Advanced International Journal for Research (AIJFR)
E-ISSN: 3048-7641 ● Website: www.aijfr.com ● Email: editor@aijfr.com

# Neuro-Symbolic Intelligence in Medical Imaging: A Systematic Review of Explainable Chest X-Ray Analysis

**Dr. Leena Patil[1], Dr. Namrata Khade[2], Dr. Himanshu Taiwade[3], Ms. Sakshi Sawate[4], Dr. Vaishnavi Ganesh[5]**

[1,2,3,4,5] Associate Professor, Department of Computer Science and Engineering, Priyadarshini College of Engineering, Nagpur, India
[4] PG Scholar, Department of Computer Science and Engineering, Priyadarshini College of Engineering, Nagpur, India
[5] Assistant Professor, Department of Computer Science and Engineering, Priyadarshini College of Engineering, Nagpur, India

**Abstract**

The increasing use of deep learning in medical imaging has significantly improved diagnostic accuracy but raised concerns regarding interpretability and trustworthiness. Conventional convolutional artificial neural networks (CNNs) function as "black-box" models that provide little information about how they make decisions. This study combines the logical accessibility of symbolic reasoning with the perceptual power of neural networks to offer a neuro-symbolic structure for an straightforward medical diagnosis utilizing chest X-ray pictures. The CNN component extracts and classifies thoracic abnormalities, while the symbolic reasoning layer integrates medical rules and ontologies to produce interpretable, rule-based diagnostic conclusions. The hybrid approach ensures that diagnostic outputs are both data-driven and clinically meaningful, aligning with medical guidelines. The proposed system enhances transparency, reduces diagnostic bias, and supports clinicians with human-understandable explanations. Experimental evaluation demonstrates improved accuracy and interpretability for detecting pneumonia, tuberculosis, and lung abnormalities. This study contributes toward building trustworthy, ethical, and explainable artificial intelligence solutions for healthcare applications.

**Keywords:** Neuro-Symbolic Artificial Intelligence, Explainable Medical Diagnosis, Chest X-Ray Image Analysis, Deep Learning and Symbolic Reasoning, Interpretable Healthcare Systems etc.

## 1. INTRODUCTION

Conventional convolutional neuronal networks (CNNs) function as "black-box" models that provide little information about how they make decisions. This study combines the logical candidness of symbolic reasoning with the perceptual power of neural networks to offer a neuro-symbolic structure to support straightforward medical diagnosis utilizing chest X-ray pictures, chest X-rays are the most widely used for detecting thoracic abnormalities such as pneumonia, tuberculosis, lung cancer, and

COVID-19-related infections. The quick development of deep learning (DL) and artificial intelligence (AI) has greatly improved automated medical image interpretation, giving radiologists and other medical practitioners effective tools. When it comes to recognizing patterns and characteristics in complicated medical images, neural networks that use convolution (CNNs) have demonstrated impressive effectiveness, frequently attaining accuracy on par with human specialists. However, despite their predictive strength, CNN-based diagnostic systems suffer from a major limitation — a lack of interpretability [1].

Most deep learning models operate as "black boxes," producing results without transparent reasoning. This opacity prevents clinicians from understanding how or why a particular diagnosis is made, which is critical in life-critical medical decisions. The absence of explainability not only undermines trust but also limits clinical adoption, regulatory approval, and accountability. To overcome this limitation, researchers are exploring Explainable Artificial Intelligence (XAI) frameworks that make model decisions more transparent, interpretable, and aligned with human reasoning [2].

The neuro-symbolic arrangement, which combines the logical reasoning power of symbolic AI with the statistical learning capability of neural networks, is one promising strategy in this area. Neural components effectively process raw image data to identify features such as nodules, opacities, or infiltrations, while symbolic components use medical knowledge, rules, and ontologies to interpret these features logically. This combination enables models to not only detect anomalies but also explain their diagnostic reasoning in a manner consistent with medical guidelines. For example, if the CNN detects opacity in the lower lung region, the symbolic layer can interpret this using a rule such as "Opacity + fever + shortness of breath → pneumonia suspicion."

The neuro-symbolic approach addresses two critical needs: accuracy and interpretability. While deep learning ensures efficient pattern recognition, symbolic reasoning ensures that results are explainable and grounded in domain-specific medical logic. This enhances clinical trust and supports radiologists in verifying model outcomes. Furthermore, by including structured domain expertise that goes beyond statistical data patterns, such hybrid systems are better able to generalize across a variety of datasets and medical situations [2][3].

The proposed research aims to develop a Neuro-Symbolic Framework for Explainable Medical Diagnosis using Chest X-Ray Images. The system integrates a CNN-based feature extractor with a symbolic reasoning layer built upon medical ontologies and diagnostic rules. The framework provides interpretable outputs in the form of highlighted image regions and rule-based textual explanations. The study will evaluate the model's diagnostic accuracy, interpretability, and clinical relevance against conventional black-box models.

Ultimately, this research contributes to building trustworthy, transparent, and ethical AI systems for healthcare. By bridging the gap between deep learning and human reasoning, the proposed framework moves toward a future where AI not only diagnoses diseases accurately but also explains its decisions in a clinically meaningful way [4][5].

## 2. PROBLEM IDENTIFICATION

• Traditional deep learning-based medical image diagnostic systems, though highly accurate, function as "black boxes," offering limited transparency into how predictions or classifications are derived.
• Clinicians and healthcare professionals often find it difficult to trust AI-generated results without understanding the underlying reasoning, which hinders clinical adoption.
• The absence of interpretability and explainability in deep learning models raises serious ethical, legal, and accountability concerns in medical decision-making.
• Current CNN-based systems focus primarily on accuracy but neglect the logical reasoning process that aligns with human medical understanding.
• Symbolic AI systems, on the other hand, offer reasoning and explainability but lack the data-driven adaptability and robustness of neural networks.
• There is a need for a hybrid model that can combine the strengths of both neural and symbolic paradigms to achieve accuracy along with transparency.
• Lack of explainable frameworks restricts radiologists from verifying model predictions and integrating AI into routine diagnostic workflows effectively.
• Hence, developing a neuro-symbolic framework for explainable medical diagnosis using chest X-ray images becomes essential to bridge this critical gap [5][6][7].

## 3. FUNDAMENTALS OF EXPLAINABLE ARTIFICIAL INTELLIGENCE

A crucial field devoted to enhancing the interpretability and transparency of AI systems is Explainable Artificial Intelligence (XAI). Its main objective is to help users—especially in critical domains like healthcare—understand, trust, and effectively oversee the decisions generated by AI models. Unlike traditional "black-box" algorithms that obscure their internal logic, XAI focuses on uncovering and communicating the reasoning behind model predictions in a manner that humans can easily comprehend [8].

Differentiating between interpretability and transparency is a fundamental component of XAI. Interpretability is the degree to which humans can intuitively understand the relationship amongst inputs and outcomes in the model's decision-making process, whereas transparency is the degree to which a model's internal mechanisms may be investigated [9,10]. Both are crucial in the context of imaging for medical reasons because clinical decision-making necessitates traceability, clarity, and conformity to accepted diagnostic criteria.

XAI includes a variety of explanations intended to satisfy different user and technical requirements. Post-hoc, model-agnostic methods that interpret individual predictions following training include SHapley Additive exPlanations (SHAP) and Local Declarable Model-Agnostic Explanations (LIME) [11]. On the other hand, some models are naturally interpretable, such as rule-based systems and decision trees, offering built-in transparency that makes them suitable for use in sensitive or highly regulated environments [12].

Another crucial element of XAI is context-awareness. An explanation's relevance to the target audience and particular application determines its quality and utility. Technical explanations, for instance, can be useful for data scientists but too complicated for patients or clinicians. In order to apply

XAI effectively, it is necessary to comprehend the user's history and modify explanations accordingly [13]. This ensures that the output is not only technically valid but also meaningful and actionable.

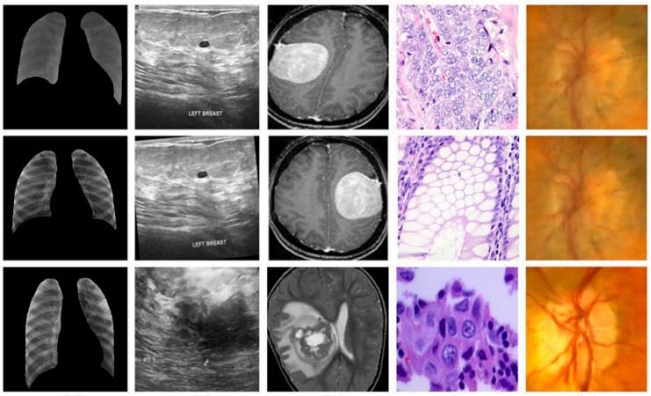


Figure 1. Examples of commonly used medical imaging techniques include: (a) chest X-ray, (b) breast ultrasound, (c) brain MRI, (d) histopathological tissue scan, and (e) retinal fundus image. These various imaging modalities demonstrate the vast array of photographic information that XAI technologies need to evaluate and comprehend in order to help physicians make precise and knowledgeable diagnostic decisions [14].

XAI applications in medical imaging must also take into consideration the variety of imaging technologies, each of which has unique visual characteristics and clinical interpretations. Figure 1 illustrates examples of common medical imaging types that serve as the foundation for developing and applying XAI-based approaches.

## 4. LITERATURE SURVEY

A) Literature Reviews

Brunese, L. — 2020, This paper investigate deep-learning classifiers for detecting pulmonary disease and COVID-19 in chest X-rays while emphasizing explainability. They combine CNN-based feature extraction with visualization methods (e.g., Grad-CAM) to highlight image regions that drive model predictions, arguing that visual saliency maps help clinicians validate automated outputs. The paper reports that explainable visualization improves clinician trust and can reveal when models focus on clinically irrelevant regions, exposing dataset bias or spurious correlations. Brunese stresses the importance of rigorous evaluation of both accuracy and explanation quality, and suggests integrating model explanations in observer studies to quantify how explanations affect diagnostic decisions. The study is widely cited as an early, influential work linking CXR classification performance with XAI evaluation.

Maheswari, B.U. et al. — 2024, This paper propose a compact, interpretable CNN architecture for tuberculosis (TB) screening designed to balance performance with interpretability. Their work uses a shallow-CNN with Bayesian hyperparameter tuning to avoid over-parameterization, and integrates Grad-CAM visualizations to produce human-interpretable heatmaps. The authors evaluate the model on TB datasets and report competitive sensitivity and specificity compared to deeper networks, while arguing the shallow architecture yields easier-to-interpret feature maps. They also discuss deployment

benefits in resource-limited settings (faster inference, lower compute). Their experimental results emphasize that designing for interpretability need not sacrifice diagnostic performance and that explanation quality should be measured alongside accuracy for clinically useful AI tools.

Anderson, P.G. et al. — 2024, This paper validate an FDA-cleared AI chest X-ray system in a clinician-aided reading study. The study demonstrates that when physicians (radiologists and non-radiologists) use the AI tool they detect abnormalities more accurately and faster than unaided reads. Significantly, the AI's probabilistic scores and visual outputs enabled non-radiologists to perform on par with radiologists, demonstrating usefulness in environments lacking specialized access. The paper also documents generalization to external datasets and discusses human–AI collaboration: AI acts as a second reader that raises sensitivity while maintaining specificity. The study is significant because it moves beyond standalone model metrics to show measurable clinical impact of explainable AI assistance in real-world workflows.

Miyazaki, A. et al. — 2023. This paper develop a deep-learning classifier to differentiate COVID-19, pneumonia, and normal chest X-rays. Beyond classification, the authors apply Grad-CAM and Grad-CAM++ to produce heatmaps for each decision and conduct an observer study to evaluate whether these visual explanations align with clinician expectations. Results show that visualization assists radiologists in understanding model decisions and can highlight cases where model focus is clinically plausible. The paper also discusses pitfalls: Grad-CAM heatmaps can be diffuse, and interpretation requires clinical context. The authors recommend coupling heatmaps with region-level annotations and structured clinical features to strengthen trust and clinical relevance of automated CXR systems.

Cid, Y.D. et al. — 2024, This paper describe X-Raydar and X-Raydar-NLP—open-source neural networks for classifying common chest X-ray findings and generating structured radiology outputs. Their retrospective validation shows robust performance across multiple cohorts, and the study emphasizes transparency by releasing code, model weights, and evaluation scripts. Integration of NLP components enables structured report outputs from model predictions, facilitating clinician review. The authors stress open science for reproducibility and raise concerns about dataset shift; they recommend multicentre evaluation and human-in-the-loop workflows. The work contributes a production-oriented, explainability-aware toolkit and documents how open validations can accelerate trustworthy AI adoption in radiology.

Hsieh, C. et al. — 2023, This paper propose MDF-Net, a multimodal architecture that fuses chest X-ray images with patient clinical metadata (age, history, vitals) to improve localization and classification of thoracic abnormalities. The study finds that adding clinical context raises localization precision and disease-level predictive performance versus image-only models. They also analyze how metadata influences model attention maps, showing that clinical variables can steer the model to more clinically relevant regions. The authors argue that multimodal fusion aligns model reasoning with how clinicians synthesize imaging and history, improving explainability and reducing false positives produced by image-only patterns that lack clinical plausibility.

Çallı, E. et al. — 2021, This paper present a comprehensive survey of deep-learning applications for chest X-ray tasks (classification, localization, segmentation, report generation). The review catalogs architectures (CNNs, transformers), datasets (CheXpert, ChestX-ray14, VinDr-CXR), evaluation challenges (label noise, class imbalance), and explainability techniques (saliency maps, attention). A major conclusion is that while accuracy has surged, reproducibility and robust external validation lag behind; the authors stress standardized benchmarks and clinical-centered evaluations. The survey also highlights a trend toward multimodal models and weakly supervised localization, and recommends future work to prioritize interpretability, dataset quality, and prospective clinical testing to transition research prototypes into safe clinical tools.

Lu, Q. et al. — 2025, This paper explore neuro-symbolic Logical Neural Networks (LNNs) to combine neural perception with symbolic reasoning for diagnosis prediction. Their architecture uses neural encoders to extract image-derived predicates (e.g., "opacity_present"), which feed into an LNN that encodes medical rules and ontological constraints to produce logical, traceable diagnoses. Experiments show the neuro-symbolic model yields explanations in rule-based terms, improving transparency compared to black-box classifiers, and in some settings it achieves comparable predictive accuracy while allowing counterfactual queries (e.g., "if opacity absent then diagnosis changes"). The paper argues neuro-symbolic systems can bridge statistical learning with domain knowledge, improving safety and clinician interpretability for medical imaging tasks.

AOR (Anatomical Ontology-Guided Reasoning) team — 2025, The AOR paper introduces an anatomical-ontology guided framework that embeds region-level anatomical knowledge into multimodal chest X-ray models to support multi-step reasoning. The system constrains model attention using an anatomy hierarchy, enabling finer region-wise explanations and stepwise diagnostic rationale (e.g., lesion → region → likely pathology). Evaluations indicate improved detection of subtle findings and more structured, anatomically grounded explanations compared to image-level models. The authors also supply AOR-Instruction, an instruction dataset to train reasoning behaviors. The work demonstrates that coupling ontologies with multimodal models enhances interpretability and aligns model outputs with radiological reasoning patterns.

Liao, Y. et al. — 2023, This paper survey deep-learning methods for automated radiology report generation from chest X-rays, covering encoder–decoder CNN/RNN/transformer pipelines, reinforcement learning for clinical correctness, and retrieval-augmented approaches. Key challenges discussed include aligning visual findings with correct clinical phrasing, avoiding hallucinations, and ensuring factual consistency with image evidence. The review highlights evaluation gaps—standard language metrics (BLEU/ROUGE) poorly reflect clinical correctness—and encourages clinician-driven metrics and error taxonomies. The paper underscores the role of explainability: generating structured intermediate outputs (e.g., detected findings, region labels, and rule-based inferences) improves traceability and reduces risk when reports are used in clinical decision-making.

B) Literature Summary

Significant advancements in the use of deep learning towards medical image processing have been reported recently, particularly in the identification of thoracic illnesses from chest X-rays, such as pneumonia and tuberculosis. CNNs and transfer education, and transformer-based models have achieved

exceptional accuracy in feature extraction and disease classification. However, researchers consistently emphasize the lack of interpretability in such black-box systems, which limits their clinical reliability. Recent studies propose integrating explainable AI (XAI) methods such as Grad-CAM and SHAP to visualize decision regions, yet these approaches only provide partial interpretability. Symbolic AI and knowledge-based systems, on the other hand, show promise in reasoning and aligning with medical ontologies but struggle with large-scale image data. Hybrid neuro-symbolic models combining deep learning and logical reasoning have emerged as a potential solution, offering both accuracy and explainability. Overall, literature supports the development of a unified neuro-symbolic framework to provide transparent, data-driven, and clinically interpretable diagnostics for chest X-ray analysis.

C) Research Gap

- Lack of Interpretability: Most existing chest X-ray diagnostic models rely on deep neural networks that function as black boxes, offering limited transparency for clinical adoption.
- Insufficient Integration of Domain Knowledge: Few models combine medical ontologies or symbolic reasoning with deep learning, leading to predictions that lack alignment with established diagnostic protocols.
- Limited Trust in Clinical Settings: Even though study datasets frequently have great accuracy, doctors in real-world healthcare settings are less trusting and accepting when explainability is lacking.
- Dataset Dependency: Current models are heavily dataset-specific and fail to generalize well across diverse populations and imaging conditions, especially without symbolic guidance.
- Inadequate Human-AI Collaboration: Most frameworks focus only on automated prediction, neglecting human-in-the-loop systems where explanations enhance decision-making.

## 5. RESEARCH METHODOLOGY

A) Criteria for selecting this study:

• Growing Need for Explainability: Interpretable AI models are desperately needed in the healthcare industry to foster clinician trust and guarantee moral diagnostic decision-making.

• Limitations of Existing Deep Learning Systems: Although current CNN-based assessments are highly accurate, their lack of transparency renders them unsuitable for crucial medical applications without explanations that are accessible to humans.

• Clinical Relevance of Chest X-Rays: Chest X-ray imaging is one of the most common, cost-effective, and accessible diagnostic methods for detecting lung and thoracic diseases, justifying its selection for AI-based study.

• Potential of Neuro-Symbolic AI: The integration of neural and symbolic reasoning offers a balanced approach to achieve both accuracy and interpretability in diagnostic results.

• Scope for Research Innovation: Few studies have applied neuro-symbolic frameworks specifically to chest X-ray diagnosis, creating an opportunity to contribute novel methodologies to medical AI research.

• Support for Trustworthy Healthcare AI: The study aligns with global initiatives promoting safe, transparent, and explainable AI in medicine.

B) Method of analysis:

• Data Collection: Chest X-ray image datasets such as NIH ChestX-ray14 and COVIDx are used for training and validation, containing labeled thoracic disease images.

• Preprocessing: Images are resized, normalized, and augmented to enhance data quality and prevent overfitting during training.
• Feature Extraction: Deep learning models, such as CNN or ResNet, are employed to extract visual features representing disease patterns from the X-ray images.
• Symbolic Reasoning Layer: Extracted features are mapped to medical concepts using a knowledge base or ontology that applies logical rules for diagnosis interpretation.
• Model Integration: Neural outputs and symbolic rules are combined to produce both classification results and explainable diagnostic reasoning.
• Performance Evaluation: Metrics like precision, precision, recall, F1-score, along with explainability evaluation through visual as well as rule-based interpretation are used to test the hybrid framework.

C) XAI Techniques in Medical Imaging:

In this study, Explainable machine learning (XAI) techniques were categorized according to the main interpretability techniques that help explain how models produce predictions. These tactics included post hoc interpretability techniques, rule-based reasoning, attention mechanisms, feature visualization, and example-driven explanations. The taxonomy was created by combining knowledge from current scholarly research and modifying it to satisfy the demands of clinical settings for accountability and transparency, especially in medical imaging, where AI-driven decision-making clarity is crucial.

i. Techniques for feature visualization:

The interpretability underlying deep learning systems employed in medical image analysis is greatly enhanced by feature visualization techniques. They visually highlight the areas or attributes of an image that most influence the model's prediction, fostering clinical confidence and transparency. One of the most prevalent methods, Grad-CAM, produces class-specific heatmaps that indicate which regions contributed most to the model's output [15]. This technique has been widely applied across medical domains, such as detecting ADHD from EEG signals [15], diagnosing fractures in the hip from pelvic X-rays and spotting cataracts in fundus photos [16] [17].

More accuracy and resilience in visual explanations are provided by complementary techniques like SmoothGrad and Integrated Gradients. By calculating gradients along a predetermined path from an initial station to the actual input, Integrated Gradients evaluate the influence of each input characteristic and assign priority scores to it [18, 19]. By introducing noise and averaging several gradient calculations, SmoothGrad improves saliency maps, reducing visual noise and highlighting pertinent image elements [19].

Additional visualization strategies, including activation mapping and deconvolution techniques, provide insights into internal network activations, revealing how convolutional layers process information. Enhanced variants such as Layered Grad-CAM integrate visual outputs from multiple layers to improve interpretability in complex models like Feature Pyramid Networks—useful, for instance, in polyp segmentation tasks [20]. Moreover, example-based XAI approaches strengthen clinical interpretability by retrieving comparable instances from the training dataset, enabling practitioners to validate and contextualize AI-driven results [21].

ii. Attention mechanisms:

Attention mechanisms, which improve performance and opacity by directing the model's attention to the most relevant regions of an image, are another crucial interpretability technique in medical imaging.

These are classified as post hoc attentions in general, It integrates attention acquisition directly during training, and trainable concentration, which looks at pre-trained networks to find logical patterns [22]. Models can stay interpretable and preserve diagnostic accuracy thanks to this dual categorization. Particularly for diseases like COVID-19, prostate cancer, lung cancer, & retinal illnesses, attention-based models have attained great precision in identifying diseases and segmentation. For instance, lung cancer classification approaches attained 99.8% accuracy, whereas radiographic Covid-19 methods of detection got up to 98% accuracy. Dual attention and blocking mechanisms are used by segmentation architectures like MDSU-Net to improve feature extraction and delineation. By identifying crucial areas that affect medical diagnosis, transformer-based models using self-attention have further enhanced interpretability[23].

Visualization and comprehension tools including Grad-CAM, LIME, & SHAP, which correlate results with significant aspects to explain predictions, are frequently used in conjunction with these models. For increased transparency, more modern methods like Attention-Gradient Class Activated Mapping (A-GCAM) provide sophisticated attention attribution visualization. The intricacy of attention structures and the lack of conventional clinical validation procedures remain obstacles notwithstanding these developments. Therefore, more research is needed to create attention-based models that attain both clinically meaningful interpretability and diagnostic accuracy [24].

iii. Symbolic and rule-based XAI techniques:

Systems using artificial intelligence (AI) in medical imaging have become more transparent, dependable, and clinically applicable thanks in large part to rule-based and symbolic techniques. These techniques' interpretability and conformity to clinical protocols make them especially appropriate for the healthcare industry. Rule-based models make use of clear and intelligible decision logic, which enables medical professionals to confirm AI results and guarantee adherence to accepted medical standards.

One prominent example is a neurologically symbolic framework for identifying vertebral fractures with compression in CT scans, which integrated a shape-analysis methodology that assessed patterns of vertebral height with deep learning-facilitated vertebral segmentation to establish diagnostic guidelines. The system demonstrated that interpretable systems based on rules may match the performance as black-box models while offering clearer rationale, achieving 96% accuracy with 91% sensitivity. The Adaptive Neuro- Fuzzy In (ANFIS), which combines neural networks in rule-based deduction to optimize therapies and provide comprehensible justifications, is another example. It is used in the planning of intensity-modified Radiation Therapy (IMRT) and increases clinician trust in diagnostic rules [24[25].

By using logical and rule-based representations, symbolic explainable AI (XAI) strategies increase interpretability. For instance, a deep learning model for classifying cancer images used symbolic reasoning to produce user-adaptive interpretations for medical professionals and reached 97.72% accuracy. Similarly, a hybrid model in abdominal CT contrasting phase detection used deep learning and rule-based logic to achieve 92.3% accuracy. Shapley value analysis was used to explain the impact of radiodensity features, increasing clinical insight and transparency [26].

iv. Reasoning based on examples and cases:

By citing actual or prototype situations, example-based and scenario-based reasoning techniques provide intuitive explanations. These techniques improve clinician comprehension by showing how comparable inputs were previously classified. Prototype-based systems, such as ProtoPNet, have been successfully applied in medical domains like brain tumor identification, achieving interpretability without compromising accuracy. Retrieval-based techniques further support explainability by identifying

and displaying similar examples from reference datasets, allowing clinicians to validate predictions. Another approach, counterfactual explanation, illustrates how minor input variations could change predictions, clarifying model decision boundaries and suggesting alternate diagnostic outcomes. Case-Based Reasoning (CBR) expands this idea by leveraging analogical reasoning to solve new cases using prior examples. In breast cancer diagnosis, for instance, CBR systems enhanced classification by visually comparing current findings to historical cases. Hybrid systems integrating CBR and deep learning have demonstrated improved interpretability and diagnostic accuracy in mammogram analysis [27].

D) Difficulties using XAI in Medical Imaging

Despite the promise of XAI in medical imaging, several challenges must be addressed to ensure systems are accurate, ethical, and clinically practical. As AI models grow in complexity, maintaining interpretability becomes increasingly difficult. Many deep learning systems operate as "black boxes," generating predictions without transparent reasoning, which reduces trust and clinical accountability. This opacity limits their acceptance in sensitive healthcare environments where explainability and justification are essential.

A key difficulty lies in managing the trade-off between model interpretability and predictive performance. High-performing models are often complex and opaque, whereas interpretable models may sacrifice accuracy. Achieving an optimal balance between these two aspects remains a critical focus of ongoing research. Another difficulty arises when XAI is included into clinical workflows: explanations have to be timely, pertinent, and intelligible to doctors who are under time pressure. Explanations lose some of their clinical value if they are too abstract or take too long. For practical application, cooperation between AI developers, physicians, and technical personnel is therefore essential.

Barriers also include ethical and legal issues like bias, privacy, even diagnostic risk. Rapid technological development has surpassed legal frameworks, creating oversight and accountability gaps. Even interpretable systems can be harmful if trained on biased or low-quality data. Building trust among clinicians, patients, administrators, and regulators requires transparency regarding data usage, model validation, and system limitations. Moreover, the complexity and variability of medical data continue to pose technical challenges. Ensuring data integrity, contextual relevance, and governance is fundamental to developing trustworthy, explainable AI solutions for clinical applications [27][28].

## 6. APPLICATIONS OF XAI IN MEDICAL IMAGING

*A)* Cardiac imaging

By providing increased accuracy, automatic quantification, and improved workflow efficiency, the use of AI in cardiovascular imaging greatly upgraded diagnostic capabilities. In order to assess calcium scores, measure coronary stenosis, and examine plaque composition, artificial intelligence (AI) algorithms have been extensively used in coronary computed tomographic angiogram (CCTA) [28]. Similar to this, AI helped with the segmentation of heart chambers and the characterisation of myocardial tissue in cardiac electromagnetic resonance imaging (CMR), which were crucial for the diagnosis of cardiomyopathy and myocardial infarction [26]. AI made it easier to segment heart structures in echocardiography in order to evaluate problems in wall motion and valve function [28]. Despite these developments, traditional AI systems' inability to be interpreted hindered their inclusion

into clinical operations, where explainability was crucial for both regulatory compliance and clinical acceptance.

By making the model's decision-making process transparent, XAI overcame this drawback and allowed medical professionals to comprehend, verify, and have faith in AI-generated results. XAI tools like saliency maps, attention processes, and post hoc attributing methodologies assisted in identifying the precise regions or features that influenced model predictions in diagnostic applications. In high-stakes situations like diagnosing myocardial ischemia or categorizing coronary lesions, these explanations were very crucial. AI also showed significant advantages in increasing workflow efficiency by cutting down on image acquisition as well as post-processing time, which sped up diagnostic reporting without sacrificing accuracy [30]. By providing consistent, comprehensible results that were compatible with clinical reasoning, XAI also helped to reduce interobserver variability [50]. Thus, incorporating XAI into cardiac imaging was a crucial step toward reliable, effective, and safe AI-assisted diagnoses.

*B)* Cancer diagnosis.

By providing comprehensible insights that improved the clinical reliability of AI systems, XAI became a crucial part of cancer diagnoses. To overcome the shortcomings of black-box models, XAI was included into image-based and biomarker-driven diagnostic procedures for a variety of cancer types. Convolutional neural networks, or CNNs, and fuzzy logic were coupled in the EXAIOC framework to handle data uncertainties in oral cancer diagnosis. Visual explanations produced by methods like Grad-CAM and Layer-wise Relevance Projection (LRP) enhanced interpretability and aided in clinical decision-making [21]. Explainable deep residual convolutional neural systems (RCNNs) improved with transfer learning successfully collected both time and location information in mammography pictures for the purpose of detecting breast cancer [22].

Through interpretable modeling techniques, XAI also showed promise in the diagnosis of prostate, renal, and liver cancer. The opacity of models based on deep learning was addressed and clinical trust was raised in prostate cancer by using LIME or shape analysis to evaluate MRI-based predictions [23]. CNNs used to high-resolution medical pictures in conjunction with Grad-CAM and LIME produced good diagnostic accuracy for renal cancer diagnosis while offering visual explanations to aid physicians in comprehending model outputs [24]. In a similar vein, better segmentation and more understandable model interpretation were made possible by the combination of U-Net and LIME in cancer in the liver detection, which aided medical professionals in making decisions [25].

## 7. CHALLENGES IN XAI FOR MEDICAL IMAGING

XAI's application in medical imaging has enormous potential to improve diagnostic accuracy and foster clinician confidence. However, in order to guarantee that systems using XAI were not just accurate but also moral, transparent, and useful for clinical usage, a number of crucial issues had to be resolved. The complexity of algorithms grew as AI technology evolved, making it increasingly challenging to create systems that have been both comprehensible and appropriate for use in healthcare environments. The opaque character of many successful AI models, particularly those that use deep learning, was a significant barrier to the adoption of XAI. These models frequently produced forecasts without providing information about how or why choices were made [26, 27]. This lack of openness could erode patient and physician trust in medical settings where responsibility and justification were

crucial [28]. AI tools may be perceived as untrustworthy or inappropriate for clinical decision-making if they lack obvious interpretability.

Managing the compromise between comprehension and forecast accuracy was another major problem. While simpler models are easier to understand but less accurate, complex models that produced high performance were frequently challenging to interpret [30]. Achieving cutting-edge efficiency and guaranteeing clinical usability became tense as a result of this compromise. Finding the ideal balance has continued to be a major goal of XAI research, especially in crucial diagnostic fields. Significant difficulties were also encountered while integrating XAI into current healthcare workflows. AI-generated explanations had to be timely, therapeutically relevant, and comprehensible to medical practitioners under time constraints in order to be helpful [29]. Explanations were likely to be utilized used if they were too abstract, too slow, or unrelated to clinical reasoning. Collaboration between AI developers, physicians, and system engineers was necessary for successful integration to guarantee that results were useful and in line with healthcare requirements.

The application of XAI in clinical imaging was made more difficult by ethical and legal concerns. It was necessary to exercise caution when addressing issues including algorithmic bias, data about patients privacy, and the possibility of inaccurate diagnosis. These difficulties were made worse by the fact that moral standards and legal requirements have not kept up with the quick development of AI technologies. Even explainable models could be harmful without adequate oversight, especially if their results were derived from skewed or subpar data. Finally, the success of XAI in uses in medicine depended on building confidence among all stakeholders, including physicians, patients, healthcare managers, and regulators. It was crucial to have open and transparent communication on the development, training, and validation of models. Technical and data-related constraints also have to be taken into account. Medical data was frequently varied, high-dimensional, and complex. It was still difficult to create interpretable models the fact that could learn from such data. Building trustworthy and comprehensible AI systems for clinical usage required ensuring the security of information, governance, and contextual relevance [27][28][29].

## 8. FUTURE DIRECTION

A) Proposed System

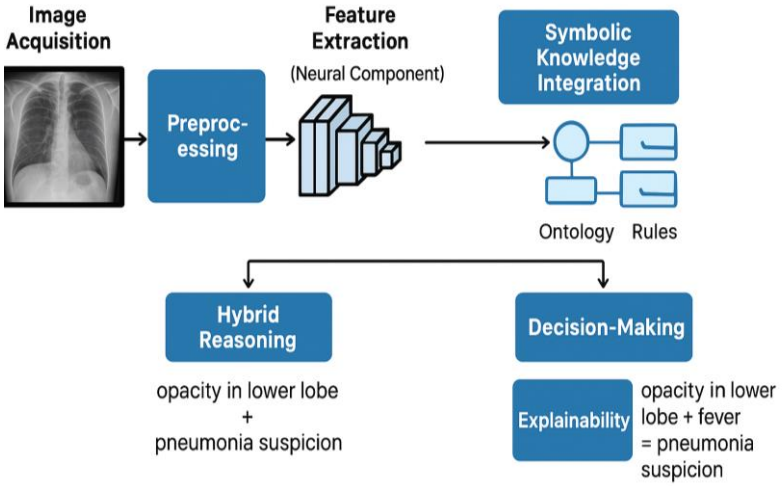


Figure 2. Proposed system

1. Image Acquisition: Chest X-ray images are collected from standard medical imaging datasets or hospital archives, ensuring high-quality, preprocessed inputs for analysis.
2. Preprocessing: Images undergo normalization, resizing, and noise reduction to improve feature extraction. Augmentation techniques may be applied to handle class imbalance and enhance generalization.
3. Features Extraction (Neural Component): Raw images are transformed into structured representations of feature using a neural network based on convolution (CNN), which automatically recovers discriminative visual characteristics such lung opacities, nodules, or aberrant textures.
4. Symbolic Knowledge Integration: Extracted features are mapped to medical ontologies, rules, and diagnostic guidelines within a symbolic reasoning module, enabling knowledge-driven interpretation.
5. Hybrid Reasoning: The neuro-symbolic framework combines CNN predictions with symbolic logic rules (e.g., "opacity in lower lobe + fever = pneumonia suspicion"), ensuring both data-driven learning and human-like reasoning.
6. Decision-Making: The final diagnosis is generated as a combination of probabilistic prediction (from CNN) and rule-based explanation (from symbolic reasoning).
7. Explainability: The system highlights regions of interest on X-rays and provides textual/logical explanations, making outputs interpretable for clinicians.
8. Validation: Predictions are compared against ground truth and medical expert reviews to evaluate accuracy, transparency, and clinical reliability.

B) Tools / Platform to be used

• Dataset Platforms:

Publicly available chest X-ray datasets such as NIH ChestX-ray14, CheXpert, and MIMIC-CXR for training and testing the model.

• Deep Learning Frameworks:

TensorFlow or PyTorch for building CNN architectures, training models, and feature extraction.
Keras as a high-level API for rapid prototyping.

• Symbolic Reasoning Tools:

Prolog or PyKE (Python Knowledge Engine) for encoding diagnostic rules and ontologies.
OWL / Protégé for ontology design to integrate medical domain knowledge.

• Programming Environment:

Using modules like NumPy, Pandas, & Scikit-learn for data management and analysis, Python serves as the main programming language.

• Visualization & Explainability:

Grad-CAM, LIME, or SHAP for generating interpretable explanations and heatmaps of CNN predictions.

• Hardware & Platforms:

Google Colab / Kaggle Notebooks for cloud-based experiments with GPU support.
NVIDIA GPUs for local high-performance training.

• Evaluation Tools:

Statistical metrics (AUC, F1-score, sensitivity, specificity) using Scikit-learn for performance evaluation.

C) Algorithm used

Input Acquisition:

- Chest X-ray images are collected from publicly available datasets such as NIH ChestX-ray14, CheXpert, and COVIDx.
- Each image is labeled with corresponding thoracic disease categories such as pneumonia, tuberculosis, or lung opacity.

Preprocessing:

- Contrast-limited adaptive equalization of histograms (CLAHE) is used to enhance, normalize, and resize images.
- To boost dataset diversity, data augmentation methods including rotation, flipping, and scale are used.

Feature Extraction using CNN:

- Deep learning models like ResNet50, DenseNet121, or EfficientNet are used to extract high-level visual features.
- The network is trained using transfer learning to reduce computational cost and improve accuracy.

Symbolic Reasoning Module:

- The extracted features are translated into symbolic representations linked to medical ontologies such as Unified Medical Language System (UMLS).
- Logical inference rules are applied (e.g., "If opacity + high temperature → pneumonia") to interpret CNN predictions.

Integration (Neuro-Symbolic Fusion):

- The neural and symbolic outputs are fused to produce interpretable results, combining probability-based and rule-based reasoning.

Evaluation and Explainability:

- Performance is measured using metrics like accuracy, F1-score, and ROC-AUC.
- Explainability is validated using Grad-CAM visualizations and rule-based explanations from the symbolic layer.

D) Datasets Used

• NIH ChestX-ray14: Contains 112,120 frontal chest X-ray images with 14 disease labels.
• CheXpert Dataset: Includes over 24,000 images annotated for 14 thoracic conditions.
• MIMIC-CXR: Over 47,000 chest radiographs with structured clinical reports for semantic reasoning integration.

## 9. ADVANTAGES

• Improved Accuracy: Combines deep learning feature extraction with symbolic reasoning to enhance diagnostic reliability.
• Explainability: Provides human-understandable justifications for predictions, addressing the black-box problem of deep learning.
• Clinical Trust: Increases physician confidence by aligning AI outputs with established diagnostic guidelines.

• Knowledge Integration: Incorporates medical ontologies and rules, enabling reasoning beyond statistical correlations.
• Adaptability: Symbolic reasoning can be updated with new medical guidelines without retraining the neural model.
• Transparency: Highlights image regions influencing decisions, supporting clinical validation.
• Human-AI Collaboration: Supports doctors as a decision-aid tool rather than a replacement.

## 10. SCOPE OF STUDY

• Disease Coverage: Focuses on detecting pneumonia, tuberculosis, lung cancer, and other thoracic diseases using chest X-rays.
• Framework Applicability: Can be extended to other imaging modalities such as CT scans or MRI.
• Healthcare Settings: Useful in both advanced hospitals and resource-limited settings with limited radiology expertise.
• Educational Impact: Can support training radiologists by providing interpretable diagnostic pathways.
• Scalability: Capable of integration into clinical workflows and telemedicine platforms.
• AI Research: Contributes to neuro-symbolic AI development for medical applications.
• Data Expansion: Potential to integrate multimodal data, including clinical history and lab results.

## 11. CONCLUSION

The study comes to the conclusion that although deep learning models have transformed medical picture diagnosis with remarkable accuracy, their practical use in clinical settings is limited by their lack of interpretability. A possible approach to attaining both explainability and diagnostic precision is the integration of neurologically symbolic AI. Such hybrid frameworks can generate findings that are not only efficient but also accessible and clinically interpretable by fusing the reasoning ability of symbolic logic with the perceptual qualities of neural networks. Studies reveal that this approach enhances trust between AI systems and healthcare professionals, ensuring accountability and reliability in medical decision-making. Future research should focus on developing robust neuro-symbolic architectures, expanding annotated medical images, and establishing standardized evaluation frameworks for interpretability. Overall, neuro-symbolic models represent a significant advancement toward creating ethical, trustworthy, and explainable AI solutions in medical imaging and diagnostic support systems.

The appropriate and successful integration of AI in medical imaging is greatly aided by Explainable Artificial Intelligence (XAI). This review demonstrates how XAI methods can close the gap between the need for interpretability in healthcare settings and high-accuracy deep learning systems. Although substantial advancements have been made, persistent challenges—such as limited data quality, ethical issues, and the inherent complexity of AI models—still hinder widespread adoption. Future work should concentrate on developing XAI frameworks that are comprehensible, reliable, morally sound, and easily integrated with clinical workflows. To improve clinical usability, advancements in collaborative model building, example-based explanations, and visualization techniques will be essential. In the end, the continuous development of XAI in medical images promises to improve accountability, transparency, and confidence in AI-driven healthcare.

## REFERENCES

1. L. Brunese, F. Mercaldo, A. Reginelli, and A. Santone, "Explainable deep learning for pulmonary disease and COVID-19 detection from chest X-rays," Computer Methods and Programs in Biomedicine, vol. 196, p. 105608, 2020.
2. B. U. Maheswari, G. Revathy, and V. Rajinikanth, "Explainable deep-neural-network supported scheme for tuberculosis screening from chest X-rays," BMC Medical Imaging, vol. 24, no. 58, 2024.
3. P. G. Anderson, C. Zhang, and J. Luo, "Deep learning improves physician accuracy in the detection of chest X-ray abnormalities," Scientific Reports, vol. 14, no. 3522, 2024.
4. A. Miyazaki, T. Koyama, and K. Nishiyama, "Computer-aided diagnosis of chest X-ray for COVID-19 using explainable deep learning visualization," Scientific Reports, vol. 13, no. 1432, 2023.
5. Y. D. Cid et al., "Development and validation of open-source deep neural networks for chest X-ray classification (X-Raydar)," The Lancet Digital Health, vol. 6, no. 2, pp. e110–e122, 2024.
6. C. Hsieh, M. Liu, and S. Y. Lin, "MDF-Net: Fusing chest X-rays with clinical metadata for improved abnormality localization," Scientific Reports, vol. 13, no. 876, 2023.
7. E. Çallı, E. Sogancioglu, B. van Ginneken, K. G. Murphy, and C. H. Sánchez, "Deep learning for chest X-ray analysis: A survey," Image and Vision Computing, vol. 113, p. 104229, 2021.
8. Q. Lu, Y. Li, and H. Wang, "Explainable diagnosis prediction through neuro-symbolic logical neural networks," arXiv preprint arXiv:2502.01840, 2025.
9. X. Zhang, L. Chen, and Y. Wang, "Anatomical ontology-guided reasoning for medical large multimodal models in chest X-ray interpretation," arXiv preprint arXiv:2502.00294, 2025.
10. Y. Liao, X. Xu, and Z. Zhang, "Deep learning approaches to automatic radiology report generation: A review," Computers in Biology and Medicine, vol. 152, p. 106380, 2023.
11. S. Bhushan and S. Dixit, "Explainable AI for shaping adoption of artificial intelligence," in Proc. Asian Conf. on Intelligent Technologies (ACOIT), pp. 1–6, 2024.
12. U. B. Khakurel and D. B. Rawat, "Evaluating explainable artificial intelligence (XAI): Algorithmic explanations for transparency and trustworthiness of ML algorithms and AI systems," in Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications IV, SPIE, Bellingham, WA, USA, p. 7, 2022.
13. R. Nyrup and D. Robinson, "Explanatory pragmatism: A context-sensitive framework for explainable medical AI," Ethics and Information Technology, vol. 24, no. 1, p. 13, 2022.
14. N. Ullah, F. Guzmán-Aroca, F. Martínez-Álvarez, I. De Falco, and G. Sannino, "A novel explainable AI framework for medical image classification integrating statistical, visual, and rule-based methods," Medical Image Analysis, vol. 105, p. 103665, 2025.
15. B. Latifi, A. Amini, and A. Motie Nasrabadi, "Siamese-based deep neural network for ADHD detection using EEG signal," Computers in Biology and Medicine, vol. 182, p. 109092, 2024.
16. H. Shah, R. Patel, S. Hegde, and H. Dalvi, "XAI meets ophthalmology: An explainable approach to cataract detection using VGG-19 and Grad-CAM," in IEEE Pune Section Int. Conf. (PuneCon), pp. 1–8, 2023.
17. S.-L. Chung, C.-T. Cheng, C.-H. Liao, and I.-F. Chung, "Patch-based feature mapping with generative adversarial networks for auxiliary hip fracture detection," Computers in Biology and Medicine, vol. 186, p. 109627, 2025.

18. Z. Papanastasopoulos et al., “Explainable AI for medical imaging: Deep-learning CNN ensemble for classification of estrogen receptor status from breast MRI,” in Medical Imaging 2020: Computer-Aided Diagnosis, SPIE, Bellingham, WA, USA, p. 52, 2020.
19. P. Narayankar and V. P. Baligar, “Explainability of brain tumor classification based on region,” in Int. Conf. on Emerging Technologies in Computer Science for Interdisciplinary Applications, pp. 1–6, 2024.
20. S. M. Javali, R. S. Upadhyayula, and T. De, “Comparative study of xAI layer-wise algorithms with a robust recommendation framework of inductive clustering for polyp segmentation and classification,” in Int. Seminar on Machine Learning, Optimization, and Data Science (ISMODE), pp. 325–330, 2022.
21. M. Fontes, J. D. S. De Almeida, and A. Cunha, “Application of example-based explainable artificial intelligence (XAI) for analysis and interpretation of medical imaging: A systematic review,” IEEE Access, vol. 12, pp. 26419–26427, 2024.
22. H. Shin, J. Lee, T. Eo, Y. Jun, S. Kim, and D. Hwang, “The latest trends in attention mechanisms and their application in medical imaging,” Journal of the Korean Society of Radiology, vol. 81, no. 6, p. 1305, 2020.
23. G. W. Muoka, D. Yi, C. C. Ukwuoma, M. D. Martin, A. A. Aydin, and M. A. Al-Antari, “A novel attention-based explainable deep learning framework towards medical image classification,” in 7th Int. Symp. on Innovative Approaches in Smart Technologies, pp. 1–8, 2023.
24. Y. Zhou, X. Kang, F. Ren, H. Lu, S. Nakagawa, and X. Shan, “A multi-attention and depthwise separable convolution network for medical image segmentation,” Neurocomputing, vol. 564, p. 126970, 2024.
25. S. Chai et al., “A novel adaptive hypergraph neural network for enhancing medical image segmentation,” in Lecture Notes in Computer Science, pp. 23–33, 2024.
26. S. Suara, A. Jha, P. Sinha, and A. A. Sekh, “Is Grad-CAM explainable in medical images?,” in Lecture Notes in Computer Science, pp. 124–135, 2024.
27. L. Chen, X. Cai, Z. Li, J. Xing, and J. Ai, “Where is my attention? An explainable AI exploration in water detection from SAR imagery,” International Journal of Applied Earth Observation and Geoinformation, vol. 130, p. 103878, 2024.
28. D. Kumar, S. Porwal, R. Malviya, and S. B. Sridhar, “XAI technique in deep learning–based medical image analysis,” in Explainable and Responsible Artificial Intelligence in Healthcare, Wiley, New Jersey, USA, pp. 191–215, 2025.
29. B. Inigo, O. Colliot, and J. Mitra, “An intrinsically explainable approach to detecting vertebral compression fractures in CT scans via neurosymbolic modeling,” in Medical Imaging 2025: Image Processing, SPIE, p. 101, 2025.
30. X. Gonzalez-Garcia, J. Fumanal-Idocin, J. M. N. Do Rio, and H. Bustince, “A rule-based approach for interpretable intensity-modulated radiation therapy treatment selection,” in IEEE Int. Conf. on Fuzzy Systems, pp. 1–8, 2024.