. Mohammed Saidul Islam, I. Hussain, Md Mezbaur Rahman, Se Jin Park, and Md Azam Hossain, Explainable artificial intelligence model for stroke prediction using EEG signal, Sensors, vol. 22, no. 24, Dec. 2022.

44. Z. U. Ahmed, K. Sun, M. Shelly, and L. Mu, Explainable artificial intelligence (XAI) for exploring spatial variability of lung and bronchus cancer (LBC) mortality rates in the contiguous USA, Scientific Reports, vol. 11, no. 1, Dec. 2021.

45. Samantapudi, R. K. R. (2025). Enhancing search and recommendation personalization through user modeling and representation. International Journal of Computational and Experimental Science and Engineering, 11(3), 6246–6265. https://doi.org/10.22399/ijcesen.3784

46. F. Ullah, J. Moon, H. Naeem, and S. Jabbar, Explainable artificial intelligence approach in combating real-time surveillance of COVID19 pandemic from CT scan and X-ray images using ensemble model, Journal of Supercomputing, vol. 78, no. 17, pp. 19246–19271, Nov. 2022.

47. F. Ahmed, M. Asif, M. Saleem, U. F. Mushtaq, and M. Imran, Identification and Prediction of Brain Tumor Using VGG-16 Empowered with Explainable Artificial Intelligence, International Journal of Computational and Innovative Sciences, vol. 2, no. 2, pp. 24–33, Jun. 2023, Available: ttps:// index.php/IJCIS/article/view/69.

48. Hussain and R. Jany, Interpreting stroke-impaired electromyography patterns through explainable artificial intelligence, Sensors, vol. 24, no. 5, Mar. 2024.

49. M. Westerlund, J. S. Hawe, M. Heinig, and Heribert Schunkert,Risk prediction of cardiovascular events by exploration of molecular data with explainable artificial intelligence, International Journal of Molecular Sciences, vol. 22, no. 19, Oct. 2021.

50. S. I. Nafisah and G. Muhammad, Tuberculosis detection in chest radiograph using convolutional neural network architecture and explainable artificial intelligence, Neural Computing and Applications, vol. 36, no. 1, pp. 111–131, Jan. 2024.

51. Sanjana, V. Sowmya, E. A. Gopalakrishnan, and K. P. Soman, Explainable artificial intelligence for heart rate variability in ECG signal, Healthcare Technology Letters, vol. 7, no. 6, pp. 146–154, Dec. 2020.

52. Schweizer et al., Analysing cerebrospinal fluid with explainable deep learning: From diagnostics to insights, Neuropathology and Applied Neurobiology, vol. 49, no. 1, Feb. 2023.

53. Gimeno et al., Explainable artificial intelligence for precision medicine in acute myeloid leukemia, Frontiers in Immunology, vol.13, Sep. 2022.

54. Anwer Mustafa Hilal et al., Modeling of explainable artificial intelligence for biomedical mental disorder diagnosis, Computers, Materials and Continua, vol. 71, no. 2, pp. 3853–3867, 2022.

55. Samanta Knapič, A. Malhi, R. Saluja, and K. Främling, Explainable artificial intelligence for human decision support system in the medical domain, Machine Learning and Knowledge Extraction, vol. 3, no. 3, pp. 740–770, Sep. 2021.

56. Q. Hu et al., Explainable artificial intelligence-based edge fuzzy images for COVID-19 detection and identification Applied Soft Computing, vol. 123, Jul. 2022.

57. Bader Aldughayfiq, F. Ashfaq, N. Z. Jhanjhi, and M. Humayun, Explainable AI for retinoblastoma diagnosis: Interpreting deep learning models with LIME and SHAP, Diagnostics, vol. 13, no. 11, Jun. 2023.

58. Jeong Kyun Kim, Myung Nam Bae, K. Lee, Jae Chul Kim, and Sang Gi Hong, Explainable artificial intelligence and wearable sensor-based gait analysis to identify patients with osteopenia and sarcopenia in daily life, Biosensors, vol. 12, no. 3, Mar. 2022.

59. T. Mahmud, K. Barua, Sultana Umme Habiba, Nahed Sharmen, Mohammad Shahadat Hossain, and K. Andersson, An explainable AI paradigm for alzheimer's diagnosis using deep transfer learning, Diagnostics, vol. 14, no. 3, Feb. 2024.

60. S. D. Mohanty, D. Lekan, T. P. McCoy, M. Jenkins, and P. Manda, Machine learning for predicting readmission risk among the frail: Explainable AI for healthcare, Patterns, vol. 3, no. 1, Jan. 2022.

61. J. Ma et al., Towards trustworthy AI in dentistry, Journal of Dental Research, vol. 101, no. 11, pp. 1263–1268, Oct. 2022.

62. C. Duckworth et al., Using explainable machine learning to characterize data drift and detect emergent health risks for emergency department admissions during COVID-19, Scientific Reports, vol. 11, no. 1, Dec. 2021.

63. Merry, P. Riddle, and J. Warren, A mental models approach for defining explainable artificial intelligence, BMC Medical Informatics and Decision Making, vol. 21, no. 1, Dec. 2021.

64. Aslam, Explainable artificial intelligence approach for the early prediction of ventilator support and mortality in COVID-19 patients, Computation, vol. 10, no. 3, Mar. 2022.

65. L. M. Thimoteo, M. M. Vellasco, J. Amaral, K. Figueiredo, Cátia Lie Yokoyama, and E. Marques, Explainable artificial intelligence for COVID-19 diagnosis through blood test variables, Journal of Control, Automation and Electrical Systems, vol. 33, no. 2, pp. 625–644, Apr. 2022.

66. Moreno-Sánchez, Improvement of a prediction model for heart failure survival through explainable artificial intelligence, Frontiers in Cardiovascular Medicine, vol. 10, 2023.

67. Salih Sarp, Murat Kuzlu, E. Wilson, U. Cali, and O. Guler, The enlightening role of explainable artificial intelligence in chronic wound classification, Electronics (Switzerland), vol. 10, no. 12, Jun. 2021.