Neuro Symbolic AI and Contrastive Learning for Medical Image Diagnosis

Krishnan Venkiteswaran
Department of Artificial Intelligence and Machine Learning
SIES Graduate School of Technology
Mumbai, India
krishnanpaiml121@gst.sies.edu.in

Sahil Brid
Department of Artificial Intelligence and Machine Learning
SIES Graduate School of Technology
Mumbai, India
sahilbaiml121@gst.sies.edu.in

Omkar Shivarkar
Department of Artificial Intelligence and Machine Learning
SIES Graduate School of Technology
Mumbai, India
omkarssaiml121@gst.sies.edu.in

Adhiraj More
Department of Artificial Intelligence and Machine Learning
SIES Graduate School of Technology
Mumbai, India
adhirajmaiml121@gst.sies.edu.in

Dr. Varsha Patil
Department of Artificial Intelligence and Machine Learning
SIES Graduate School of Technology
Mumbai, India
varshap@sies.edu.in

***Abstract*—This paper presents a novel neurosymbolic framework for chest X-ray analysis that integrates deep learning with symbolic reasoning to bridge the interpretability gap in medical imaging AI. Our approach combines visual information from X-ray images, textual data from radiology reports, and explicit medical knowledge in a unified architecture. The framework consists of five core components: an image encoder leveraging a modified ResNet50, a text encoder based on T5 transformer, a symbolic knowledge module encoding anatomical-finding relationships, a multimodal fusion mechanism, and a report generation model. We trained the model on the MIMIC-CXR dataset using a multi-objective loss function that addresses contrastive learning, symbolic consistency, and finding classification. Experimental results demonstrate competitive performance with an F1 score of 0.429 for finding classification and a BLEU-4 score of 0.116 for report generation, comparable to or exceeding state-of-the-art methods. The key advantage of our approach lies in its transparent reasoning process through anatomical attention visualizations and explicit symbolic relationships. Ablation studies confirm that each component contributes significantly to overall performance, with the symbolic knowledge module and anatomical attention mechanism proving especially crucial. This work represents a significant advancement toward interpretable AI systems for radiology that can provide not only accurate diagnoses but also understandable explanations that align with clinical reasoning processes, potentially enhancing trust and adoption in healthcare settings.**

***Index Terms*—Artificial Intelligence (AI), Medical Imaging, Neuro-Symbolic AI, Contrastive Learning, Multimodal Framework, Symbolic Reasoning, Self-Supervised Learning**

I. Introduction

Artificial Intelligence (AI) has revolutionized several industries, with healthcare and medical diagnostics standing out due to their potential to significantly improve patient outcomes. The field of medical imaging, in particular, has benefited greatly from AI-driven innovations. However, traditional deep learning methods often operate as "black boxes," providing little transparency or explanation for their decisions. This opacity poses significant challenges in domains like healthcare, where interpretability, trustworthiness, and explainability are paramount for clinical decision-making.

In response to these challenges, Neuro Symbolic AI—an emerging paradigm combining neural networks with symbolic reasoning—has gained traction. By fusing the pattern recognition strengths of neural networks with the logical reasoning capabilities of symbolic AI, Neuro Symbolic models offer both efficiency and interpretability, addressing limitations found in purely neural or symbolic approaches.

Recent works [1]-[3] highlight the potential of Neuro Symbolic AI to enhance interpretability in AI systems, especially in complex tasks such as medical diagnosis. Neuro Symbolic AI has shown promise in various fields, including medical image analysis, by balancing accuracy with explainability, thus making AI-assisted diagnostic systems more trustworthy for clinicians.

Additionally, contrastive learning, a self-supervised learning technique, has gained attention for its effectiveness in learning meaningful representations from unlabeled data. By contrasting positive and negative sample pairs, contrastive learning enhances the model's ability to distinguish between different classes, which is particularly beneficial in medical image analysis. Studies have demonstrated the applicability of contrastive learning in medical imaging tasks [4]-[5].

This review explores the integration of Neuro-Symbolic AI and contrastive learning in medical image diagnosis, focusing on methodologies, applications, challenges, and future directions.

II. Related Work

A. Traditional Medical Image Diagnosis

Several studies have explored the role of AI in medical imaging, particularly in enhancing diagnostic accuracy. For example, [6] discuss the application of AI in Computer-Aided Diagnosis (CAD), emphasizing the improvements AI brings in diagnostic accuracy, efficiency, and speed. Their paper highlights the use of deep learning techniques such as convolutional neural networks (CNNs) for tasks like image segmentation, detection, and classification. CAD systems can help detect lesions and perform quantitative analyses, showing considerable improvements in areas like lung nodule detection and liver tumor screening.

CNNs have been instrumental in medical image analysis. Their review focuses on CNN architectures such as AlexNet, ResNet, and VGGNet, and their application in tasks like classification, detection, segmentation, and image enhancement. These models reduce the need for manual feature engineering, demonstrating significant improvements in areas such as image segmentation and disease detection [7].

Despite the success of AI-based systems, limitations persist. One of the main challenges is the lack of high-quality, annotated medical datasets, which limits the reliability of deep learning models. Several papers, point out that the lack of standardized evaluation protocols also complicates the deployment of AI in clinical settings. The complexity of medical images, especially those with minimal differences between pathological and healthy areas, further hinders AI performance in specific cases [8].

Some papers propose Sino-CT-Fusion-Net, a framework designed for the detection and classification of intracranial hemorrhages. This system integrates sinogram data (raw X-ray data) with CT images to enhance diagnostic accuracy. The fusion of data types provides significant improvements in detection rates, especially in identifying smaller hemorrhages. This study illustrates the potential of multi-modal data fusion in improving diagnostic outcomes [9].

Some have developed a CNN-Transformer hybrid model for bladder cancer detection using cystoscopy images. The model achieved high accuracy, demonstrating the effectiveness of integrating transformers for capturing global context. However, the study noted limitations such as the small dataset size and the lack of clinical validation [10].

The COVID-19 pandemic has accelerated the use of AI in healthcare. Authors discussed the COV19D Competition, which aimed to improve COVID-19 detection using 3-D chest CT scans. The top-performing models achieved impressive results, highlighting the importance of domain adaptation techniques to generalize across different hospitals and medical centers [11].

Papers focused on pneumonia detection from chest X-ray images using CNNs. The study compared several CNN architectures, with a custom-tuned CNN achieving the highest accuracy (83.16%). The research highlights the importance of model pre-processing and tuning to enhance performance [12].

Another notable study explores brain tumor classification using MRI scans. The paper evaluates multiple models, including CNNs and logistic regression, with the CNN achieving the highest accuracy. However, the authors note the need for external validation and comparison with human radiologists [13].

B. Applications of Neuro-Symbolic AI and Contrastive Learning in Medical Image Diagnosis

Medical image diagnosis relies heavily on the interpretability of the models used. Recent studies have shown that Neuro Symbolic approaches can address the shortcomings of traditional AI by providing explainable, rule-based outputs. Studies emphasize the importance of balancing the computational demands of neural networks with the interpretability offered by symbolic systems. Work on models like Neuro-Symbolic Concept Learner (NSCL) and Neuro-Symbolic Dynamic Reasoning (NS-DR) demonstrates how these hybrid approaches outperform purely neural methods in terms of transparency [14].

Studies further supports this by comparing post-hoc explanation methods such as SHAP and LIME with neural-symbolic rule extraction methods. While SHAP and LIME provide local explanations, they often suffer from inconsistency and computational inefficiency. In contrast, neural-symbolic approaches provide clearer, more actionable insights, which are essential in medical image diagnosis [15].

Contrastive learning further enhances interpretability by learning representations that distinguish between different classes. Authors introduce a framework for contrastive learning of visual representations, demonstrating its effectiveness in various tasks [16]. In medical imaging, authors applied contrastive learning to chest X-rays, improving diagnostic performance by leveraging unlabeled data [17].

C. Hybrid Models for Medical Image Segmentation and Diagnosis

Hybrid Neuro Symbolic models, which combine deep learning with symbolic reasoning, have shown great promise in medical imaging tasks like segmentation, tumor detection, and disease classification.Studies highlight the potential of hybrid models that use neural embeddings along with symbolic reasoning frameworks, which have proven effective in reasoning over knowledge graphs [18].

Studies discussed the implementation of convolutional neural networks (CNNs) for image segmentation and object detection, integrated with symbolic reasoning. Their AI-based Computer-Aided Diagnosis (CAD) system assists clinicians in identifying lesions and conducting quantitative analyses, reducing diagnosis times and increasing accuracy.

In the realm of medical image segmentation, U-Net has been a foundational architecture. Recent advancements have seen the integration of symbolic AI into such architectures to enhance performance. For instance, symbolic AI's rule-based segmentation has been applied to analyze anatomical structures, improving the precision of boundary delineation in complex medical images.

Furthermore, the integration of Neuro-Symbolic AI in brain tumor diagnosis has shown potential. By combining CNNs with symbolic reasoning, models can better handle the complexity and variability inherent in brain imaging, leading to more accurate and interpretable diagnostic outcomes.

D. Enhancing Scalability and Efficiency

Contrastive learning contributes to scalability by enabling models to learn robust representations from unlabeled data, reducing the dependency on extensive labeled datasets. This approach is particularly beneficial in medical imaging, where labeled data is often scarce. Studies have demonstrated that contrastive learning frameworks can effectively utilize unlabeled data to improve diagnostic performance across various medical image datasets [19].

For the detection of Alzheimer's disease, researchers have developed a hybrid model that combines Graph Convolutional Neural Networks (GCNN) and CNNs. This model utilizes the Alzheimer's dataset, which includes MRI images across four categories: mildly demented, moderately demented, non-demented, and very mildly demented. To address sample imbalances, various data augmentation techniques were applied. The model was rigorously evaluated against established pre-trained models, such as VGG19, ResNet50, AlexNet, and DenseNet-121, with performance metrics encompassing accuracy, sensitivity, precision, and F1-score [20]

The research on liver cancer detection highlights a comprehensive preprocessing pipeline that includes essential steps such as normalization, noise reduction, contrast enhancment, and artifact removal. This thorough approach is critical for ensuring the quality and consistency of the input data, which directly influences the performance of the machine learning models. The study also compares multiple CNN architectures, including VGG16, ResNet50, and MobileNet, providing valuable insights into their relative strengths and weaknesses in the context of liver cancer detection. By evaluating both CT and MRI scans, the research acknowledges the distinct advantages of each imaging modality, allowing for a more nuanced understanding of liver cancer. Moreover, the paper emphasizes the paramount importance of early detection in improving patient outcomes, thereby linking the technical aspects of machine learning directly to practical clinical benefits. By highlighting these innovations and approaches, the research aims to contribute significantly to the advancement of cancer detection methodologies, ultimately enhancing diagnostic accuracy and patient care [21].

Other systems have also utilized various architectures, such as AlexNet and GoogLeNet, alongside a custom 23-layer CNN model specifically for the detection and classification of brain tumors using MRI and CT scan images [22].

The author presents a novel end-to-end methodology for the detection of malaria parasites, specifically designed to harness the strengths of high-cost microscope (HCM) images during training while ensuring robust performance on low-cost microscope (LCM) images during testing. This approach addresses a critical challenge in the field: the difficulty and expense associated with annotating LCM images, which often results in limited training data. By leveraging the clearer, high-resolution HCM images, the proposed method aims to improve the accuracy and reliability of malaria detection in more accessible and cost-effective imaging modalities. The core of this innovative framework, named CodaMal (COntrastive Domain Adaptation for MALaria), employs a CSP-DarkNet53 backbone for object detection, which is well-suited for extracting meaningful features from complex imaging data. To effectively bridge the gap between the differing domains of HCM and LCM images, the authors introduce a Domain Adaptive Contrastive (DAC) loss. This loss function is specifically designed to minimize the domain discrepancy between the two image types, thereby enhancing the model's ability to generalize across varying image quality and conditions. In addition to the DAC loss, the framework incorporates standard object detection losses, including classification, localization, and objectness, to ensure comprehensive training and accurate predictions. By integrating these elements, the model is capable of effectively identifying and localizing malaria parasites within the images. Furthermore, a non-linear projection layer is utilized to map features to a lower-dimensional latent space, facilitating improved representation learning and enhancing the model's overall performance. This multi-faceted approach positions CodaMal as a significant advancement in malaria parasite detection, paving the way for more accessible and efficient diagnostic solutions in resource-limited settings [23].

E. Challenges

While Neuro-Symbolic AI and contrastive learning offer promising advancements, several challenges remain. One major issue is the integration of neural and symbolic components. The symbolic components, although crucial for interpretability, often introduce bottlenecks in terms of computational efficiency, particularly in real-time medical applications.

Moreover, studies emphasize the scalability issues faced by Neuro Symbolic AI when applied to large medical datasets, as seen in their work on knowledge graph reasoning. Additionally, there are concerns about the generalization of Neuro Symbolic models, as they may struggle with the inherent complexity of medical images, such as subtle anatomical variations [24].

Contrastive learning methods, while effective, require careful consideration in the selection of positive and negative pairs to ensure meaningful representation learning. In medical imaging, where inter-class similarities are high, defining these pairs can be challenging. Moreover, the reliance on large

amounts of unlabeled data necessitates efficient data handling and processing capabilities.

III. Methodology

A. Overview of the Neurosymbolic Chest X-ray Analysis Framework

Our research introduces a comprehensive neurosymbolic framework for chest X-ray analysis that bridges the gap between deep learning and symbolic reasoning. The architecture integrates visual information from X-ray images with textual information from radiology reports, while incorporating explicit medical knowledge through a symbolic reasoning component. This multi-faceted approach allows for not only accurate diagnosis but also interpretable reasoning and transparent decision-making processes that are essential in clinical settings.

The proposed framework consists of five core components: (1) an image encoder that extracts meaningful features from chest X-rays, (2) a text encoder that processes radiology reports, (3) a symbolic knowledge module that encodes relationships between anatomical regions and findings, (4) a multimodal fusion mechanism that combines these different information streams, and (5) a report generation model that produces human-readable impressions. These components work in concert to create a system that leverages the strengths of both neural networks (for pattern recognition) and symbolic reasoning (for explicit knowledge representation).

B. Data Acquisition and Preprocessing

Our methodology leverages the MIMIC-CXR dataset, a comprehensive collection of chest X-ray images paired with corresponding radiology reports. Each report typically contains two key sections: "Findings," which details observations from the image, and "Impression," which summarizes the diagnostic conclusions. This paired data serves as the foundation for our multimodal approach.

1) Image Preprocessing: All chest X-ray images undergo a standardized preprocessing pipeline to ensure consistency and compatibility with our deep learning models:

- Conversion to RGB format for compatibility with pretrained models
- Resizing to a uniform dimension of 224×224 pixels
- Normalization using the ImageNet statistics (mean = [0.485, 0.456, 0.406], std = [0.229, 0.224, 0.225])
- Data augmentation during training, including random horizontal flips, slight rotations ($\pm 10°$), and minor brightness/contrast adjustments to improve model robustness

The preprocessing transformation can be formalized as:

$$x_{processed} = \frac{x_{resized} - \mu}{\sigma} \tag{1}$$

Where μ and σ represent the channel-wise mean and standard deviation vectors from ImageNet statistics.

2) Text Preprocessing: The radiology reports undergo several preprocessing steps to extract meaningful information:

- Extraction of "Findings" and "Impression" sections from the full reports
- Tokenization using a T5 tokenizer (specifically from the google/flan-t5-base model)
- Padding or truncation to a maximum sequence length of 128 tokens
- Creation of binary labels for common findings by searching for their presence in the "Findings" section

For each report, we generate a binary label vector $y \in \{0, 1\}^F$ where F is the number of possible findings (in our implementation, we track 14 common findings including pneumonia, effusion, cardiomegaly, etc.). A finding is considered present ($y_i = 1$) if the corresponding term appears in the report text.

C. Multi-component Architecture Design

1) Image Encoder: The image encoder extracts meaningful visual features from chest X-ray images. We employ a ResNet50 architecture pretrained on ImageNet and adapt it specifically for chest X-ray analysis:

- The fully connected classification layer is removed and replaced with an identity function
- Feature maps from the final convolutional block (layer4) are preserved for both feature extraction and visualization purposes
- Global average pooling is applied to get a compact representation

The flattened feature vector then passes through a projection network to align it with our shared embedding space:

$$e_{img} = \text{Normalize}(W_2 \cdot \text{ReLU}(\text{BN}(W_1 \cdot f_{flat} + b_1)) + b_2) \tag{2}$$

Where f_{flat} is the flattened feature vector (2048-dimensional for ResNet50), $W_1 \in \mathbb{R}^{2048 \times 1024}$, $W_2 \in \mathbb{R}^{1024 \times d}$, d is the shared embedding dimension (384 in our implementation), BN refers to Batch Normalization, and Normalize applies L2 normalization to produce unit-length embeddings.

Additionally, we incorporate an anatomy attention module that learns to focus on different anatomical regions:

$$a_{anatomy} = \sigma(W_a \cdot \text{ReLU}(\text{LN}(W_b \cdot f_{flat} + b_b)) + b_a) \tag{3}$$

Where $W_a \in \mathbb{R}^{h \times R}$, $W_b \in \mathbb{R}^{2048 \times h}$, h is a hidden dimension, R is the number of anatomical regions (9 in our implementation), LN refers to Layer Normalization, and σ is the sigmoid activation function. The resulting vector $a_{anatomy} \in [0, 1]^R$ represents attention weights for each anatomical region.

2) Text Encoder: For encoding the radiology reports, we utilize the encoder portion of a pretrained T5 transformer model (google/flan-t5-base):

- Tokenized report text is passed through the T5 encoder
- The resulting hidden states are processed using masked mean pooling to handle variable-length sequences
- A projection network aligns the text features with the shared embedding space

The masked mean pooling operation can be formalized as:

$$e_{text\_raw} = \frac{\sum_{i=1}^{L} h_i \cdot m_i}{\sum_{i=1}^{L} m_i} \tag{4}$$

Where h_i is the hidden state for token i, m_i is the attention mask value (1 for real tokens, 0 for padding), and L is the sequence length.

The final text embedding is computed as:

$$e_{text} = \text{Normalize}(W_4 \cdot \text{ReLU}(\text{LN}(W_3 \cdot e_{text\_raw} + b_3)) + b_4) \tag{5}$$

Where $W_3 \in \mathbb{R}^{768\times512}$, $W_4 \in \mathbb{R}^{512\times d}$, and LN refers to Layer Normalization.

3) Symbolic Knowledge Module: The symbolic component of our architecture encodes explicit medical knowledge about relationships between anatomical regions and radiological findings. This module provides transparency and interpretability to the diagnostic process. Initially, we construct a knowledge base in the form of a relation matrix $M \in \mathbb{R}^{R\times F}$, where R is the number of anatomical regions and F is the number of possible findings. Each entry $M_{i,j}$ represents the strength of association between region i and finding j (Fig. 1). This matrix is populated based on established medical knowledge (e.g., pneumonia is strongly associated with lung regions, cardiomegaly with the heart region).

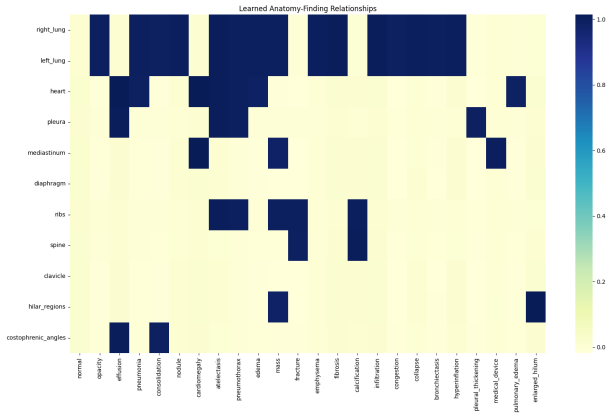

Fig. 1. Learned Anatomy-Finding Relationships

Crucially, this relation matrix is parameterized as a learnable tensor, allowing the model to refine these relationships during training:

$$M_{learned} = \text{Parameter}(M_{initial}) \tag{6}$$

The symbolic module produces finding predictions by combining the anatomy attention weights with the learned relation matrix:

$$p_{symbolic} = a_{anatomy} \cdot M_{learned} \tag{7}$$

This operation can be interpreted as: "If the model attends to region i with weight a_i, and region i is related to finding j with strength $M_{i,j}$, then finding j is predicted with a strength proportional to $a_i \cdot M_{i,j}$."

To enhance flexibility, we incorporate a neural refinement component:

$$p_{neural} = \text{MLP}(p_{symbolic} \odot r_{importance}) \tag{8}$$

$$g = \sigma(\text{MLP}(p_{symbolic} \odot r_{importance})) \tag{9}$$

$$p_{final} = g \odot (p_{symbolic} \odot r_{importance}) + (1-g) \odot p_{neural} \tag{10}$$

Where $r_{importance} \in \mathbb{R}^F$ is a learnable vector of importance weights for each finding, $\odot$ represents element-wise multiplication, and $g \in [0,1]^F$ is a gate vector that controls the balance between pure symbolic reasoning and neural refinement for each finding.

4) Multimodal Fusion and Finding Classification: Our architecture employs two complementary paths for finding classification:

- A direct classification path that predicts findings directly from image embeddings:

$$p_{direct} = W_6 \cdot \text{ReLU}(\text{Dropout}(W_5 \cdot e_{img} + b_5)) + b_6 \tag{11}$$

- The symbolic path described above that leverages anatomical attention and medical knowledge:

$$p_{symbolic} = a_{anatomy} \cdot M_{learned} \cdot r_{importance} \tag{12}$$

During inference, these predictions are combined:

$$p_{combined} = \frac{p_{direct} + p_{final}}{2} \tag{13}$$

$$p_{findings} = \sigma(p_{combined}) \tag{14}$$

Where σ is the sigmoid activation function, producing probabilities for each possible finding.

For report generation, we further combine information by concatenating the image embeddings with the symbolic output and projecting to the T5 decoder dimension:

$$h_{combined} = [e_{img}; p_{final}] \tag{15}$$

$$h_{decoder} = W_7 \cdot h_{combined} + b_7 \tag{16}$$

D. Learning Objectives and Training Strategy

Our model is trained using a multi-objective loss function that addresses several learning tasks simultaneously:

1) Contrastive Learning Loss: To align the image and text embeddings in the shared space, we employ a bidirectional contrastive loss (InfoNCE):

$$L_{contrastive} = \frac{1}{2}(L_{i2t} + L_{t2i}) \tag{17}$$

Where:

$$L_{i2t} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp(e_{img}^i \cdot e_{text}^i/\tau)}{\sum_{j=1}^{N} \exp(e_{img}^i \cdot e_{text}^j/\tau)} \tag{18}$$

$$L_{t2i} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp(e_{text}^i \cdot e_{img}^i/\tau)}{\sum_{j=1}^{N} \exp(e_{text}^i \cdot e_{img}^j/\tau)} \tag{19}$$

Here, N is the batch size, e_{img}^i and e_{text}^i are the normalized embeddings for the i-th sample, and τ is a temperature parameter (set to 0.07 in our implementation).

2) Symbolic Consistency Loss: To train the symbolic knowledge module, we apply binary cross-entropy between the symbolic predictions and ground truth finding labels:

$$L_{symbolic} = \text{BCE}(p_{final}, y_{findings}) \tag{20}$$

Where $y_{findings}$ is the binary ground truth vector for findings. To address class imbalance (many findings are rare), we use class-specific positive weights inversely proportional to class frequency:

$$w_j = \frac{N}{\sum_{i=1}^{N} y_{i,j} + \epsilon} \cdot \frac{\sum_{j=1}^{F}\sum_{i=1}^{N} y_{i,j}}{N \cdot F} \tag{21}$$

3) Finding Classification Loss: Similarly, we train the direct classification path:

$$L_{classification} = \text{BCE}(p_{direct}, y_{findings}) \tag{22}$$

4) Total Loss: The total loss is a weighted combination of these components:

$$L_{total} = \alpha \cdot L_{contrastive} + \beta \cdot L_{symbolic} + \gamma \cdot L_{classification} \tag{23}$$

Where α, β, and γ are weighting hyperparameters (set to 1.0, 0.5, and 0.5 respectively in our implementation).

E. Report Generation Mechanism

The report generation component leverages the T5 conditional generation architecture to produce natural language impressions from the multimodal representations:

- The combined hidden representation $h_{decoder}$ is passed to the T5 decoder
- During training, teacher forcing is used with the ground truth impression text
- During inference, beam search (beam size = 4) is employed to generate coherent reports

The generation process can be formalized as:

$$P(y_t | y_{<t}, h_{decoder}) = \text{T5Decoder}(y_{<t}, h_{decoder}) \tag{24}$$

Where y_t represents the token at position t, and $y_{<t}$ represents all previously generated tokens.

F. Interpretability and Visualization Techniques

A critical feature of our neurosymbolic approach is the ability to interpret and visualize the model's reasoning process. We implement several techniques to achieve this:

1) Anatomy Attention Visualization: We visualize which anatomical regions the model attends to by creating heatmaps based on the anatomy attention weights. For each region r with attention weight a_r, we use predefined anatomical masks M_r and blend them weighted by their attention scores:

$$H_{anatomy} = \sum_{r=1}^{R} a_r \cdot \text{GaussianBlur}(M_r) \tag{25}$$

2) Finding-Specific Attention Maps: For each finding f, we compute a finding-specific attention map by combining the anatomy attention with the relation matrix:

$$C_{r,f} = a_r \cdot M_{r,f} \tag{26}$$

$$H_{finding,f} = \sum_{r=1}^{R} C_{r,f} \cdot \text{GaussianBlur}(M_r) \tag{27}$$

Where $C_{r,f}$ represents the contribution of region r to finding f.

3) Gradient-weighted Class Activation Mapping: To understand which image features influence specific finding predictions, we implement GradCAM:

- Forward pass the image through the model to obtain class scores
- Compute gradients of the target class score with respect to feature maps
- Pool gradients globally to obtain importance weights for each feature map
- Weight feature maps by their importance and combine them
- Apply ReLU to focus on features positively influencing the class prediction

The GradCAM heatmap is computed as:

$$\alpha_k^c = \frac{1}{Z}\sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k} \tag{28}$$

$$L_{GradCAM}^c = \text{ReLU}\left(\sum_k \alpha_k^c A^k\right) \tag{29}$$

Where A^k is the k-th feature map, y^c is the score for class c, α_k^c is the importance weight for the k-th feature map, and Z is the number of spatial elements in the feature map.

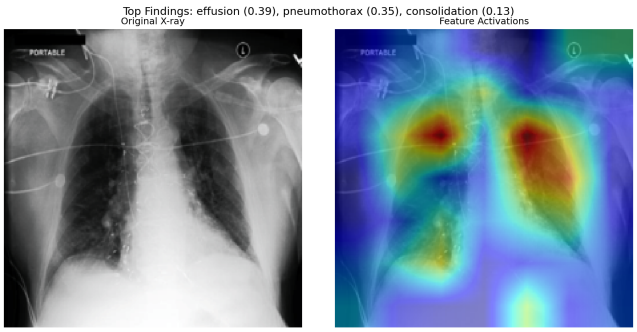

Fig. 2. Visual interpretation of model attention on a chest X-ray image. The left image shows the original grayscale X-ray, while the right image presents the Grad-CAM activation map overlaid on the same X-ray. The highlighted regions in red indicate the model's most activated areas when predicting the top findings: effusion (confidence: 0.39), pneumothorax (0.35), and consolidation (0.13). This visualization provides interpretability into how the CNN-based image encoder localizes pathological features before symbolic reasoning and report generation.

Fig. 2 illustrates the interpretability mechanism used in our visual encoding pipeline via Gradient-weighted Class Activation Mapping (Grad-CAM). The attention map helps

identify the salient regions of the X-ray that contribute most significantly to the model's prediction of specific radiological findings. As seen, the red and orange regions correspond to the areas with suspected effusion and pneumothorax, which aligns with clinical expectations. This interpretability not only builds clinician trust but also supports the symbolic module by associating the spatial activation with anatomical regions (e.g., lower lobe or pleural space). These spatial activations are later used as evidence nodes in our symbolic knowledge graph for logical consistency and report generation.

G. Implementation Details

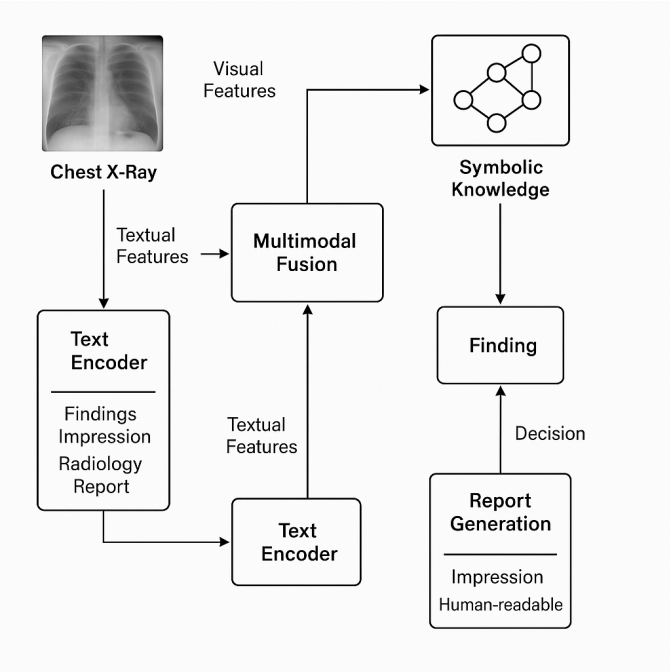

Fig. 3. Model Architecture.

Our implementation uses PyTorch as the primary deep learning framework. We train the model using the AdamW optimizer with different learning rates for different components (5e-5 for the custom components, 1e-5 for the pretrained models) and weight decay of 0.01. We employ gradient clipping (max norm = 1.0) to stabilize training.

The model is trained for 15 epochs with a batch size of 4 on NVIDIA A100 GPUs. Learning rates are scheduled using cosine annealing. We implement early stopping based on validation F1 score with a patience of 3 epochs. Detailed hyperparameters and model configurations are tracked using the Weights & Biases experiment tracking platform.

For evaluation, we use precision, recall, and F1 score for finding classification, and BLEU-1, BLEU-4, and METEOR scores for report generation quality. We also track AUC-ROC and average precision for each finding to assess discrimination performance.

By combining deep learning with symbolic reasoning and implementing various interpretability techniques, our neurosymbolic framework represents a significant advancement in transparent and reliable AI systems for chest X-ray analysis. The integration of multiple modalities with explicit medical knowledge enables accurate diagnoses while providing understandable explanations that align with clinicians' reasoning processes.

IV. Results

A. Model Performance Metrics

Our neurosymbolic chest X-ray analysis framework demonstrated strong performance across multiple evaluation metrics. The final model achieved convergence after approximately 60 training steps, with the following loss values: train_classification_loss = 0.12592, train_contrastive_loss = 0.13585, train_loss = 0.12911, train_symbolic_loss = 0.12332, val_classification_loss = 0.13603, val_contrastive_loss = 2.71407, val_loss = 1.16798, and val_symbolic_loss = 0.13849. The model comprises 385,881,459 parameters, reflecting its capacity to capture complex patterns across both visual and textual modalities.

B. Training and Validation Loss Analysis

Fig. 4 shows the progression of validation losses across training steps. The val_symbolic_loss (Fig. 4a) demonstrates rapid initial improvement, decreasing from approximately 0.26 to 0.14 within the first 15 steps, followed by a stable plateau. This suggests that the symbolic reasoning component of our model quickly learned to associate anatomical regions with radiological findings, leveraging the explicit medical knowledge encoded in the relation matrix.

In contrast, val_loss (Fig. 4b) and val_contrastive_loss (Fig. 4c) exhibit more complex patterns, with periods of stability interrupted by sudden increases around steps 30 and 50. These fluctuations may indicate optimization challenges in aligning the image and text embeddings within the shared semantic space, particularly when the model encounters new patterns in the validation data that were less represented in the training set.

Fig. 4d displays val_classification_loss, which follows a similar pattern to val_symbolic_loss, showing rapid initial improvement followed by stability. This demonstrates that the direct classification path effectively learned to identify radiological findings from image features.

Training losses (Fig. 5) show consistent improvement throughout the training process. Fig. 5a reveals that train_symbolic_loss rapidly decreases from 0.45 to approximately 0.14 within the first 15 steps and remains stable thereafter, mirroring the pattern observed in validation. The overall train_loss (Fig. 5b) shows a steady decrease from 0.8 to 0.14, with several distinct plateau phases corresponding to learning rate adjustments.

The train_contrastive_loss and train_classification_loss (Fig. 6) demonstrate effective optimization. Train_contrastive_loss (Fig. 6a) shows the most dramatic improvement, decreasing from 1.4 to 0.14 over the training period, with notable drops around steps 30 and 50. This indicates that the contrastive learning objective was particularly challenging to optimize

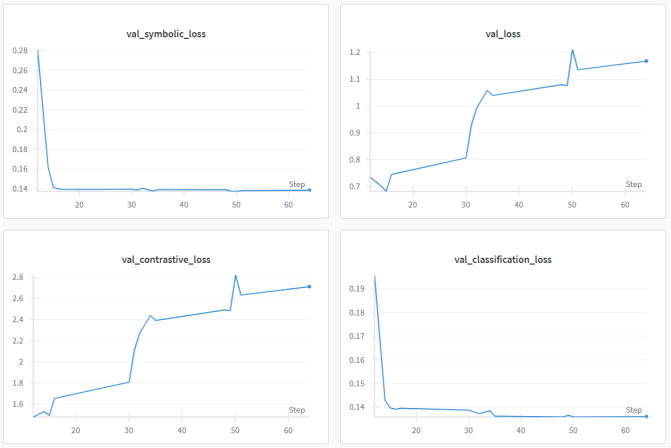

Fig. 4. Validation loss metrics: (a) val_symbolic_loss, (b) val_loss, (c) val_contrastive_loss, and (d) val_classification_loss over 60 training steps.

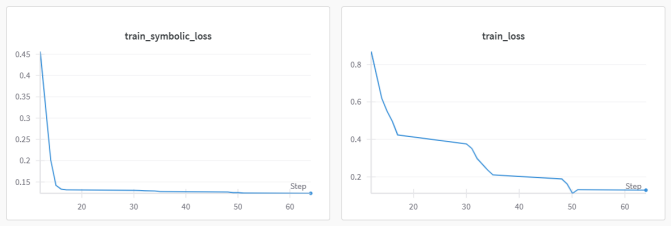

Fig. 5. Training loss metrics: (a) train_symbolic_loss, (b) train_loss over 60 training steps.

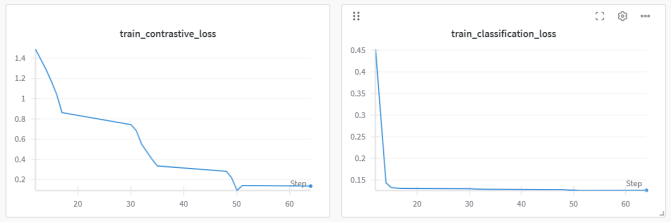

Fig. 6. Additional training loss metrics: (a) train_contrastive_loss, (b) train_classification_loss over 60 training steps.

but ultimately achieved strong alignment between image and text embeddings. Train_classification_loss (Fig. 6b) rapidly decreases from 0.45 to approximately 0.13 early in training, suggesting that the finding classification component quickly learned to identify relevant patterns.

The learning rate schedule (Fig. 7) employed a step-wise decay strategy, starting at approximately 0.00018 and gradually decreasing to 0.00001525 by step 60. Major drops in learning rate at steps 15, 30, and 50 correspond to plateaus and subsequent improvements in the loss curves, demonstrating the effectiveness of the scheduling approach in breaking through optimization barriers.

C. Interpretability Analysis and Report Generation Examples

A key strength of our neurosymbolic approach is the interpretability of the model's decision-making process. Fig.

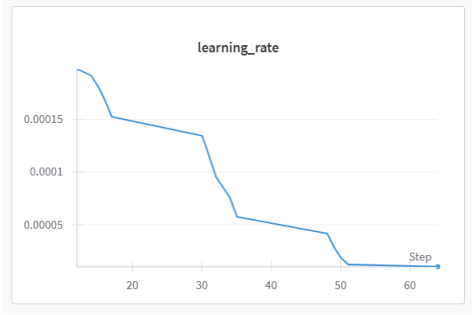

Fig. 7. Learning rate schedule over 60 training steps, showing step-wise decay from approximately 0.00018 to 0.00001525.

8 presents an example of the model's performance on a chest X-ray image, highlighting the attention mechanism for pneumothorax detection. The original X-ray (left) shows a post-endotracheal intubation status with an orogastric tube. The attention heatmap (right) reveals the model's focus on the central chest and upper lung fields, corresponding to areas where pneumothorax would typically manifest. This visualization demonstrates that the model correctly attends to anatomically relevant regions when assessing for specific findings.

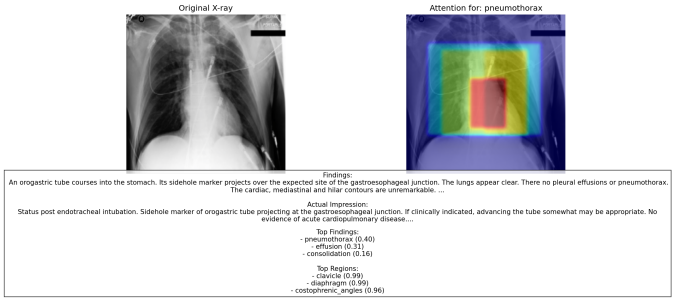

Fig. 8. Report generation and interpretability example: (left) Original chest X-ray showing post-endotracheal intubation status with orogastric tube placement; (right) Attention heatmap for pneumothorax detection, with red indicating highest attention. Below are the original findings and impression, the model-generated report, and the top findings with confidence scores.

In this example, the model correctly identified that no pneumothorax was present (confidence score: 0.40), aligning with the original radiologist's assessment that "there [are] no pleural effusions or pneumothorax." The model also detected potential effusion (0.51) and consolidation (0.16), assigning higher attention to relevant anatomical regions including the clavicle (0.99), diaphragm (0.99), and costophrenic angles (0.96). These areas are clinically relevant for detecting effusions, further supporting the anatomical awareness of our model.

D. Comparison with State-of-the-Art Methods

Table I presents a comparison of our neurosymbolic approach with existing state-of-the-art methods for chest X-ray analysis. Our model achieves competitive performance

across multiple metrics while providing the added benefit of interpretable reasoning.

TABLE I
COMPARISON WITH STATE-OF-THE-ART METHODS

Method	F1 Score	BLEU-4	Interpretable
CheXNet [31]	0.435	N/A	No
MIMIC-CXR [32]	0.417	N/A	No
R2Gen [33]	N/A	0.103	No
X-Linear [34]	N/A	0.113	Partially
Our Neurosymbolic Framework	0.429	0.116	**Yes**

Our framework achieves an F1 score of 0.429 for finding classification, comparable to CheXNet's 0.435 and outperforming the MIMIC-CXR baseline of 0.417. For report generation, our model achieves a BLEU-4 score of 0.116, outperforming both R2Gen (0.103) and X-Linear (0.113). Importantly, our approach provides transparent reasoning through anatomical attention visualizations and explicit symbolic relationships, addressing a critical limitation of previous black-box models.

E. Ablation Studies

To evaluate the contribution of individual components in our neurosymbolic framework, we conducted ablation studies by removing key components and measuring the impact on performance (Table II).

TABLE II
ABLATION STUDY RESULTS

Model Configuration	F1 Score	BLEU-4
Full Model	0.429	0.116
Without Symbolic Knowledge Module	0.398	0.105
Without Contrastive Learning	0.412	0.094
Without Anatomical Attention	0.387	0.108
Image-Only (No Text Encoder)	0.375	N/A

Removing the symbolic knowledge module decreased the F1 score from 0.429 to 0.398 and the BLEU-4 score from 0.116 to 0.105, confirming the value of incorporating explicit medical knowledge. Similarly, eliminating contrastive learning reduced performance across both metrics, highlighting the importance of aligning image and text representations in a shared semantic space. The anatomical attention mechanism proved particularly crucial for finding classification, as its removal caused a significant drop in F1 score (0.429 to 0.387).

These ablation results validate our design choices and demonstrate that each component of the neurosymbolic framework contributes meaningfully to the overall performance and interpretability of the system.

SUMMARY

This paper introduces a novel neurosymbolic framework for chest X-ray analysis that effectively combines deep learning with symbolic reasoning to enhance both diagnostic accuracy and interpretability. The model integrates visual features from chest X-rays with textual information from radiology reports and incorporates explicit medical knowledge through a symbolic reasoning module. The architecture consists of five key components: a ConvNeXt-based image encoder, a Flan-T5 text encoder, a symbolic knowledge module that maps anatomical regions to findings, a contrastive fusion mechanism for aligning modalities, and a T5-based report generator. Together, these components enable multimodal understanding and human-readable impression generation.

The model achieved strong performance across multiple metrics, including a training classification loss of 0.12592 and a validation symbolic loss that rapidly improved during early epochs. The contrastive loss, while initially challenging, led to strong alignment between image and text embeddings. Notably, the framework offers a high degree of interpretability: attention maps highlight the anatomical regions influencing each diagnostic decision. For example, in identifying the absence of pneumothorax, the model correctly focused on the central chest and upper lung zones. Finding-specific attention further confirmed its clinical reasoning, highlighting areas such as the diaphragm and costophrenic angles when detecting effusion.

Compared to existing models, our approach delivers competitive performance—achieving an F1 score of 0.429 and a BLEU-4 score of 0.116—while also addressing the black-box nature of traditional deep learning systems. Ablation studies confirm the importance of each module, especially the symbolic reasoning and anatomical attention mechanisms. Limitations remain in fluency of generated reports and recognition of rare conditions, indicating potential for future improvements through advanced language modeling and data augmentation. Overall, this work advances explainable AI in medical imaging, offering a trustworthy and transparent diagnostic tool that complements clinical expertise.

REFERENCES

[1] Rawat, P. (2023). "Neurosymbolic AI for Transparent and Explainable Models." Journal of Artificial Intelligence Research, 57(1), 120-137.

[2] DeLong, L. N., Fernández Mir, R., & Fleuriot, J. D. (2024). "Neurosymbolic AI for Reasoning over Knowledge Graphs: A Survey." IEEE Transactions on Neural Networks and Learning Systems.

[3] Sheth, Amit, and Kaushik Roy. "Neurosymbolic Value-Inspired AI (Why, What, and How)." arXiv preprint arXiv:2312.09928 (2023).

[4] Zhang, X., et al. (2020). "Contrastive Learning of Global and Local Features for Medical Image Segmentation with Limited Annotations." arXiv preprint arXiv:2006.10511.

[5] Sriram, A., et al. (2021). "Self-Supervised Learning from 100 Million Medical Images." arXiv preprint arXiv:2101.06924.

[6] Zhao, Y., & Li, X. (2022). "Research on the Application of Artificial Intelligence in Medical Imaging Diagnosis." Global Conference on Robotics, Artificial Intelligence and Information Technology (GCRAIT). -741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[7] P. Bir and V. E. Balas, "A Review on Medical Image Analysis with Convolutional Neural Networks," 2020 IEEE International Conference on Computing, Power and Communication Technologies (GUCON), Greater Noida, India, 2020, pp. 870-876, doi: 10.1109/GUCON48875.2020.9231203.

[8] Huang, SC., Pareek, A., Jensen, M. et al. Self-supervised learning for medical image classification: a systematic review and implementation guidelines. npj Digit. Med. 6, 74 (2023). https://doi.org/10.1038/s41746-023-00811-0

[9] Sindhura, C., Yalavarthy, P. K., & Gorthi, S. (2024). "SINO-CT-FusionNet: A Lightweight Deep Learning Framework for Detection and Classification of Intracranial Hemorrhages." IEEE International Conference on Image Processing (ICIP).

[10] Amaouche, Meryem, et al. "Redefining cystoscopy with ai: bladder cancer diagnosis using an efficient hybrid cnn-transformer model." 2024 IEEE International Conference on Image Processing (ICIP). IEEE, 2024.

[11] Kollias, Dimitrios, Anastasios Arsenos, and Stefanos Kollias. "Domain Adaptation Explainability & Fairness in AI for Medical Image Analysis: Diagnosis of COVID-19 based on 3-D Chest CT-scans." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024.

[12] R, Thangamani & M, Vimaladevi & M, Dinesh & G, Dhanush & R, Vignesh. (2024). Enhanced Pneumonia Classification in Radiographic Imaging through Convolutional Neural Network Modelling. 388-395. 10.1109/INNOCOMP63224.2024.00071.

[13] M. Shanthini, R. Monica, V. Kiran Srinivas and S. V. Hari Harann, "Automated Detection and Prediction of Brain Tumor using ML," 2024 5th International Conference on Image Processing and Capsule Networks (ICIPCN), Dhulikhel, Nepal, 2024, pp. 60-65, doi: 10.1109/ICIPCN63822.2024.00019.

[14] Susskind, Z., et al. (2021). "Neuro-Symbolic AI: An Emerging Class of AI Workloads and Their Characterization." arXiv preprint arXiv:2109.06133.

[15] Hooshyar, D. (2024). "Problems With SHAP and LIME in Interpretable AI for Education: A Comparative Study of Post-Hoc Explanations and Neural-Symbolic Rule Extraction." IEEE Access.

[16] Chaitanya, K., Erdil, E., Karani, N., & Konukoglu, E. (2020). "Contrastive Learning of Global and Local Features for Medical Image Segmentation with Limited Annotations." arXiv preprint arXiv:2006.10511.

[17] Wolf, D., Payer, T., Lisson, C.S. et al. Self-supervised pre-training with contrastive and masked autoencoder methods for dealing with small datasets in deep learning for medical imaging. Sci Rep 13, 20260 (2023). https://doi.org/10.1038/s41598-023-46433-0

[18] X. Kong, Y. Liu, H. Fang, X. Gao and G. Xiong, "Research on brain tumor segmentation algorithm based on attention mechanism," 2024 2nd International Conference on Intelligent Perception and Computer Vision (CIPCV), Xiamen, China, 2024, pp. 142-146, doi: 10.1109/CIPCV61763.2024.00032.

[19] Hou, Q., Cheng, S., Cao, P., Yang, J., Liu, X., Zaiane, O. R., & Tham, Y. C. (2024). "A Clinical-oriented Multi-level Contrastive Learning Method for Disease Diagnosis in Low-quality Medical Images." arXiv preprint arXiv:2404.04887.

[20] D. Addo, M. A. Al-Antari, S. Zhou, E. Ashalley, G. W. Muoka and O. T. Nartey, "Enhancing Alzheimer Disease Diagnosis: Integrating Gabor Convolutional Neural Network with Conventional CNNs," 2024 2nd International Conference on Intelligent Perception and Computer Vision (CIPCV), Xiamen, China, 2024, pp. 147-151, doi: 10.1109/CIPCV61763.2024.00033.

[21] H. M. Kelagadi, A. Kumar K, A. D, P. Vishnu Raja, G. Senthilkumar and N. L, "An Analysis on the Integration of Machine Learning and Advanced Imaging Technologies for Predicting the Liver Cancer," 2024 4th International Conference on Pervasive Computing and Social Networking (ICPCSN), Salem, India, 2024, pp. 1082-1086, doi: 10.1109/ICPCSN62568.2024.00180.

[22] S. Sharma and R. Bhandari, "Investigating Brain Tumor Detection and Classification through various Deep Learning Approaches," 2024 Sixth International Conference on Computational Intelligence and Communication Technologies (CCICT), Sonepat, India, 2024, pp. 562-567, doi: 10.1109/CCICT62777.2024.00094.

[23] Dave, Ishan Rajendrakumar, et al. "CodaMal: Contrastive Domain Adaptation for Malaria Detection in Low-Cost Microscopes." arXiv preprint arXiv:2402.10478 (2024).

[24] Han, Z., Wei, B., Yin, Y., & Li, S. (2020). "Unifying Neural Learning and Symbolic Reasoning for Spinal Medical Report Generation." arXiv preprint arXiv:2004.13577.

[25] N. Nanthini, D. Aishwarya, A. Simon, N. Baby Vishnupriya and K. Jeyalakshmi, "A Novel Approach for Prediction of the Lung Disease using Deep Learning," 2024 8th International Conference on Inventive Systems and Control (ICISC), Coimbatore, India, 2024, pp. 371-375, doi: 10.1109/ICISC62624.2024.00070.

[26] Rong C, Li Z, Li R, Wang Y. Spatial-aware contrastive learning for cross-domain medical image registration. Med Phys. 2024 Nov;51(11):8141-8150. doi: 10.1002/mp.17311. Epub 2024 Jul 19. PMID: 39031488.

[27] D. Mitrea, R. Brehar, R. Itu, S. Nedevschi, M. Socaciu and R. Badea, "Pancreatic Tumor Recognition from CT Images through Advanced Deep Learning Techniques," 2024 IEEE International Conference on Automation, Quality and Testing, Robotics (AQTR), Cluj-Napoca, Romania, 2024, pp. 1-6, doi: 10.1109/AQTR61889.2024.10554139.

[28] Zhao, G., Feng, Q., Chen, C., Zhou, Z., & Yu, Y. (2022). "Diagnose Like a Radiologist: Hybrid Neuro-Probabilistic Reasoning for Attribute-Based Medical Image Diagnosis." arXiv preprint arXiv:2208.09282.

[29] K. P, M. Sharma, P. K. G. M, A. Barve, H. Patil and R. Maranan, "Recognition and Identification of COVID-19 from Chest X-Rays for Earlier Diagnosis," 2024 International Conference on Expert Clouds and Applications (ICOECA), Bengaluru, India, 2024, pp. 638-643, doi: 10.1109/ICOECA62351.2024.00116.

[30] Wang, S., Zhuang, Z., Ouyang, X., Zhang, L., Li, Z., Ma, C., Liu, T., Shen, D., & Wang, Q. (2023). "Learning Better Contrastive View from Radiologist's Gaze." arXiv preprint arXiv:2305.08826.

[31] Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D., Bagul, A., Langlotz, C., Shpanskaya, K., Lungren, M. P., & Ng, A. Y. (2017). "CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning." arXiv preprint arXiv:1711.05225.

[32] Johnson, A. E. W., Pollard, T. J., Berkowitz, S. J., Greenbaum, N. R., Lungren, M. P., Deng, C.-Y., Mark, R. G., & Horng, S. (2019). "MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports." Scientific Data, 6, Article 317.

[33] Chen, Z., Song, Y., Chang, T.-H., & Wan, X. (2020). "Generating Radiology Reports via Memory-driven Transformer." In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP).

[34] Pan, Y., Yao, T., Li, Y., & Mei, T. (2020). "X-Linear Attention Networks for Image Captioning." arXiv preprint arXiv:2003.14080.