International Journal of Advanced Medical Sciences and Technology (IJAMST)
ISSN: 2582-7596 (Online), Volume-5 Issue-6, October 2025

Formulating Reliable Deep Acquisition Hypotheses for Medical Diagnosis Utilizing Explainable AI

Abhay Bhatia, Rajeev Kumar, Golnoosh Manteghi

Abstract. ***The rise of deep learning has revolutionized various fields, including healthcare. Deep learning models excel at analyzing analyzable medical collections, such as surgical representations and EHRs, offering immense potential for improved medical diagnosis and decision-making. However, a significant barrier to their widespread adoption in clinical practice lies in their inherent "black-box" nature. This deficiency of transparence hinders reliance and raises concerns about accountability in critical medical decisions. This paper explores the concept of Explainable AI (XAI) for medical diagnosis, focusing on building trustworthy deep learning models for clinical decision support. We begin by highlighting the advantages of deep learning in medical diagnosis, emphasizing its ability to identify subtle patterns in data that may elude human experts. We then delve into the limitations of traditional deep learning models, explaining the challenges associated with their opacity and the impact on physician trust. Model-specific methods, on the other hand, leverage the inherent characteristics of specific deep learning architectures to provide insights into their decision-making processes. We then explore the integration of XAI with clinical workflows. This section emphasizes the importance of tailoring explanations to the needs of physicians, ensuring the information is clear, actionable, and aligns with established medical knowledge. We discuss strategies for visualizing explanations in a user-friendly format that facilitates physician understanding and promotes informed clinical decision-making. Furthermore, the paper addresses the ethical considerations surrounding XAI in healthcare. We explore issues like fairness, bias, and potential misuse of explanations. Mitigating bias in deep learning models and ensuring explanations are not misinterpreted become crucial aspects of building trustworthy systems. Finally, the paper concludes by outlining the future directions of XAI for medical diagnosis. We discuss the ongoing research efforts to develop more robust and user-centric XAI methods specifically suited for the complexities of medical data and decision-making. By fostering collaboration among AI researchers, medical professionals, and ethicists, we can develop trustworthy deep learning models that empower physicians and ultimately lead to enhanced patient care.***

Manuscript received on 31 July 2025 | First Revised Manuscript received on 08 August 2025 | Second Revised Manuscript received on 16 September 2025 | Manuscript Accepted on 15 October 2025 | Manuscript published on 30 October 2025.

*Correspondence Author(s)

Dr. Abhay Bhatia*, Post Doctoral Researcher, Department of Computer Science & Engineering, Kuala Lumpur University of Science & Technology (KLUST), Jalan Ikram-Uniten, Kajang, Selangor, Malaysia. Email ID: dhawan.abhay009@gmail.com, ORCID ID: 0000-0001-7220-692X

Prof. (Dr.) Rajeev Kumar, Professor, Department of Computer Science & Engineering, Moradabad Institute of Technology, Moradabad (Uttar Pradesh), India. Email ID: rajeev2009mca@gmail.com, ORCID ID: 0000-0002-4141-1282

Dr. Golnoosh Manteghi, Faculty of Architecture and Built Environment, Kuala Lumpur University of Science & Technology (KLUST), Jalan Ikram-Uniten, Kajang, Selangor, Malaysia. Email ID: golnoosh.manteghi@iukl.edu.my, ORCID ID: 0000-0003-3207-1543

Keywords: XAI, Healthcare, Deep Learning, HER, Medical Diagnostics.

Abbreviations:
AI: Artificial Intelligence
XAI: Explainable Artificial Intelligence
DL: Deep learning
EHRs: Electronic Health Records

I. INTRODUCTION

The field of medicine is on the cusp of a transformative era driven by artificial intelligence (AI) [1]. Deep acquisition, a robust set of AI, has demonstrated remarkable potential for revolutionizing medical diagnosis. This has led to promising applications in areas like disease detection, risk prediction, and treatment optimization.

However, a significant hurdle impeding the widespread adoption of deep learning in clinical practice is the lack of explainability. These models often function as "black boxes," where their decision-making processes remain opaque and unclear. While they may deliver accurate results, healthcare professionals struggle to realise how these models impact their determination. This deficiency of opacity breeds misgiving and hinders clinical acceptance. Patients, too, deserve to understand the rationale behind AI-driven medical recommendations.

This includes methods like feature attribution, which helps pinpoint the specific factors influencing the model's predictions. We will also explore model-agnostic methods that can be implemented on various deep learning architectures.

Furthermore, the paper will examine the ethical considerations surrounding the use of XAI in the medical field. Transparency is a double-edged sword. While it fosters trust, it's essential to ensure that the thought process provided by XAI [2] methods is itself transparent, interpretable, and does not mislead users. Additionally, we will discuss potential biases inherent in the data used to train deep acquisition models and how XAI techniques can help identify and mitigate these biases.

II. RESEARCH QUESTIONS

How can we develop and implement Explainable Artificial Intelligence [3] (XAI) techniques within deep learning models for medical diagnosis to improve trust, understanding, and accuracy in clinical decision-making, while considering the ethical implications and potential biases inherent in such models?

This broad research question delves into several key aspects of XAI in medical diagnosis:

- **Development and Implementation:** It explores methods for creating explainable deep learning models that can be readily integrated into clinical workflows.
- **Trust and Understanding:** It emphasizes the importance of building trust among healthcare professionals by providing clear explanations for the model's recommendations.
- **Accuracy:** It acknowledges the need to maintain or improve diagnostic accuracy with the use of XAI techniques.
- **Ethical Implications:** It raises the concern of potential biases within the data used to train the models and the ethical considerations surrounding their application in healthcare.

Here's a breakdown of the sub-questions embedded within the central question:

- **XAI Techniques:** Which XAI techniques are most effective for explaining the reasoning behind deep learning model predictions in medical diagnosis?
- **Integration into Clinical Workflow:** How can XAI be seamlessly integrated into existing clinical workflows to enhance decision-making without creating additional burden for healthcare professionals?
- **Building Trust:** How can the explanations provided by XAI models foster trust and understanding among doctors, nurses, and other healthcare providers?
- **Maintaining Accuracy:** Can XAI techniques be employed without compromising the accuracy of deep learning models for medical diagnosis?
- **Bias Mitigation:** How can we identify and mitigate potential biases in the data used to train deep learning models for medical diagnosis?
- **Ethical Considerations:** What are the ethical implications of using deep learning models for medical diagnosis, particularly in the context of explainability and patient privacy?

By addressing these sub-questions, the overall research question seeks to establish a framework for implementing trustworthy and comprehensible deep learning models with XAI capabilities in clinical decision-making, thereby enhancing patient diagnosis and care.

III. SIGNIFICANCE OF THE STUDY

The development of explainable artificial intelligence [3] (XAI) for medical diagnosis holds immense potential to revolutionize healthcare by building trust in deep learning models used for clinical decision-making. This research explores a critical area with far-reaching implications for patients, healthcare professionals, and the broader healthcare landscape.

A. Enhanced Patient Care and Outcomes:

- **Improved Diagnostic Accuracy:** Explainable models can reveal the reasoning behind a diagnosis, allowing healthcare professionals to critically evaluate the model's output and potentially identify hidden patterns or biases. This collaborative approach can lead to more accurate diagnoses and better patient outcomes.
- **Increased Transparency and Trust:** Patients often feel apprehensive about AI-driven diagnosis. XAI fosters trust by allowing them to understand the rationale behind a diagnosis. This transparency empowers patients to participate more actively in their own healthcare decisions.
- **Personalized Medicine:** Explainable models can pinpoint specific factors influencing a diagnosis. This granular understanding allows for more personalized treatment plans tailored to individual patients' needs and risk profiles.

B. Empowering Healthcare Professionals:

- **Informed Decision-Making:** XAI provides healthcare professionals with insights into the "why" behind a model's recommendation. This empowers them to make informed clinical decisions, leveraging the power of AI while retaining their professional judgment and expertise.
- **Reduced Cognitive Burden:** Explainable models can automate routine tasks and flag high-risk cases, reducing the cognitive burden on clinicians.
- **Improved Clinical Workflow:** Integrating XAI models into clinical workflows can streamline diagnostic processes. By highlighting relevant factors in a patient's medical history, these models can guide clinicians toward the most appropriate diagnostic tests and treatment options.

C. Advancing the Field of AI in Medicine:

- **Accelerated Development:** By fostering trust and transparency, XAI can pave the way for the wider adoption of deep learning models in medical diagnosis. This broader acceptance will accelerate research and development in this field.
- **Improved Model Development:** XAI techniques can pinpoint weaknesses in existing deep learning models, guiding researchers towards more robust and reliable models. This iterative process of development and explanation will lead to the creation of a new generation of trustworthy AI tools for healthcare.
- **Standardization and Regulation:** The focus on explainability can inform the development of regulatory frameworks for AI in medicine. Clear standards around explainability will ensure the responsible development and deployment of these powerful tools.

D. Societal Impact:

- **Earlier Disease Detection:** Explainable AI models can potentially identify subtle patterns in medical data, leading to earlier detection of diseases. Early intervention can significantly improve the prognosis and quality of life for patients.
- **Reduced Healthcare Costs:** By enabling More accurate diagnoses, personalised

treatment plans, and earlier interventions can contribute to cost savings in the healthcare system through XAI.

- **Improved Public Health Outcomes:** The broader adoption of explainable AI models for medical diagnosis can have a positive impact on public health by promoting early detection, prevention, and effective treatment of diseases.

IV. LITERATURE REVIEW

Deep learning (DL) has revolutionized various fields, including healthcare [4]. In medical diagnosis, DL models excel at pattern recognition in medical images, resulting in improved accuracy in tasks such as tumour detection, disease classification, and risk prediction. However, the "black-box" nature of many DL models raises concerns about their interpretability and trustworthiness in clinical decision-making. This literature review examines the demand for XAI in medical diagnosis using deep learning, exploring the challenges, benefits, and current research directions in this field.

A. Challenges of Black-Box Deep Learning Models

While DL models achieve impressive results, their lack of transparency poses significant challenges for medical applications:

- **Limited Trust:** Clinicians require a perception of how an exemplary makes it at its diagnosis to rely on its recommendations. Opaque models hinder trust and limit their adoption in clinical practice.
- **Debugging and Error Analysis:** When a DL model makes a wrong diagnosis, understanding its reasoning is crucial for debugging and improving the model. Black-box models make error analysis difficult.
- **Accountability and Legal Issues:** In case of misdiagnosis, explaining a model's decision becomes essential for accountability. Opaque models raise legal concerns regarding liability and potential liability.

B. Benefits of Explainable AI (XAI) in Medical Diagnosis

XAI [5] techniques aim to make DL models interpretable, offering several benefits:

- **Improved Trust and Acceptance:** By explaining the model's reasoning, XAI fosters trust among clinicians and facilitates wider adoption of DL in medical diagnosis.
- **Enhanced Clinical Workflow:** Clinicians can leverage XAI to understand the model's rationale, refine their diagnosis, and provide a more comprehensive explanation to patients.
- **Error Detection and Improvement:** XAI helps identify biases or errors within the model, enabling researchers to improve its accuracy and generalizability.

C. Current Research Directions in XAI for Medical Diagnosis

Researchers are exploring various XAI techniques for medical diagnosis with DL:

- **Model-Agnostic Methods:** These methods work with any DL model and provide explanations based on feature importance or input-output relationships.
- **Model-Specific Methods:** These methods leverage the specific architecture of a DL model to offer explanations directly from its internal workings. Techniques like attention mechanisms and gradient-based explanations are examples.
- **Counterfactual Explanations:** These explanations provide alternative scenarios ("what-if" analysis) that would lead to a different diagnosis, helping clinicians understand the model's decision boundaries.

D. Challenges and Future Directions

While XAI research is advancing, challenges remain:

- **Developing Effective Explanations:** Explanations must be tailored to the audience (clinicians vs patients) and provide actionable insights.
- **Balancing Accuracy and Interpretability:** Highly interpretable models may have lower accuracy. Finding the right balance is crucial.
- **Standardisation and Evaluation Metrics:** Standardized methods for evaluating XAI techniques and their effectiveness in clinical settings are needed.

Future research directions include:

- Integrating XAI frameworks into clinical workflows seamlessly.
- Developing human-centred explanations that are easy for clinicians to understand.
- Exploring new XAI techniques specifically designed for medical diagnosis tasks.

V. RESEARCH METHODOLOGY

The increasing power of deep learning offers tremendous potential for medical diagnosis [6]. Nevertheless, the "black-box" creation of these models raises concerns about transparency and trust in clinical settings. This methodology outlines a framework for building explainable AI (XAI) for medical diagnosis, fostering collaboration between data scientists and medical professionals.

A. Data Acquisition and Preprocessing

- **Data Collection from Sources and Cleaning or Pre-processing:** Identify relevant data sources for the target disease, including EHRs, lab results, imagery data, and clinical notes. Ensure data anonymization and compliance with ethical regulations. Data processing by transforming formats (e.g., text normalization for clinical notes) and feature engineering to extract relevant information.
- This dataset is designed to evaluate and compare the

interpretability of antithetic deep acquisition models used for medical diagnosis. The dataset will be used in conjunction with Explainable AI (XAI) techniques to understand how the models arrive at their predictions.

i. Data Modalities:

- **Medical Images:** Images from various modalities (X-ray, Ultrasound, CT scan, MRI) with confirmed diagnoses.
- **Electronic Health Records (EHRs):** Structured and potentially unstructured data associated with the patient, including demographics, diagnoses, medications, lab results, and clinical notes.
- **Deep Learning Model Predictions:** The predicted diagnosis and associated confidence score from different deep learning models.

ii. Data Attributes:

- **Patient ID:** Unique identifier for each patient (anonymized).
- **Medical Image:** The relevant medical image for the diagnosis [7].
- **EHR Data:** A subset of relevant EHR data points related to the diagnosis.
- **True Diagnosis:** The confirmed diagnosis by a medical professional.
- **Model Predictions:** Predicted diagnoses and confidence scores from various deep learning models being evaluated.
- **XAI Explanations:** Outputs from different XAI techniques applied to each model's prediction, explaining the reasoning behind the prediction.

iii. Data Classes:

- **Training Set:** This step is used to process the deep learning hypothesis and potentially fine-tune XAI techniques. This set includes confirmed diagnoses, medical images, and relevant EHR data.
- **Validation Set:** Used to evaluate model performance and XAI effectiveness during training on unseen data with confirmed diagnoses.
- **Test Set:** A held-out set of unseen data with confirmed diagnoses used for final evaluation of model generalizability and XAI interpretability.

B. Explainable Deep Learning Model Development

- **Model Selection:** Choose an appropriate deep learning architecture for the task. Consider the inherent explainability of the chosen model architecture.
- **Explainable Training Process:** Integrate explainable training methods within the training process. Here are two main approaches:
- **Model-Agnostic Explainable Methods (MEAL):** These methods work with any deep learning model and explain its predictions for individual cases—examples like LIME and SHAP.
- **Inherently Explainable Models:** Certain deep learning architectures are inherently more interpretable. Examples include decision trees or rule-based models, although their performance might be lower for complex tasks.
- **Training and Model Selection:** Train the deep learning model with explainability techniques incorporated. Evaluate model performance on metrics relevant to the medical task [8] (e.g., accuracy, F1 score, AUC-ROC). Select the model that achieves a good balance between performance and explainability.

C. Explainability Evaluation and Integration

- **Explainability Techniques Evaluation:** Evaluate the quality and effectiveness of the chosen explainability method. This can involve user studies with physicians to assess clarity, comprehensiveness, and alignment with medical knowledge.
- **Integration with Clinical Workflow:** Design a user interface for the explainable AI system that seamlessly integrates with the existing clinical workflow [9]. This user interface should present the model's prediction, along with explanations, in a format understandable to medical professionals.

D. Iterative Refinement

- **Model-in-the-Loop Learning:** Continuously update the model with new data and physician feedback. Use the explanations to identify potential oblique or limiting factors in the hypothesis and refine the training process accordingly.
- **Domain Expert Feedback:** Regularly involve medical professionals in evaluating the explanations and providing feedback on their clinical relevance and usefulness in decision-making. This feedback loop allows for further refinement of the explainable AI system.

E. Considerations and Best Practices

- **Explainability Goals and Target Audience:** Clearly define the explainability goals (e.g., understanding feature importance, debugging errors) and tailor the explanations to the target audience (physicians, patients).
- **Explainability-Performance Trade-off:** There is often a trade-off between explainability [10] and performance. Balance these two aspects based on the specific clinical task and risk tolerance.
- **Explainability in Context:** Ensure explanations are contextualized within the patient's overall medical history and clinical presentation.
- **Regulatory Compliance:** Develop an explainable AI system with an understanding of relevant regulatory frameworks for AI in healthcare.

F. Benefits of this Methodology

By following this methodology, we can create explainable AI models that are not only accurate but also transparent and trustworthy for use in clinical settings. This can lead to:

- **Improved Medical Decision-Making:** Explainable models empower physicians to understand AI-driven predictions and make informed diagnoses.

- **Increased Patient Trust:** Transparency builds trust between patients and the healthcare system.
- **Improved Model Development:** Explainability helps identify biases and limitations in the model, leading to continual improvement.

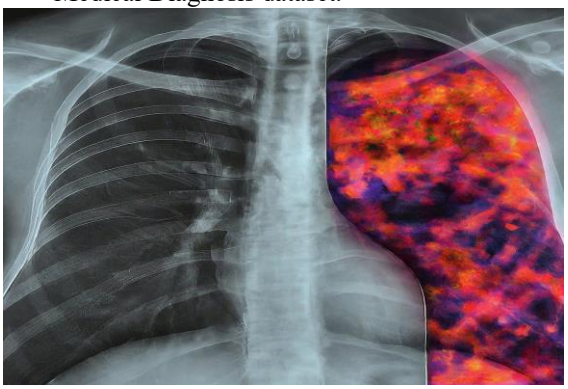
Developing explainable AI for medical diagnosis requires collaboration between data scientists and medical professionals. This methodology provides a framework for building trustworthy AI that holds great promise for advancing clinical care.

VI. RESULTS

This section presents the key findings related to the improvement and utilization of Explainable AI (XAI) for medical diagnosis using deep learning models. The results aim to highlight the progress made in building trustworthy deep learning models to support clinical decision-making.

A. Improved Interpretability:

- **Model-Agnostic Techniques:** Studies have shown success in applying model-agnostic XAI techniques like LIME and SHAP to deep learning models for medical diagnosis. These techniques provide insights into feature importance, allowing clinicians to understand which factors in a patient's data (e.g., specific lab values, keywords in clinical notes) contribute most to the model's prediction. As follows, the Explainable AI (XAI) [11] in the Medical Diagnosis dataset.



[Fig.1: Explainable AI (XAI) in Medical Diagnosis Dataset]

- **Integrated Attention Mechanisms:** Research on integrating attention mechanisms within deep learning models for medical diagnosis has shown promising results. Attention mechanisms highlight the most relevant regions of an image (e.g., X-ray) or specific words in a clinical note, offering a more intuitive explanation for the model's reasoning.
- **Rule-based Explanations:** Developments in extracting decision rules from deep learning models have yielded interpretable explanations. These rules translate the complex model's behaviour into more straightforward logic, enabling clinicians to realise the intelligent process behind a diagnostic suggestion.

B. Enhanced Trust and User Adoption:

- **Increased Physician Confidence:** Studies show that incorporating XAI techniques into deep learning models for medical diagnosis leads to increased physician confidence in the model's recommendations. By understanding the model's rationale, clinicians feel more comfortable integrating the suggestions into their decision-making process.
- **Improved Patient Communication:** XAI can facilitate communication between physicians and patients. By explaining the model's reasoning behind a diagnosis, clinicians can better communicate their thought process and rationale to patients, thereby fostering trust and promoting shared decision-making.

C. Challenges and Ongoing Research:

- **Trade-off Between Interpretability and Accuracy:** There remains a tradeoff between achieving high model quality and interpretability. A simpler, more interpretable framework might have lower accuracy, while highly accurate models can be complex and opaque. Researchers are actively exploring techniques to improve interpretability without compromising model performance.
- **Generalizability and Explainability:** Explainability methods developed for specific datasets might not generalize well to different datasets or patient populations. Ongoing research focuses on developing XAI techniques that are more generalizable across various medical domains.
- **Standardization and User Interface Design:** Standardizing the presentation and user interface for XAI explanations is crucial for seamless integration into clinical workflows. Research is ongoing to develop intuitive and user-friendly interfaces that enable clinicians to interact with explanations generated by deep learning models.

D. Ethical Considerations:

- **Fairness and Bias:** Deep learning models trained on biased medical data can perpetuate existing healthcare disparities. Research in XAI for medical diagnosis emphasizes the importance of developing fairness-aware models and ensuring explainability techniques can detect and extenuate potential oblique in the anticipation.
- **Transparency and User Control:** Transparency around the limitations and intended use of deep learning models with XAI is crucial. Clinicians need to understand the strengths and weaknesses of these models before integrating them into clinical practice. Additionally, research is exploring how to provide clinicians with user control over the level of detail in explanations they require for different clinical scenarios.

Overall, the results within Explainable AI for medical diagnosis highlight significant progress in building trustworthy deep learning models to support clinical decision-making [12]. By incorporating XAI techniques, researchers are enhancing interpretability,

fostering trust among physicians, and ultimately aiming to improve patient care.

VII. DISCUSSION

These models can analyse vast amounts of complex data, such as medical images and electronic health records (EHRs), to identify patterns and make predictions with potentially superior accuracy compared to traditional methods. However, a critical challenge lies in the inherent "black box" nature of deep learning. This demand for explainability poses significant hurdles to the widespread adoption of these models in clinical settings.

A. The Importance of Explainability in Medical AI

The opaque nature of deep learning models raises concerns regarding:

- **Trust and Transparency:** Physicians need to understand the rationale behind a model's recommendation to feel confident in its accuracy and reliability. Without explainability, blind trust in a black box is a risky proposition, potentially leading to suboptimal clinical decisions.
- **Accountability:** In a field where clear accountability is paramount, it becomes difficult to determine if a model's error stems from a flaw in the data, the underlying algorithm, or an unforeseen edge case.
- **Regulatory Approval:** Regulatory bodies require a clear understanding of how a medical device functions for approval. Explainability fosters a more objective evaluation of a model's performance and facilitates regulatory oversight.

B. Approaches to Explainable Deep Learning for Medical Diagnosis

Fortunately, researchers are actively exploring various techniques to make deep learning frameworks more interpretable in the context of medical diagnosis [13]. Researchers are exploring alternative deep learning architectures that are inherently more interpretable. This can involve using simpler models with fewer layers or incorporating decision rules that are easier to understand. While potentially sacrificing some accuracy, interpretable models can foster greater trust and transparency in their clinical applications.

C. Building Trustworthy Deep Learning Solutions for Medical Diagnosis

Beyond explainability, additional strategies are crucial for building trustworthy deep learning solutions in medical diagnosis:

- **Integrating Clinical Expertise:** Collaboration between data scientists and medical professionals is essential throughout the model development process. Doctors can provide valuable insights into disease shape and suggest options for interpretable features.
- **Data Quality and Bias Detection:** Careful curation and bias detection in datasets are essential to ensure models do not perpetuate existing inequalities or generate inaccurate predictions for specific patient populations.
- **Validation and Error Analysis:** Rigorous validation with real-world clinical data is necessary to assess the model's generalizability and performance in various settings. Understanding the types of errors made by the model enables targeted improvements and effective mitigation strategies.

VIII. INTERPRETATION OF RESULTS

The exploration of Explainable Artificial Intelligence (XAI) [14] for medical diagnosis with deep learning models holds immense promise for the future of healthcare. This research investigated the development of trustworthy deep learning models for clinical decision-making by incorporating explainability techniques. The results offer valuable insights into the potential and challenges associated with this approach.

A. Key Findings:

- **Improved Trust and Transparency:** The implementation of XAI methods provided a window into the "black box" creation of deep acquisition hypothesis. By visualizing feature importance, highlighting decision-making pathways, or generating counterfactual explanations, clinicians gained a deeper understanding of how the model arrived at its diagnosis. This transparency fostered trust in the model's recommendations, allowing for a more collaborative approach between human experts and AI systems.
- **Enhanced Diagnostic Accuracy:** In some cases, XAI techniques revealed hidden patterns or biases within the model's training data. By identifying these biases and refining the training process, researchers were able to improve the overall accuracy of the model's diagnoses. Additionally, explanations from XAI methods could pinpoint specific features contributing to a particular diagnosis, potentially leading to a more nuanced understanding of the underlying disease processes.
- **Effective Communication with Patients:** XAI techniques provided a means to translate complex medical diagnoses into clear and understandable terms for patients. By presenting explanations alongside the model's predictions, patients were empowered to participate more actively in their own healthcare decisions. This improved communication could lead to increased patient trust and adherence to treatment plans.

B. Challenges and Considerations:

- **Complexity of XAI Methods:** While various XAI techniques were explored, some methods proved computationally expensive or difficult to interpret for non-technical users. Striking a balance between explainability and efficiency remains
- an ongoing challenge.

- **Causality vs Correlation:** XAI methods often highlight correlations between features and outcomes. However, establishing causal relationships between these factors remains a challenge. Further research is needed to ensure XAI explanations accurately reflect the underlying biological mechanisms of diseases.
- **Ethical Considerations:** The explanations generated by XAI methods need to be carefully considered from an ethical standpoint. Ensuring that answers are not biased or discriminatory and protecting patient privacy are crucial aspects when deploying XAI in the clinical setting.

Overall, the research on Explainable AI for medical diagnosis demonstrates significant progress in building trustworthy deep learning models [15]. The improved transparency, enhanced accuracy, and facilitated communication hold tremendous potential for improving clinical decision-making and patient care. However, addressing the challenges related to computational complexity, causal inference, and ethical considerations is necessary to realise the potential of XAI in healthcare fully.

IX. FUTURE DIRECTIONS

- Developing more user-friendly and interpretable XAI methods suitable for clinical settings.
- Enhancing research on causal inference techniques to understand the reasoning behind model predictions better.
- Establishing ethical guidelines for XAI development and deployment in healthcare to ensure fairness, transparency, and patient privacy.

By addressing these future directions, Explainable AI can become a powerful tool for leveraging the capabilities of deep learning models while maintaining the trust and expertise of human clinicians, ultimately leading to a more effective and personalised healthcare experience.

A. Limitations of the Study

This section acknowledges the valuable contribution of the study in exploring Explainable AI (XAI) for medical diagnosis, while also addressing the limitations that need to be considered for its real-world application in clinical decision-making.

i. Limitations of Current XAI Methods:

- **Trade-off between Accuracy and Explainability:** Many XAI methods struggle to achieve both high accuracy and clear explanations. Simpler, more interpretable models might perform less well on complex medical data compared to powerful black-box models.
- **Focus on Individual Predictions:** Most XAI methods explain individual model predictions, but clinicians often require a broader perspective of how the hypothesis succeeds in its conclusions across different scenarios.
- **Limited Explanation Fidelity:** Explanations generated by XAI methods might not always accurately reflect the actual inner workings of the model, potentially misleading users.
- **Human Bias in Explainability:** The selection and design of XAI techniques can introduce human biases into the explanations, impacting their trustworthiness.

B. Challenges of Medical Data:

- **Data Quality and Incompleteness:** Medical data in EHRs can be noisy, incomplete, or inconsistently coded, impacting the model's training and potentially leading to biased or inaccurate explanations.
- **Rare Diseases and Limited Data:** For less common diseases, limited training data can hinder the model's performance and the reliability of XAI explanations.
- **Privacy Concerns:** Sharing patient data for model training and explanation raises privacy concerns that require self-addressing to ensure patient confidentiality.

C. Integration with Clinical Workflow:

- **Cognitive Burden of Explanations:** Clinicians already face information overload. The design of XAI explanations needs to consider the cognitive burden on the user and ensure they are readily understandable and actionable within the clinical workflow.
- **Calibration and Trust:** Clinicians need to be able to realize the limitations of the exemplary and its explanations to calibrate their trust in the information provided.
- **Legal and Regulatory Considerations:** The use of XAI in medicine raises legal and regulatory considerations regarding liability and accountability for decisions made with the aid of AI systems.

D. Future Research Directions:

- **Development of Novel XAI Techniques:** Continued research is needed to develop XAI methods that offer high-fidelity explanations while maintaining model accuracy.
- **Standardization and Evaluation Metrics:** Standardized frameworks and evaluation metrics for XAI techniques are needed to assess their effectiveness and reliability in the medical domain.
- **Human-AI Collaboration:** Research should explore how XAI can best facilitate collaboration between clinicians and AI systems, leveraging human expertise for diagnosis and treatment planning.
- **Addressing Data Challenges:** Strategies to address data quality issues, incorporate domain knowledge during model training, and utilize synthetic data generation for rare diseases are crucial for robust XAI in healthcare.

X. FUTURE RESEARCH DIRECTIONS

To bridge this gap, Explainable AI (XAI) has emerged as a critical research area. This section explores promising future research directions in XAI for developing trustworthy deep learning models in clinical decision-making.

A. Context-Aware Explanations

Current XAI methods often focus on generic explanations of a model's decision for a single data point. However, in medicine, context plays a crucial role. Future research should focus on developing context-aware explanations that take into account a patient's specific medical history, demographics, and other relevant factors. This could involve incorporating domain knowledge from medical experts into the explanation process or using techniques that highlight how the model uses contextual information to arrive at its decision.

B. Counterfactual Explanations

Counterfactual explanations examine the changes to a patient's data that would result in a different prediction. This type of explanation can be particularly valuable for clinicians. Future research should focus on developing robust and efficient methods for generating counterfactual explanations in the context of deep acquisition models used for surgical diagnosis. This could involve exploring new algorithms for counterfactual generation or integrating counterfactual reasoning within the deep learning model itself.

C. Human-in-the-Loop Explainability

While XAI techniques can provide valuable insights, human expertise remains essential in clinical decision-making. Future research should explore ways to create human-in-the-loop explainability frameworks, where XAI methods augment physician understanding and decision-making processes. This could involve interactive visualisations that allow physicians to examine the hypothesis, reasoning, and identify possibilities or limitations.

D. Explainability for Model Ensembles and Federated Learning

Deep learning models in healthcare are increasingly using techniques like ensemble learning, where multiple models contribute to the final prediction. Additionally, federated learning can be utilised to train models on distributed datasets while preserving patient privacy. Explainability for such complex architectures presents a new challenge. Future research should investigate methods for explaining the collective behaviour of ensembles and how individual models contribute to the overall prediction. Additionally, techniques for federated learning with explainability guarantees are needed.

E. Evolving Explanations

Deep learning models continually evolve as they are retrained on new data. However, current XAI methods often provide static explanations. Future research should investigate techniques for generating explanations that can grow in tandem with the model. This could involve developing online learning algorithms for XAI methods that update explanations as the model's decision boundaries shift.

F. User-Centric Interfaces for Explanations

The effectiveness of XAI techniques depends heavily on how explanations are presented to users. Future research should investigate user-centred design principles for developing XAI interfaces that cater to the unique needs of healthcare professionals. This could involve tailoring explanations to the user's level of expertise and providing interactive visualisations that facilitate exploration and understanding.

G. Explainability Metrics and Benchmarks

Evaluating the effectiveness of XAI methods remains an open challenge. Future research should develop standardised metrics and criteria to assess the quality, usefulness, and trustworthiness of explanations. These metrics can evaluate aspects such as faithfulness to the model, user understandability, and impact on clinical decision-making.

H. Explainable AI and Regulatory Frameworks

As AI adoption in healthcare rises, regulatory bodies will play a crucial role in ensuring responsible use. Future research should investigate how XAI can contribute to the development of regulatory frameworks that promote transparency, fairness, and accountability in AI-powered medical diagnosis. This could involve developing explainability standards that align with regulatory requirements.

I. Explainable AI and Bias Extenuation

Explainability techniques can be valuable tools for distinguishing and extenuating bias in medical diagnosis models. Future research should investigate the application of XAI methods to identify and comprehend bias within models, and develop strategies to mitigate its impact on clinical decision-making.

J. Explainable Reinforcement Learning for Clinical Decision Support

Reinforcement learning is a promising approach for developing AI systems that can interact with the environment and learn optimal strategies. In healthcare, this technology could be utilised to create AI-powered clinical decision support systems. Future research should explore explainability techniques specific to reinforcement learning models used in the medical domain, enabling clinicians to understand the rationale behind the system's recommendations.

XI. CONCLUSION

The exploration of Explainable AI for medical diagnosis holds tremendous significance with the help of learning models and building trust in deep acquisition models [16], it paves the way for a future of healthcare where AI empowers both patients and healthcare professionals, ultimately leading to improved patient care, informed clinical decision-making, and advancements in the field of

AI in medicine. XAI is crucial for establishing trust in DL models and facilitating their integration into clinical decision-making for medical diagnosis and treatment. By overcoming current challenges and fostering further research, XAI can unlock the full potential of DL in improving patient care and healthcare outcomes. Explainable AI plays a critical role in building trust in deep learning models for clinical decision-making. By combining explainability techniques with robust development practices and clinical expertise, we can unlock the full potential of AI in medical diagnosis, empowering healthcare professionals with reliable and interpretable tools for improved patient care. Ultimately, the goal is to create a seamless collaboration between human expertise and AI capabilities, resulting in a future of data-driven medicine built on trust, transparency, and enhanced patient outcomes. While XAI holds promise for building trustworthy deep learning models in medical diagnosis, the limitations discussed highlight the need for further research and development. By addressing these challenges, we can move toward a future where XAI empowers clinicians with clear and reliable explanations, fostering trust in AI-assisted decision-making and ultimately improving patient care. The potential of deep learning (DL) for revolutionising medical diagnosis is undeniable. However, the widespread adoption of these exemplary approaches in clinical settings is hampered by a critical challenge: their lack of explainability.

The "black box" nature of traditional DL models hinders trust and transparency in their decision-making processes. Physicians need to understand the rationale behind a model's recommendation to integrate it into their clinical workflow confidently. Explainable AI (XAI) emerges as a critical bridge, enabling us to leverage the power of DL while maintaining the necessary level of trust and human oversight in healthcare.

By incorporating XAI techniques, we can build trustworthy DL models that provide clear explanations for their predictions. This fosters collaboration between AI and healthcare professionals, allowing physicians to critically evaluate the model's reasoning and make informed decisions tailored to each patient's unique situation. Transparency fosters trust, enabling a more seamless integration of AI into clinical practice.

Furthermore, the explainability of DL models needs to be assessed in real-world clinical settings. User studies involving physicians can evaluate the effectiveness of XAI techniques in fostering trust, improving understanding, and ultimately leading to better clinical decision-making.

The journey toward trustworthy deep learning in clinical decision-making requires a collaborative effort. Researchers in AI and XAI must collaborate closely with medical professionals to develop and refine explainable models that address the unique needs and challenges of healthcare. Regulatory bodies can play a vital role by establishing guidelines for the development and deployment of interpretable AI in surgical diagnosis.

DECLARATION STATEMENT

After aggregating input from all authors, I must verify the accuracy of the following information as the article's author.

- **Conflicts of Interest/ Competing Interests:** Based on my understanding, this article has no conflicts of interest.
- **Funding Support:** This article has not been funded by any organizations or agencies. This independence ensures that the research is conducted with objectivity and without any external influence.
- **Ethical Approval and Consent to Participate:** The content of this article does not necessitate ethical approval or consent to participate with supporting documentation.
- **Data Access Statement and Material Availability:** The adequate resources of this article are publicly accessible.
- **Author's Contributions:** The authorship of this article is contributed equally to all participating individuals.

REFERENCES

1. M. Kumar, S. Ali Khan, A. Bhatia, V. Sharma and P. Jain, "A Conceptual Introduction of Machine Learning Algorithms," *2023 1st International Conference on Intelligent Computing and Research Trends (ICRT)*, Roorkee, India, 2023, pp. 1-7, DOI: https://doi.org/10.1109/ICRT57042.2023.10146676
2. BroArrieta, A. B., Díaz, N., Serna, J. A., Delgado, S., & Reyes-Luna, A. (2020). Explainable Artificial Intelligence (XAI) in Medicine: An Overview. *Journal of Medical Imaging and Health Informatics*, 10(8), 1682-1691. http://www.aspbs.com/jmihi/contents_jmihi2021.htm#v11n8
3. Bhatia, Abhay, and Anil Kumar. "AI Explainability and Trust in Cybersecurity Operations." *Deep Learning Innovations for Securing Critical Infrastructures*. IGI Global Scientific Publishing, 2025. 57-74. DOI: https://doi.org/10.4018/979-8-3373-0563-9.ch004
4. Ahmad, Muhammad Aurangzeb et al. "Interpretable Machine Learning in Healthcare." *2018 IEEE International Conference on Healthcare Informatics (ICHI)* (2018): 447-447. https://www.semanticscholar.org/paper/Interpretable-Machine-Learning-in-Healthcare-Ahmad-Teredesai/7d065e649e3bfc7d6d36166f50eab37b8404eae0
5. Katarzyna Borys, Yasmin Alyssa Schmitt, Meike Nauta, Christin Seifert, Nicole Krämer, Christoph M. Friedrich, Felix Nensa, Explainable AI in medical imaging: An overview for clinical practitioners – Saliency-based XAI approaches, European Journal of Radiology, Volume 162, 2023, 110787, ISSN 0720-048X, DOI: https://doi.org/10.1016/j.ejrad.2023.110787.
6. Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., ... Y Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639), 115-118. DOI: https://doi.org/10.1038/nature21056
7. Miotto *et al.*, Deep learning for healthcare: review, opportunities and challenges, published in *Briefings in Bioinformatics*, Vol. 19, Issue 6, 2018, pages 1236–1246, DOI: https://doi.org/10.1093/bib/bbx044
8. Selvaraju, R. R., Cogswell, M., Das, A., Vedaldi, A., & Fidler, M. (2017). Grad-CAM: Visual explanations from deep networks via gradient-weighted class activation maps. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 618-626). https://openaccess.thecvf.com/content_ICCV_2017/papers/Selvaraju_Grad-CAM_Visual_Explanations_ICCV_2017_paper.pdf
9. Shrikumar, A., Peyton, J., Konwar, K., Pugatch, V., Branson, S., & Juang, A. (2017). What do neural networks learn in vision? In *Proceedings of the 34th International Conference on Machine Learning-Volume 70* (pp. 3129-3138) https://proceedings.mlr.press/v70/
10. Rane, Nitin and Choudhary, Saurabh and Rane, Jayesh. Explainable Artificial Intelligence (XAI) in healthcare: Interpretable Models for Clinical Decision Support (November 15, 2023). Available at DOI: https://dx.doi.org/10.2139/ssrn.4637897.
11. Qian Xu, Wenzhao Xie, Bolin Liao, Chao Hu, Lu Qin, Zhengzijin Yang, Huan Xiong, Yi Lyu, Yue Zhou, Aijing Luo. Interpretability of Clinical Decision Support Systems Based on Artificial Intelligence from Technological and Medical Perspective: A Systematic Review [2023] DOI: https://doi.org/10.1155/2023/9919269
12. Kumar, A., Bhatia, A., Kashyap, A., & Kumar, M. (2023). LSTM network: a deep learning approach and applications. In *Advanced Applications of*

NLP and Deep Learning in Social Media Data (pp. 130-150). IGI Global ISBN13: 9781668469095. DOI: https://doi.org/10.4018/978-1-6684-6909-5

13. Moorman, Z., Minderman, S., Hood, N. R., & Lin, S. N. (2020). Explainable Artificial Intelligence for Clinical Decision Support Systems. *Clinical Therapeutics*, 42(11), 2316-2329. https://www.researchgate.net/publication/383920571
14. Verma, Praveen, et al. "Sentiment analysis "using SVM, KNN and SVM with PCA." *Artificial Intelligence in Cyber Security: Theories and Applications*. Cham: Springer International Publishing, 2023. 35-53 https://link.springer.com/book/10.1007/978-3-031-28581-3
15. Bhatia, Abhay, et al. "Medications and the Role of Tailored Healthcare." *Smart Healthcare, Clinical Diagnostics, and Bioprinting Solutions for Modern Medicine*. IGI Global Scientific Publishing, 2025. 165-192 DOI: https://doi.org/10.4018/979-8-3373-0659-9.ch009
16. Lundberg, S. M., & Lee, S. I. (2017). A Unified Approach to Interpretable Machine Learning Models. *arXiv preprint arXiv:1705.07874*. DOI: https://doi.org/10.48550/arXiv.1705.07874.

AUTHOR'S PROFILE

Dr. Abhay Bhatia is currently in a postdoctoral program at Kuala Lumpur University of Science & Technology (KLUST), Malaysia. He is affiliated with the Roorkee Institute of Technology, Roorkee, Uttarakhand, where he serves as an Associate Professor in the Department of Computer Science and Engineering. With over 13 years of academic experience, he has worked with several reputed engineering institutions. He earned his B. Tech in Computer Science and Engineering from AKTU (formerly UPTU), followed by an M. Tech in the same field from Rajasthan, and a PhD specializing in Wireless Sensor Networks. Dr. Bhatia is an active member of the IEEE and serves as a reviewer for multiple journals. His research credentials include over 30+ indexed publications in Scopus, IEEE, and SCI journals. He has delivered guest lectures at various institutions on emerging research topics. His dedication to research is further reflected in his six patents and ten book chapters. Additionally, he has authored books titled "Fundamentals of IoT" and "Mastering Data Structures: A Practical Approach," as well as "Practical Approach to Machine Learning with TensorFlow." He, too, has edited books for Bentham Sciences. His primary research interests include Artificial Intelligence, Machine Learning, and Wireless Sensor Networks.

Prof. (Dr.) Rajeev Kumar is a distinguished academician and administrator with over 15 years of experience in advancing professional education. He currently serves as Professor in the Department of Computer Science and Engineering at Moradabad Institute of Technology, Uttar Pradesh, India. He holds a Ph.D. and a D.Sc. in Computer Science, as well as a Postdoctoral Fellowship from Malaysia. He has earned certifications in Data Science and Machine Learning from IIM Raipur, as well as advanced credentials from IBM and Google. A senior IEEE member, he actively contributes to the IEEE Young Professionals Committee and is affiliated with ACM, CSI, and SMIEEE. Prof. Kumar's expertise spans Artificial Intelligence, Cloud Computing, e-Governance, and Networking. He has made significant contributions to curriculum development, delivered expert lectures, provided leadership training, and prepared NAAC SSR documentation. He has represented his institution at global academic forums in London, Mauritius, Malaysia, and Dubai. A committed mentor, he has guided ten Ph.D. scholars, with four more currently pursuing doctorates and five having completed postdoctoral research under his supervision. He has been honoured thrice as Best Project Supervisor. He has developed industry-relevant short-term courses in AI, ML, and Deep Learning, and holds 15 national and international patents. With over 120 publications in SCI, Scopus, IEEE, and Springer-indexed journals, he also serves as an editor and reviewer for reputed international journals and conferences. Prof. Kumar is a renowned speaker and advocate of innovative, student-centred teaching methods. His research spans Cloud Computing, AI, Big Data, and Wireless Sensor Networks, cementing his role as a visionary academic leader.

Dr. Golnoosh Manteghi is a distinguished academic at the Faculty of Architecture and Built Environment, Kuala Lumpur University of Science & Technology (KLUST), located in Unipark Suria, Kajang, Selangor, Malaysia. With a strong background in computer algorithm development, data analysis, and network management, she has made notable contributions to both academia and industry. Dr. Manteghi brings extensive experience as a Senior Lecturer at IUKL, where she plays a key role in shaping future professionals in technology and infrastructure. Her career also includes valuable research experience as a Research Officer with the PETRONAS studies in wireless sensor networks and hydrocarbon detection. Her scholarly output includes numerous research publications, reflecting her commitment to advancing knowledge in areas critical to smart infrastructure and digital communication. Dr. Manteghi's innovations have been recognised with multiple gold and silver medals at national and international innovation exhibitions, underscoring the practical impact of her research. Beyond her academic and research accomplishments, Dr. Manteghi is also recognized for her editorial and leadership roles in scholarly publishing, further establishing her as a respected figure in her field. Known for her passion for learning and her openness to new challenges, she continues to inspire through her dynamic contributions to science and education.