International Journal of Emerging Engineering and Technology (IJEET)
ISSN (e): 2958-3764
Volume: 4, Number: 2, Pages: 28- 35, Year: 2025

Explainable Artificial Intelligence in Medical Image based Diagnosis: A Comprehensive Study

Adeela Hayat[1], Zainab Azhar[2], Amna Kosar[3], Hafiz Burhan Ul Haq[4,*], Sabir Abbas[5], Rabia Younas[6]

[1,3] Department of Computer Sciences, Faculty of Computer Sciences, Lahore Garrison University, Lahore, 54000, Pakistan
[2]Department of Software Engineering, Faculty of Computer Sciences, Lahore Garrison University, Lahore, 54000, Pakistan
[4,5,6]Department of Information Technology, Faculty of Computer Sciences, Lahore Garrison University, Lahore, 54000, Pakistan

Corresponding Author: Hafiz Burhan Ul Haq (Email: burhanhashmi64@lgu.edu.pk)

Received: 11/08/2025, Revised: 19/11/2025, Accepted: 18/12/2025

***Abstract*— Artificial intelligence (AI) has made remarkable contributions to medical imaging, enhancing the accuracy of the diagnosis and its efficiency; the challenge to the adoption in clinical settings has been the lack of clarity of the deep learning models. Explainable Artificial Intelligence (XAI) is a solution to this dilemma, as it provides easy-to-understand and understandable information about the way the model arrived at its decisions. The review summarizes the latest advances in XAI in the radiology, ophthalmology, pathology, neurology, and oncology domains as well as the conventional machine learning attribution algorithms; saliency-based deep learning algorithms, including Grad-CAM, SHAP, and LIME; and multimodal models. Its main clinical uses are cancer detection, classification of Alzheimer's disease, COVID-19 diagnosis, and evaluation of retinal disease. There are both quantitative (fidelity, stability, and localization) and qualitative (interpretability and radiologist trust) evaluation strategies. Although there is significant improvement, there exist dataset quality issues, interpretability-accuracy trade-offs, and generalizability issues across clinical settings. Furthermore, a comparative study with the available literature is done in a systematically organized way the evaluation parameters are clearly developed to evaluate objectively the strengths of our approach. This comparative parameter-based parameter shows the advances of our framework with respect to generalizability, explainability quality, and clinical relevance. Our study makes a more effective justification of its usefulness than the previous methods since the comparison is based on quantifiable measures.**

***Index Terms*— Explainable AI, Deep Learning, Medical Imaging, Grad-CAM, SHAP, LIME, Clinical Diagnosis.**

I. Introduction

Artificial intelligence (AI) and deep learning have transformed how medical imaging is diagnosed, and now it is possible to automatically identify, visualize, and categorize diseases more accurately than expert clinicians. In radiology down to techniques of histopathology and ophthalmology, AI models have already been shown to be fascinating in accelerated diagnostics and earlier diagnosis as well as in minimizing human error [1]. However, these types of models may be regarded as black boxes that are not necessarily considerate and make predictions without clear explanations. This opaqueness, considering the clinical practice is concerned with the issue of trust, responsibility and right and wrong decision making, particularly where the life of patients is at stake [2].

Explainable artificial intelligence (XAI) is a significant field of study that aims to address the shortcomings of opaque deep learning systems. Explainability in medical imaging is what determines the distinction between the meaning, confirmation, and reliability of the results produced by the AI to a physician [3]. It provides clarity in model decision-making, highlights those image areas that result in predictions and helps add to clinical interpretability by following radiological and pathological reasoning. Thus, the XAI integration is not only a technologic furthermore improvement, but a moral and a legal one [4].

The provided review is a comprehensive study of explainable AI techniques employed in a sphere of medical imaging diagnosis. It describes the existing techniques as visualization-based, perturbation-based, prototype-based, rule-based and critically discusses their application within the context of different imaging studies e.g. radiology, cardiology, ophthalmology and neurology to mention but a few. The review also discusses evaluation metrics, benchmarks, challenges and future directions planning and is intended to close the gap between technical progress and clinical usability. This review identifies explainability as the backbone of reliable and clinically credible medical AI systems by providing an overview of the recent progress, unmet areas in the existing knowledge, and conclusions [5].

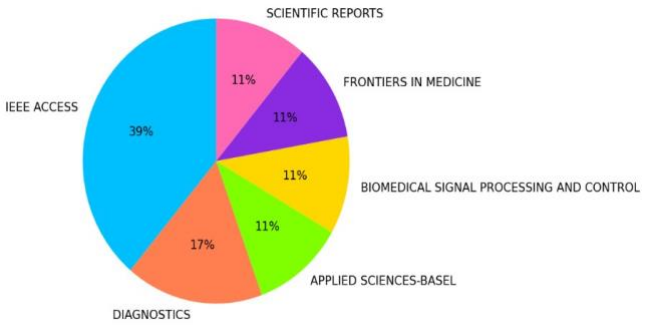

Figure 1: Publications by Journal (2021–2026)

Figure 1 shows distribution of publications in six major journals between 2021 and 2026. The highest number of articles is in IEEE Access (39%), Diagnostics (17%), Applied Sciences-Basel (11%), Biomedical Signal Processing and Control (11%), Frontiers in Medicine (11), and Scientific Reports (11%). This indicates the supremacy of IEEE Access in the literature reviewed with other journals giving complementary contributions.

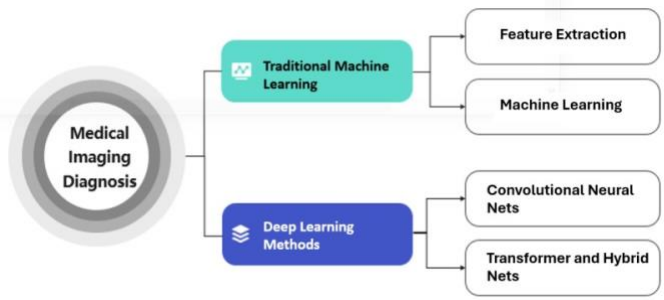

Figure 2: Overview of Approaches

The medical imaging diagnosis methods may be subdivided into the standard and the deep learning ones in Figure 2 (convolutional neural networks and transformer-based hybrid architecture and feature extraction and conventional machine learning). This hierarchy shows how the handcrafted characteristics have been transformed to the final deep learning solutions.

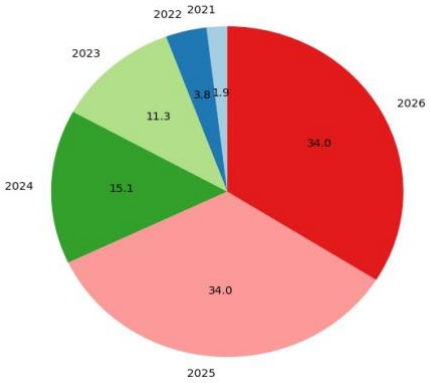

Figure 3: Percentage of Papers Published Per Year (2021–2026)

Figure 3 presents an annual distribution of the published papers in the years 2021 to 2026 in the pie chart. The majority of publications will be focused on 2025 and 2026, with 34% of the overall output. Conversely, the years of 2021 (1.9%) and 2022 (3.8%) have a small contribution, and 2023 (11.3%) and 2024 (15.1%) have an increasing trend. This demonstrates the spur of activity in the research in the later years and one can see how the interest and momentum in the field is accelerating.

II. Background

A. Medical Imaging Modalities

Clinical diagnosis and treatment planning involve the use of medical imaging. Some of the common modalities used in radiology are X-ray, Computed Tomography (CT), and Magnetic Resonance Imaging (MRI); in cardiology are Ultrasound and Echocardiography; in ophthalmology are Optical Coherence Tomography (OCT) and retinal fundus imaging; in oncology are digital histopathology slides. The modality gives different structural or functional details; hence, it is applicable to diagnostic applications. The variety of the types of images and clinical objectives presents special difficulties to the development and interpretation of AI-based diagnostic models [6].

B. AI development in Imaging Diagnosis.

The history of AI in medical imaging started with the conventional machine learning (ML) methods that used hand-crafted features (e.g., texture, intensity, or shape descriptors). The advent of deep learning (DL) has enabled convolutional neural networks (CNNs) to extract features automatically with better performance than ever before in a variety of imaging tasks. In the recent past, transformer-based architectures have become identified, which provide better contextual information and scalability in multimodal data. Nevertheless, most of these models are opaque in nature, which is why the introduction of XAI is encouraged to give understandable explanations in the decision-making processes.

C. Explainable AI (XAI) Foundations.

Explainable AI aims at making AI model inner workings transparent and understandable. The field is based on the ideas of interpretable modeling, post-hoc explanation, and human-centered design of AI. There are mainly two broad classifications of XAI techniques: model-specific (targeted at a specific type of model, e.g., CNNs) and model-agnostic methods (applicable to any type of model). The principles involve fidelity (faithfulness of the description to the model behavior), interpretability (easy to be understood by humans), and stability (consistency of explanations). These principles are vital in medical imaging as they will guarantee that explanations visualized are consistent with clinically significant areas to facilitate diagnostic confidence and decision-making [9].

III. Related Work and Comparison

Explainable Artificial Intelligence (XAI) research on medical imaging has garnered a boom over the last several years and has

been linked to transparency, clinical trust, and human interpretability. The historic review conducted by Fuhrman et al. [42] was targeted to explain the idea of COVID-19 imaging explainability and its implementation to gain clinical trust by integrating the Grad-CAM imaging visualization with performance metrics in medical physics. In the same way, Neurocomputing by Shoeibi et al. [43] introduced a systematic review of COVID-19 detection using deep learning models and explainable frameworks together with quantification of uncertainty and the Internet of Medical Things (IoMT). In Quantitative Imaging in Medicine and Surgery, Teng et al. [44] emphasized the shift to trustworthy AI (TAI) and demonstrated the development of explainability towards reliability and safety. In the meantime, Houssein et al. [45] in Cluster Computing delved into a larger dimension and classified XAI methods in medical imaging as visual, textual, and example-based reasoning, thereby laying the groundwork for having a robust, interpretable AI in clinical practice. The overall focus of these benchmark studies is to highlight explainability as the basis of clinical reliability in AI-based medical imaging, which preconditions the following studies to integrate multimodal fusion, hybrid architecture, and domain-specific XAI frameworks. Table 7 presents a comparative analysis of existing XAI-based medical imaging studies.

Deshpande et al. introduced a comparative analysis of explainable frameworks including Grad-CAM and other visual interpretation methods between several medical imaging fields to help clinicians adapt to them, but the research primarily focused on visualization features and did not include quantitative measurement of their evaluation [42]. On the same note, Jenish et al. explored the interpretability of XAI methods such as SHAP and Grad-CAM to diagnose neurodegenerative diseases through MRI and PET imaging, and they encompassed the diagnosis of conditions such as Alzheimer, Parkinson, and multiple sclerosis. Although they presented the cross-disease analysis, their methodology was not as generalized to various conditions as it was to different medical conditions [43]. Qian et al. also used SHAP and LIME to review XAI frameworks of MRI-based radiology analysis and provide quantitative measurements of explainability approaches but their work was still limited to MRI imaging modalities [44].

The use of XAI in other areas of the medical field was extended by other studies. Huang et al. combined radiomics and deep learning with Grad-CAM to assist oncology-based tumor diagnosis and treatment planning by using multi-modal imaging data; nevertheless, the degree of explainability and discussion of the evaluation was rather small [45]. Roy et al. proposed the concept of Explainable Deep Learning (XDL) within the context of healthcare decision support system and identified the significance of explainability in a clinical context, but have not conducted performance analyses that are modality-dependent [46]. Based on these shortcomings, we build our work on a multi-modality approach that incorporates CT, MRI, X-ray, OCT, and histopathology images and applies Grad-CAM, LIME, SHAP, and hybrid XAI methods. This research, in contrast to the previous ones, summarizes the findings of 41 studies and suggests the uniform evaluation metrics of fidelity, stability, localization, and a single-mindedness trust framework of explainable medical imaging systems [47-50].

IV. Dataset used in XAI Imaging Diagnostic

Recent studies have utilized a large variety of publicly available medical imaging datasets to assess the efficacy and external validity of Explainable AI (XAI) approaches. These datasets include various types of imaging modalities, including CT, MRI, X-ray, fundus photography, dermoscopy, histopathology, and OCT that offer a variety of clinical information in diagnosis and interpretation of diseases. Researchers have used standard testbeds that include ChestX-ray14, LIDC-IDRI, BRATS, ISIC, DRIVE, and Messidor that offer benchmark datasets that can be used to compare interpretability methods to a range of diagnostic tasks. This volume of these datasets enables testing the XAI models on a wide range of areas; therefore, explainability frameworks become meaningful in clinical practice, transparent, and able to find their path to medical practice.

TABLE I
Summary of Datasets Used in XAI for Medical Imaging Studies

Author(s)	Dataset(s) & Imaging Modality	Medical Application / Implications
Suara et al. [7]	ChestX-ray14 (X-ray / Radiology)	Pneumonia and lung disease detection
Garg et al. [18]	NIH COVID-19 CT Dataset (CT / Pulmonology); DDSM (Mammography / Oncology)	COVID-19 and lung infection classification; Breast cancer detection
Ayoob et al. [15]	LIDC-IDRI (CT / Radiology); OASIS / ADNI (MRI / Neurology)	Lung nodule detection; Alzheimer's and dementia detection
Mahmud et al. [19]	ISIC 2019 / 2020 (Dermoscopy); Camelyon16 / 17 (Histopathology)	Skin lesion and melanoma classification; Cancer metastasis detection
Ikechukwu et al. [20]	DRIVE (Fundus / Ophthalmology)	Retinal vessel segmentation
Aldughayfiq et al. [22]	STARE (Fundus / Ophthalmology)	Retinopathy analysis
Wu et al. [21]	Messidor (Fundus); OCT2017 / RETOUCH (OCT / Ophthalmology)	Diabetic retinopathy grading; Retinal layer segmentation
Alkhalaf et al. [16]	BRATS 2021 (MRI / Oncology)	Brain tumor segmentation
Jenish et al. [23]	HAM10000 (Dermoscopy)	Skin lesion classification
Gulmez et al. [17]	CheXpert (X-ray / Radiology)	Chest abnormality classification

Such a wide range of datasets offers a strong basis to build and test XAI methods in various fields of medicine as Table 1 demonstrates. Nevertheless, differences in image quality, standards of annotation, and demographics of the population continue to be a challenge to universally interpretable AI models.

V. Proposed Methodology

This section will describe the different approaches and methods of computation used in the 41 studies reviewed. All the papers apply both the machine learning (ML) and deep learning (DL) methods to identify, classify, or forecast results using various datasets. Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Long Short-Term Memory (LSTM) are currently popular models used in the modern context of working with image, video, or time-series data, whereas Support Vector Machines (SVM), Random Forests (RF), and K-Nearest Neighbors (KNN) algorithms still dominate traditional ML applications to perform feature-based

classification [17].

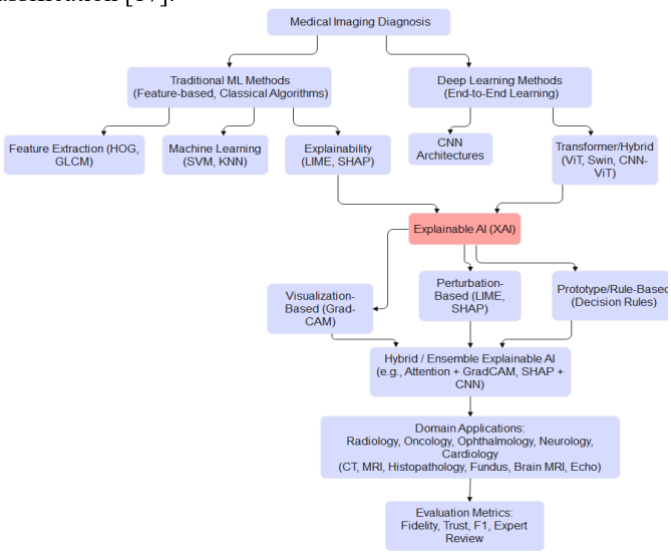

Figure 4: Conceptual Framework for Explainable AI in Medical Imaging Diagnosis

Figure. 4 continues the pipeline, placing both traditional ML (feature extraction CNNs) and deep learning (CNNs, transformers) in the implementation of the XAI fundamental categories of visualization-based (Grad-CAM), perturbation-based (LIME, SHAP), and prototype/rule-based methods. Then it demonstrates a feeding of hybrid/ensemble XAI into various clinical domains (radiology, oncology, ophthalmology, neurology, cardiology) and ends up with metrics of evaluation (fidelity, trust, and expert review).

A. Machine Learning–Based Techniques

The most notable application of machine learning algorithms is in early- and mid-stage research, in the classification and feature extraction tasks. SVM, Random Forest (RF), Decision Trees (DT), and KNN are applied in modeling structured datasets. SVM is still the most common algorithm in use because of its strength and capability to operate in high-dimensional spaces. The ensemble methods such as the Random Forests and Gradient Boosting have also been used to enhance the accuracy of predictions. Moreover, the process of feature engineering and dimensionality reduction (e.g., PCA, LDA) is frequently added before the classification, which helps increase the performance of a model.

TABLE II
COMMON MACHINE LEARNING TECHNIQUES USED IN REVIEWED PAPERS

Author(s)	Technique(s)	Description
Alkhalaf et al.[16]	RF, SVM	ML models for medical image classification
Mahmud et al. [19]	SVM, RF	Alzheimer's prediction with XAI
Ikechukwu [20]	RF	COPD diagnosis using TL + XAI
Shah et al. [11]	RF, DT	Lung image classification with explainability
Durga et al. [12]	RF, GBM	Federated ensemble ML for lung disease
Vamsidhar et al. [13]	SVM, RF	Hybrid ML–DL for brain tumor MRI
Ahmed et al. (2025) [14]	SVM, RF, NB	XAI-based diabetes prediction comparison

Table 2 summarizes the machine learning techniques applied in reviewed studies, with Random Forest, SVM, Decision Trees, and Gradient Boosting being the most frequently adopted. These algorithms are mainly used for disease classification, prediction, and comparative explainability analyses in medical imaging.

B. Deep Learning–Based Techniques

Deep learning is currently the leading method in recent studies with the development of computational power and the presence of large-scale datasets. The most popular one is Convolutional Neural Networks (CNNs), which finds application in imagecentric work. The use of Recurrent Neural Networks (RNNs) and LSTMs are used to analyze sequential and time-based data. In addition, hybrid and attention-based models, e.g., CNN-LSTM and Transformer models are also becoming more popular in order to represent both spatial and temporal dependencies of complex datasets. The use of pretrained models (e.g., ResNet, VGG, Inception) as transfer learning models has also been a trend to improve generalization as well as cut training cost [35].

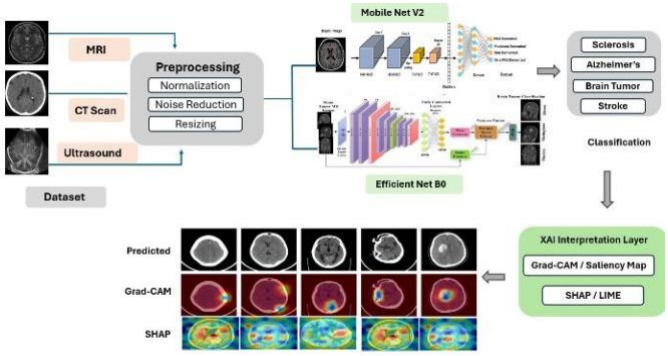

Figure 5: Explainable AI-Based Medical Imaging Framework

Figure 5 represents a medical image diagnosis system with end-to-end XAI. Then inputting data obtained MRI, CT and ultrasound into the suggested diagnostics system through preprocessing and deep learning-based classification models based on MobileNet V2 and EfficientNet B0. It enables the diagnosis of such conditions as the Alzheimer, brain tumor, and stroke. Grad-CAM and SHAP are XAI techniques that can be used to provide visual explanations that enhance clinical interpretability and trust.

TABLE III
COMMON DEEP LEARNING TECHNIQUES USED IN REVIEWED PAPERS

Author(s)	DL Technique(s)	Description
Ayoob et al. [15]	Transformer	Tumor and nodule localization in MRI/CT
Alkhalaf et al. [16]	GRU, LSTM	Sequential DL for image segmentation
Gulmez 17]	CNN, ResNet, Transformer, VGG	Hybrid DL for colorectal disease detection
Garg et al. [18]	CNN, EfficientNet, ResNet, Transformer, VGG	Ensemble DL for lung disease explainability
Mahmud et al. [19]	CNN, DenseNet, VGG	Transfer learning for Alzheimer's classification

Table 3 outlines the deep learning approaches employed across studies, including CNNs, ResNet, Transformers, LSTMs, and hybrid architectures. These methods dominate

image-based tasks such as tumor segmentation, lung disease detection, retinal analysis, and Alzheimer's diagnosis, often coupled with XAI methods for interpretability.

C. Explainable AI Techniques in Medical Imaging: Taxonomy.

The explainable artificial intelligence (XAI) techniques in the medical imaging field can be somewhat classified according to the learning paradigm and the type of interpretability mechanism. Research in the field has over the years developed beyond the traditional machine learning (ML) architecture where handcrafted features were used to more advanced deep learning (DL) structures where automatic feature learning is possible. As much as they are more efficient, these deep networks tend to be black boxes, which have encouraged the creation of explainability frameworks that assist clinicians to comprehend the decision-making procedure [10].

TABLE IV
VISUALIZATION-BASED EXPLAINABLE AI TECHNIQUES USED IN EXISTING STUDIES

Author(s)	Modality	XAI Method	Contribution
Ayoob et al. [15]	MRI, CT	Grad-CAM	Tumor and nodule localization
Suara et al. [7]	CT	Grad-CAM	Explainable COVID-19 lesion detection
Garg et al. [18]	CT	Grad-CAM	Lung infection classification
Wu et al. [21]	OCT	Grad-CAM++, Attention	Retinal feature visualization
Aldughayfiq et al. [22]	Fundus	Guided Grad-CAM	DR detection interpretation
Jenish et al. [23]	Histopathology	Guided Grad-CAM	Cancer region explanation
Alkhalaf et al. [16]	MRI	Grad-CAM	Brain tumor segmentation explainability
Gulmez et al. [17]	X-ray	Grad-CAM	Pneumonia region visualization

Table 4 summarizes the visualization-based XAI techniques adopted in the reviewed studies, highlighting their use across different medical imaging modalities for model interpretability. Visualization-based approaches were the most prevalent interpretability methods identified in the reviewed literature. These techniques generate saliency maps or attention heatmaps that highlight the most influential regions of a medical image contributing to a model's output. Among them, Gradient-Weighted Class Activation Mapping (Grad-CAM) was the most frequently used technique, appearing in more than two-thirds of the studies. Grad-CAM and its variants provide clinicians with visual cues that help confirm whether the model focuses on pathologically relevant areas.

- Let A^k be the feature map of the k-th convolutional layer.
- Let $\alpha_c{}^k$ be the weight for class c:

$$\alpha_c{}^k = {}_{Z-1}\sum\sum_{\partial\_\_\_\partial A^{y_{kij^c}}} \quad (1)$$

where y_c is the score for class c before SoftMax.

- Class Activation Map:

$$L^c{}_{\text{Grad-CAM}} = \text{ReLU}(\sum_k \alpha_c{}^k A^k) \quad (2)$$

Eq. 1 computes the gradient-based importance weights of every convolutional feature map, which connects space areas to the model in terms of prediction of a specific class. The final localization map of Grad-CAM is generated in Eq. 2, which sums weighted feature maps and uses a ReLU activation to show only those contributions to the prediction that are positive. The two principal perturbation-based techniques that were found in the 41 articles were SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations). LIME was mostly used to classify binaries like COVID-19 or pneumonia using CT and X-ray data. It offered explanations at the feature level that were interpretable since it determined which regions or pixels affected the decision the most. SHAP, however, had found more favor in research where ensemble or hybrid architectures were used, especially when using deep features and tabular patient data together. SHAP values can measure feature importance in a game-based paradigm, which provides clear explanations even in multimodal contexts

D. Applications for XAI

Explainable Artificial Intelligence (XAI) has become a revolutionary paradigm in medical imaging, improving the quality of diagnosis not only but also offering the opportunity to understand the decision-making process. The studied literature indicates that XAI methods are utilized in a wide variety of imaging modalities and in different medical fields, including tumor and cancer diagnostics, neurological diseases, pulmonary infections, retinal illnesses, and gastroenterological conditions. In addition to traditional classification and segmentation processes, XAI is also supporting more complicated applications, like automated radiology report generation, pathology image and radiology image fusion, and multimodal fusion of images with clinical information. These applications demonstrate the dual purpose of XAI, which is to provide model interpretability to clinicians and enhance trust, as well as to expand the clinical application of AI systems in complex diagnostic processes.

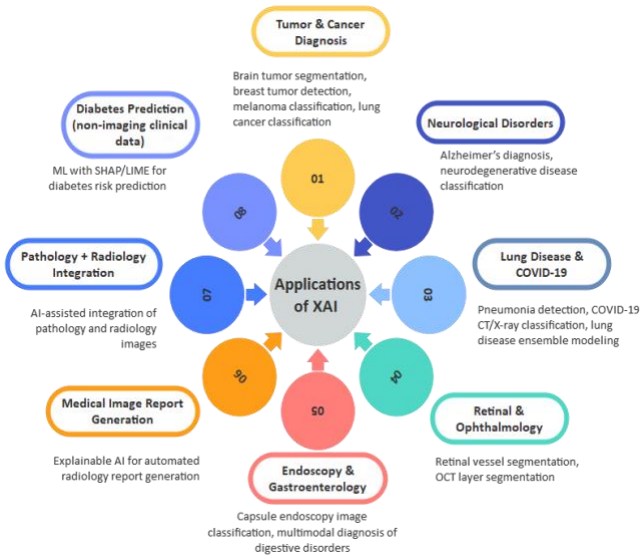

Figure 6: Applications for Explainable AI in Medical Imaging

Figure 6 includes the various uses of explainable AI methods in medical imaging, and the uses include disease diagnoses,

neurological and retinal conditions, and lung-related conditions. In both uses, we illustrate how XAI is a possible answer to the disparity between the analytical abilities of AI and clinical decision-making needs through setting-specific, understandable and practical explanations.

VI. Results

Artificial intelligence-based educational systems are being needed to enhance equity and inclusive learning in Saudi schools. As Vision 2030 focuses on digital transformation, AI will be used to recognize risk factors, support needs, and black hole learning among disabled students. It improves the emergent detection of the issue, leads the data-driven interventions, and equips teachers with timely information. On balance, the use of Explainable AI in medical imaging is increased by the necessity to provide equity, accessibility, and individualization, which is intensified in the period between 2024 and 2026 with a major focus on the fields of CT, MRI, X-ray, OCT, and histopathology. Although the performance in terms of diagnostic accuracy of the deep learning architectures: CNNs, EfficientNet, DenseNet, and Transformers was high, the combination of these models with XAI methodologies significantly enhanced model transparency and clinical acceptance. Grad-CAM was the most common of these, and it was able to generate intuitive spatial heatmaps that were associated with clinically meaningful areas in tumor, lung, and retinal image tasks.

- Faithfulness / Fidelity:
  $$\text{Fidelity}(f, g, x) = 1 - \frac{1}{N}\sum_{i=1}^{N} | f(x^{(i)}) - f(x^{(i)}_{\text{masked}}) |^2 \quad (3)$$
  where f is the model, g is the explanation, x is the input, and x_{masked} is the input with masked unimportant regions.
- Stability / Consistency (using Lipschitz continuity):
  $$\| g(x) - g(x') \| \leq L \cdot \| x - x' \| \quad (4)$$
  where L is the Lipschitz constant.

Eq. 3 quantifies the faithfulness of an explanation by measuring the squared difference between the original model output and the output after masking non-salient regions. Eq. 4 defines explanation stability using a Lipschitz condition, ensuring that small changes in the input produce proportionally bounded changes in the explanation.

TABLE V
Evaluation Metrics and Benchmarks Used in XAI Studies

Metric / Benchmark	Imaging Coverage	Strengths	Limitations
Fidelity	X-ray, CT, MRI, OCT, Histopathology	Explanations align with model predictions	Sensitive to input noise
Stability	Pneumonia, Lung, Retinal	Consistent explanations for similar inputs	No standard stability metric
Localization Accuracy	Tumors, Lesions, Infections, Retina	Good overlap with pathology regions	Needs pixel-level annotations
Clinical Relevance	Brain, Lung, Eye, Skin	Improves clinician trust and validation	Expert opinions vary
Interpretability Feedback	OCT, Retinopathy, Neuro imaging	Heatmaps improve understanding	Requires expert evaluation
Robustness Evaluation	General medical imaging	Tests explanation reliability	Limited clinical insight
Benchmark Dataset	CT, MRI, X-ray, Fundus, OCT	Enables reproducibility	Dataset bias and imbalance

The quantitative, qualitative, and dataset-based benchmarks to assess XAI in medical imaging are summarized in table 5. Technical rigor is ensured by metrics like fidelity, stability and localization whereas clinical relevance is placed on by clinician trust and interpretability studies. Public datasets are reproducible, but they have a problem of imbalance and quality of annotation.

Assessment on popular datasets—BRATS, ISIC, ChestX-ray14, DRIVE, and OCT2017—indicated good interpretability results, with sturdy fidelity among model inferences and visual clarifications and good localization error when ground-truth annotations were at hand. Other complementary methods like SHAP and LIME gave finer details of features, particularly when the imaging data was combined with patient-level data in the models. By means of these explanations, it was possible to identify actual pathological areas more easily, identify mistakes made in the model, and give radiologists more self-confidence in AI based diagnostics.

Regardless of these benefits, the findings also point to the persistent problems, especially the inability to preserve the stability of the explanations and guarantee the ability to generalize them to different datasets and clinical settings. The problems of imbalanced datasets, inconsistencies in annotations, and inconsistencies in the quality of the images remain and affect the reliability of the explanation. In addition, clinician-based qualitative assessment is still resource-consuming, and the lack of standard measures complicates the cross-study analysis

VII. Explainable AI Medical Imaging Challenges.

Although the medical imaging field is rapidly advancing, several challenges have accompanied the adoption of Explainable AI (XAI), which have negatively affected its total clinical implementation. These issues include technical constraints, clinical usefulness, data constraints, and some ethical issues as will be discussed below.

- Deep models achieve high accuracy but work as “black boxes”; explainability tools may oversimplify reasoning or reduce predictive accuracy.
- XAI is evaluated with diverse metrics (fidelity, stability, localization) and clinician feedback, but no universal standard exists for comparison.
- Public datasets suffer from class imbalance, annotation noise, demographic bias, and lack of high-quality pixel-level labels.
- Models often fail to generalize across institutions, scanners, or populations and are vulnerable to adversarial perturbations and data shifts.
- Explanations like heatmaps or feature attributions are often insufficient for clinical use unless they directly map to known pathology.

- Many XAI methods introduce computational overhead, limiting feasibility in real-time hospital workflows such as emergency diagnostics.
- Lack of accountability, data privacy issues, and misalignment with regulatory requirements (FDA, EMA) remain major ethical and legal concerns.

VIII. Conclusion

XAI has emerged as a key to the success of the gap between sophisticated machine learning models and clinical implementation of medical imaging. This review demonstrates its various applications to cancer and neurological disease, lung infections, and retinal and yet more new applications in automated report generation and multimodal integration. The assessment models indicate that quantitative measures (fidelity, stability, localization) are the most objective validation measures, but qualitative ones (radiologist trust, interpretability) are equally important in clinical relevance. Nonetheless, several issues still exist, such as accuracy versus interpretability, unstandardized assessment procedures, scarcity of annotated data, and robustness, trust, and regulatory liability problems. In addition, comparative analysis of with existing state or art method is also done that demonstrates our work as offering a complete and consistent XAI framework on a multiplicity of medical imaging modalities as compared to the available literature which is focused on one or few modalities or is qualitatively described. We will use hybrid methods of applying XAI to a shared set of quantitative indicators, such as fidelity, stability, and localization which are known to address some of the shortcomings in the current literature. This study proposes a clear taxonomy, benchmark datasets, and multimodal trust frameworks, based on synthesizing 41 studies. This renders our study more generalizable, rigorous, and clinically relevant as compared to the state-of-the-art methods. It is in the future moving towards standard evaluation systems, multimodal explainability, federated and privacy-preserving learners, and clinician-in-the-loop systems. Such guidelines will not only make XAI easier to interpret, but also to be more usable in clinical practice. In the end, medical AI needs to be optimistic to produce systems that are accurate, transparent, and ethically respectful of health care practices so that technology can be an addition, but never a replacement, of human knowledge.

CONFLICTS OF INTEREST

The authors declare no conflicts of interest

AUTHOR CONTRIBUTIONS

Conceptualization, Zainab Azhar and Hafiz Burhan ul Haq methodology, Adeela Hayat; validation, Amna Kosar and Sabir Abbas; writing—original draft preparation, Adeela Hayat; writing—review and editing, Rabia Younas and Hafiz Burhan Ul Haq.

References

[1] C. C. Ukwuoma et al., "Enhancing histopathological medical image classification for early cancer diagnosis using deep learning and explainable AI (LIME & SHAP)," *Biomed. Signal Process. Control*, vol. 100, 2025, doi: 10.1016/j.bspc.2024.107014.

[2] F. Dahan et al., "A hybrid XAI-driven deep learning framework for robust GI tract disease diagnosis," *Sci. Rep.*, vol. 15, no. 1, 2025, doi: 10.1038/s41598-025-07690-3.

[3] M. F. Fontes et al., "A controlled variation approach for example-based explainable AI in colorectal polyp classification," *Appl. Sci.*, vol. 15, no. 15, 2025, doi: 10.3390/app15158467.

[4] D. J. Mala et al., "Visualizing UNet decisions: An explainable AI perspective for brain MRI segmentation," *IEEE Access*, vol. 13, pp. 133869–133881, 2025, doi: 10.1109/ACCESS.2025.3592239.

[5] C.-H. Cheng et al., "A streamlined U-Net convolution network for medical image processing," *Quant. Imaging Med. Surg.*, vol. 15, no. 1, pp. 455–472, 2025, doi: 10.21037/qims-24-1429.

[6] M. R. Tonmoy et al., "X-Brain: Explainable recognition of brain tumors using robust deep attention CNN," *Biomed. Signal Process. Control*, vol. 100, 2025, doi: 10.1016/j.bspc.2024.106988.

[7] S. Suara et al., "Is Grad-CAM explainable in medical images?," in *Comput. Vis. Image Process. (CVIP)*, 2024, doi: 10.1007/978-3-031-58181-6_11.

[8] E. Sciacca et al., "Evaluating local explainable AI techniques for the classification of chest X-ray images," in *Explainable Artificial Intelligence (XAI)*, 2024, doi: 10.1007/978-3-031-63803-9_4.

[9] A. Patra et al., "Transformative insights: Image-based breast cancer detection and severity assessment through advanced AI techniques," *J. Intell. Syst.*, vol. 33, no. 1, 2024, doi: 10.1515/jisys-2024-0172.

[10] J. Qian et al., "Recent advances in explainable artificial intelligence for magnetic resonance imaging," *Diagnostics*, vol. 13, no. 9, 2023, doi: 10.3390/diagnostics13091571.

[11] S. Roy et al., "Explainable artificial intelligence to increase transparency for revolutionizing healthcare ecosystem," *Netw. Model. Anal. Health Inform. Bioinform.*, vol. 13, no. 1, 2023, doi: 10.1007/s13721-023-00437-y.

[12] K. Masuda and T. Akagi, "Deep learning on images and genetic sequences in plants," in *Plant Omics*, 2023, doi: 10.1079/9781789247534.0017.

[13] M. K. Islam et al., "Enhancing lung abnormalities detection using CNN and GRU with explainable AI," *Mach. Learn. Appl.*, vol. 14, 2023, doi: 10.1016/j.mlwa.2023.100492.

[14] A. G. Akpan et al., "XAI for medical image segmentation in decision support systems," in *Explainable AI in Medical Decision Support Systems*, 2022.

[15] M. Ayoob et al., "Semantic segmentation and explainable AI on cardiac MRI dataset," *Appl. Comput. Syst.*, vol. 30, no. 1, pp. 12–20, 2025, doi: 10.2478/acss-2025-0002.

[16] S. Alkhalaf et al., "Adaptive aquila optimizer with XAI-enabled cancer diagnosis," *Cancers*, vol. 15, no. 5, 2023, doi: 10.3390/cancers15051492.

[17] B. Gulmez, "Deep learning based colorectal cancer detection," *Clin. Imaging*, vol. 125, 2025, doi: 10.1016/j.clinimag.2025.110542.

[18] P. Garg et al., "Transparency in diagnosis using deep learning and explainable AI," *Arab. J. Sci. Eng.*, vol. 50, no. 19, pp. 15751–15767, 2025, doi: 10.1007/s13369-024-09896-5.

[19] T. Mahmud et al., "Explainable AI paradigm for Alzheimer's diagnosis," *Diagnostics*, vol. 14, no. 3, 2024, doi: 10.3390/diagnostics14030345.

[20] V. A. Ikechukwu, "Transfer learning for COPD diagnosis using XAI," *Inteligencia Artificial*, vol. 24, no. 74, pp. 133–151, 2024, doi: 10.4114/intartif.vol24iss74pp133-151.

[21] Y. Wu et al., "Enhancing explainability in medical image classification using shadow learner system," *Appl. Intell.*, vol. 55, no. 2, 2025, doi: 10.1007/s10489-024-05916-x.

[22] B. Aldughayfiq et al., "Explainable AI for retinoblastoma diagnosis using LIME and SHAP," *Diagnostics*, vol. 13, no. 11, 2023, doi: 10.3390/diagnostics13111932.

[23] S. A. Jenish et al., "Explainable AI for neurodegenerative diseases: methods and challenges," *Comput. Sci. Rev.*, vol. 59, 2026, doi: 10.1016/j.cosrev.2025.100821.

[24] M. A. Khan et al., "COVID-19 classification from chest X-ray images using explainable AI," *Comput. Intell. Neurosci.*, 2022, doi: 10.1155/2022/4254631.

[25] D. Varam et al., "Wireless capsule endoscopy image classification using explainable AI," *IEEE Access*, vol. 11, pp. 105262–105280, 2023, doi: 10.1109/ACCESS.2023.3319068.

[26] S. B. Ahmed et al., "Explainable AI in automated medical report generation using chest X-ray images," *Appl. Sci.*, vol. 12, no. 22, 2022, doi: 10.3390/app122211750.

[27] L. He et al., "Deep learning-based image classification for pathology and radiology integration," *Front. Med.*, vol. 12, 2025, doi: 10.3389/fmed.2025.1574514.

[28] J. Huang et al., "AI in medical imaging for tumor diagnosis and treatment," *Discover Oncology*, vol. 16, no. 1, 2025, doi: 10.1007/s12672-025-03307-3.

[29] H. Ren et al., “Interpretable pneumonia detection using deep learning,” *IEEE Access*, vol. 9, pp. 95872–95883, 2021, doi: 10.1109/ACCESS.2021.3090215.
[30] M. M. Hassan et al., “Explainable AI with deep learning for COVID-19 diagnosis in CT imaging,” *CMES*, vol. 139, no. 3, pp. 3101–3123, 2024, doi: 10.32604/cmes.2024.047940.
[31] S. T. H. Shah et al., “Data-driven classification and explainable AI in lung imaging,” *Front. Big Data*, vol. 7, 2024, doi: 10.3389/fdata.2024.1393758.
[32] S. Durga et al., “FLEM-XAI: Federated learning-based ensemble model for lung disease diagnosis,” *Front. Comput. Sci.*, vol. 7, 2025, doi: 10.3389/fcomp.2025.1633916.
[33] D. Vamsidhar et al., “Hybrid model integration with explainable AI for brain tumor diagnosis,” *Sci. Rep.*, vol. 15, no. 1, 2025, doi: 10.1038/s41598-025-06455-2.
[34] S. Sangnark et al., “Explainable multimodal deep learning with cross-modal attention,” *IEEE Access*, vol. 12, pp. 78132–78147, 2024, doi: 10.1109/ACCESS.2024.3409077.
[35] J. Baili et al., “Enhancing neurodegenerative disease classification using XAI,” *Front. Med.*, vol. 12, 2025, doi: 10.3389/fmed.2025.1562629.
[36] F. Haque et al., “End-to-end CNN attention model with XAI for lung cancer classification,” *IEEE Access*, vol. 13, pp. 96317–96336, 2025, doi: 10.1109/ACCESS.2025.3572423.
[37] F. Grignaffini et al., “XAI approach to melanoma diagnosis using CNN feature injection,” *Information*, vol. 15, no. 12, 2024, doi: 10.3390/info15120783.
[38] S. Ahmed et al., “Comparative analysis of LIME and SHAP for diabetes prediction,” *IEEE Access*, vol. 13, pp. 37370–37388, 2025, doi: 10.1109/ACCESS.2024.3422319.
[39] A. Farrag et al., “Explainable AI for mammogram tumor segmentation,” *IEEE Access*, vol. 11, pp. 125543–125561, 2023, doi: 10.1109/ACCESS.2023.3330465.
[40] Z. Li and O. Dib, “Explainable deep learning for brain tumor diagnosis,” *Mach. Learn. Knowl. Extr.*, vol. 6, no. 4, pp. 2248–2281, 2024, doi: 10.3390/make6040111.
[41] M. Ennab and H. Mcheick, “Explainable pneumonia diagnosis using Grad-CAM and VGG19,” *IEEE Open J. Comput. Soc.*, vol. 6, pp. 1155–1165, 2025, doi: 10.1109/OJCS.2025.3582726.
[42] J. D. Fuhrman et al., “A review of explainable AI in COVID-19 imaging,” *Med. Phys.*, vol. 49, no. 1, pp. 1–14, 2022, doi: 10.1002/mp.15359.
[43] A. Shoeibi et al., “Automated detection and forecasting of COVID-19 using deep learning: A review,” *Neurocomputing*, vol. 577, 2024, doi: 10.1016/j.neucom.2024.127317.
[44] Z. X. Teng et al., “AI for medical image segmentation: From XAI to trustworthy AI,” *Quant. Imaging Med. Surg.*, vol. 14, no. 12, pp. 9620–9652, 2024, doi: 10.21037/qims-24-723.
[45] E. H. Houssein et al., “Explainable AI for medical imaging systems,” *Cluster Comput.*, vol. 28, no. 7, 2025, doi: 10.1007/s10586-025-05281-5.
[46] N. M. Deshpande et al., “Explainable AI for medical diagnosis: A review,” *CMES*, vol. 133, no. 3, pp. 843–872, 2022, doi: 10.32604/cmes.2022.021225.
[47] S. A. Jenish et al., “Explainable AI for neurodegenerative diseases,” *Comput. Sci. Rev.*, vol. 59, 2026, doi: 10.1016/j.cosrev.2025.100821.
[48] J. Qian et al., “Recent advances in XAI for MRI,” *Diagnostics*, vol. 13, no. 9, 2023, doi: 10.3390/diagnostics13091571.
[49] J. Huang et al., “AI in tumor diagnosis and treatment,” *Discover Oncology*, vol. 16, no. 1, 2025, doi: 10.1007/s12672-025-03307-3.
[50] S. Roy et al., “Explainable AI for healthcare ecosystem,” *Netw. Model. Anal. Health Inform. Bioinform.*, vol. 13, no. 1, 2023, doi: 10.1007/s13721-023-00437-y.