Seeing Through Expert's Eyes: Leveraging Radiologist Eye Gaze and Speech Report with Graph Neural Networks for Chest X-ray Image Classification

Jamalia Sultana, Ruwen Qin, and Zhaozheng Yin

Stony Brook University, Stony Brook, NY 11794, USA
{jamalia.sultana,ruwen.qin,zhaozheng.yin}@stonybrook.edu

Abstract. Recently, integrating eye-tracking techniques and texts into image-based disease classification has gained traction. To address the unmet needs such as heterogeneous data alignment, information propagation and aggregation, and expert knowledge embedding, we propose an innovative expert-guided Graph Neural Network (GNN) that uses radiologists' eye-gaze data and transcribed audio reports with X-ray images during training. By distilling expert knowledge from gaze data and diagnosis reports, our GNN can achieve high accuracy using only X-ray images during inference. This approach provides a robust framework for disease diagnosis, embedded with the radiologists' insights, addressing challenges in aligning heterogeneous data, propagating local information for global decisions, and leveraging expert knowledge effectively. Additionally, the attention maps on X-ray images which are generated from the GNN model visualize the Region of Interest (ROI) for the diagnosed disease. Evaluated on two benchmark chest X-ray datasets, the proposed method outperforms state-of-the-art X-ray image classification methods.

Keywords: Multi-modal · Eye-gaze · Graph Neural Network · Text reports · Radiology

1 Introduction

Disease categorization from medical images has always been a tough task in computer vision [11]. The advancement in deep learning techniques has somewhat tackled these issues [9]. However, medical image classification remains tricky compared to natural image classification due to intricate anatomical structures overlapping in planar (2D) views [1] and limited soft tissue contrast [14].

Recent studies have tackled these challenges by combining eye-tracking techniques [26] and textual information [17] to augment the model with prior knowledge of abnormality locations. Eye-tracking methods capture radiologists' eye movement data during screenings, providing supplementary location details that spotlight potentially problematic regions [34]. By incorporating eye movement data, deep learning models can gain a more coherent understanding of disease

characteristics, thus improving diagnostic accuracy. Integrating text data, such as clinical notes and diagnostic reports, with imaging data can significantly improve model performance. For example, hybrid deep spatial and statistical feature fusion methods enhance MRI classification by incorporating textual descriptions of anatomical structures, improving contextual understanding [19]. Advances in multimodal AI, such as GPT-4 Vision for otolaryngology, integrate patient-specific textual data to refine diagnostic accuracy [27]. Vision-language pre-training models (VLPM) have advanced the integration of medical texts with imaging data, creating a unified representation space that improves understanding of text-image relationships and enhances zero-shot learning capabilities [43]. These studies highlight the synergy between textual information and X-ray images for accurate medical image classification and advocate for a multimodal approach.

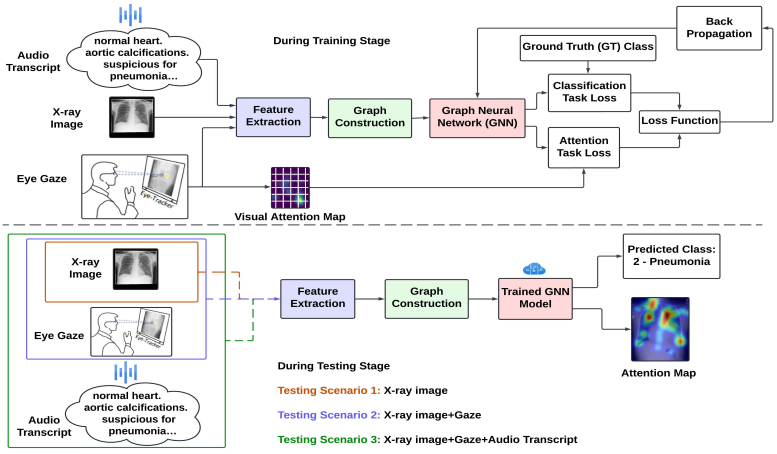

Fig. 1: An overview of our proposed method using the eye gaze and texts along with X-rays. The top side shows the training stage and the bottom side shows the workflow in the inference stage where the model has three testing scenarios. In the training stage, the tri-modality (X-ray image, eye gaze and text) is unified to create the graphs for the GNN model's input.

Despite the previous success, there are still several unsolved challenges: (1) Most of previous works focus on image+gaze or image+text pairs. When more data types are involved (image, eye gaze, text, etc.), it is more difficult to align heterogeneous data (e.g., how to align a phrase in the text to an image region and an eye gaze point?); (2) Image pixels or patches, phrases in texts, and eye gaze points represent local information. How to propagate the local information to its neighborhood and how to aggregate the local information to make a global diagnosis decision? and (3) Eye gazes and text reports are leveraged for X-ray image classification during the training stage. How to distill the expert knowledge

for the inference stage when only X-ray image is available? Facing these challenges, we introduce RET-GNN (Radiologist Eye-gaze and Text-reports Graph Neural Network), a Graph Neural Network (GNN)- based model that (1) aligns heterogeneous data types at graph node level, (2) propagates local features to neighboring nodes and assembles the features from all nodes for disease classification, and (3) learns expert knowledge via their eye gaze and text report during training and performs the inference only using X-ray images, as illustrated in Fig. 1. During training, we choose the image as the central anchor for binding eye gazes and texts, allowing for a more integrated and comprehensive analysis of medical data. A graph is designed to propagate information between different parts of the image, which facilitates the learning of relationships between various organs and enables the aggregation of information from the entire image for the disease classification. During testing, the input is only chest X-ray image without the eye gaze or speech transcript from the experts [1], and the X-ray image is classified by the trained GNN with embedded knowledge of eye gaze and diagnosis report. The GNN also generates an attention map indicating the suspect regions. The major contributions of our work are summarized as follows:

1. We design a graph that aligns data from multiple domains (image, eye gaze, and text), using X-ray images as the central anchor. This approach allows for a cohesive representation that combines various types of medical data, enhancing the overall understanding and analysis of medical images.
2. We propagate eye gaze points and their time durations into Visual Attention Maps (VAMs), which improves the graph model's ability to capture and utilize critical visual cues provided by radiologists. We propagate textual information from individual graph nodes to their neighborhoods, so text clues provided by radiologists are aware by more graph nodes. The graph enables the model to understand and leverage the relationships between different regions of the image, facilitating more accurate and robust disease classification.
3. The proposed graph neural network model learns complex relationships between different image regions via aligned heterogeneous data types (image, gaze, and text) during training, and the model embedded with expert knowledge can perform inference only using image. Evaluated on two widely used public datasets (MIMIC-CXR [21] and REFLACX [24]), the proposed RET-GNN outperforms previous state-of-the-art methods in different testing scenarios.

2 Related Work

2.1 Chest X-Ray Image Classification

Chest X-ray classification has witnessed significant advances in recent years due to the availability of large-scale public chest X-ray datasets and the development

[1] Previous works like [16,36] have used eye gaze and texts as the input during testing. We also include them as testing scenario 2 and 3 for fair comparison with previous works.

of advanced machine learning techniques. Datasets such as MIMIC-CXR [20], REFLACX [24], and others [8, 10, 29, 31] have significantly contributed to model training and evaluation. These datasets collectively provide nearly a million chest X-ray images with class annotations, enabling the development of robust chest X-ray classification algorithms. On the other side, advancements in deep learning algorithms have further enhanced the accuracy and performance of chest X-ray classification. Most methods primarily rely on unimodal chest X-ray images and propose various network architectures for analysis [4, 13, 15, 28].

2.2 Integration of Eye-Gaze Data in Medical Image Analysis

Recent studies have started to explore the integration of eye-gaze data into chest X-ray classification tasks [3, 22, 33]. Research has shown that the inclusion of human expert knowledge via eye-gaze patterns can significantly enhance the accuracy of deep learning models. Prior mainstream works typically transform eye-gaze data into visual attention maps (VAMs), which highlight radiologists' attention regions on the corresponding medical images. These works can be categorized based on their utilization of VAMs. Some studies apply VAMs to process the images and take the processed images as the model input [18, 23, 41]. Others employ a CNN-LSTM hybrid two-stream neural network, where the CNN processes the medical images, and the LSTM encodes the VAMs [21]. The second category focuses on minimizing the difference between VAMs and class activation maps (CAMs) [5,37] or the difference between VAMs and the attention maps generated by a U-Net decoder [21, 40]. More recently, after the release of the Segment Anything Model (SAM) by Meta, a human-computer interaction system called GazeSAM [35] was proposed. This system combines eye-tracking technology with SAM, enabling users to segment the object they are looking at in real-time, demonstrating the potential for real-time eye-gaze integration into routine clinical practice. Existing methods of utilizing eye-gaze data for medical image classification primarily focus on creating image and gaze pairs, limiting their ability to integrate multiple data types. Additionally, using only VAMs in deep learning models to extract gaze information may not effectively aggregate local information for making a global diagnosis decision.

2.3 Integration of Text in Medical Image Classification

Vision-language pre-training models have facilitated the convergence of medical texts with imaging data into a unified representation space, enhancing understanding and classification capabilities [43]. Notable works such as GloRIA [16] and BioVIL [6] have demonstrated the potential of aligning chest X-ray images with radiological reports through self-supervised contrastive learning, effectively capturing both local and global visual features. Additionally, models like MedCLIP [39] and CXR-CLIP [42] have advanced training methodologies by refining image-text specific loss functions, further improving classification performance. Despite advancements in vision-language pre-training models that align chest X-ray images with radiological reports, challenges remain in aligning heterogeneous

data (image, gaze, text), propagating local information for global diagnosis, and distilling expert knowledge for inference when only X-ray images are available.

3 Methodology

In this section, we describe the framework of the proposed RET-GNN (Radiologist Eye-gaze and Text-report Graph Neural Network) for disease classification. RET-GNN extracts features from image patches, eye-gaze data, and text data from radiologists' audio reports (Fig. 2). A graph is built whose node combines features from patch, gaze, position, and text embeddings (Fig. 3). Then, a GNN updates and aggregates node information to predict the disease class and generate the attention map indicating the disease regions (Fig. 4).

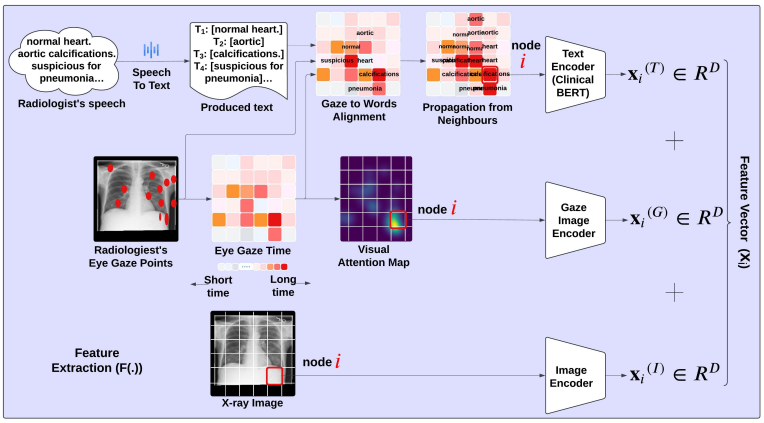

Fig. 2: An overview of feature extraction module to create patch-based node and its feature using the X-ray image, visual attention map, and speech-to-text data from the radiologist.

3.1 Graph Construction

Our proposed method integrates chest X-ray images, eye gaze data, and text information. The chest X-ray image serves as the central anchor, providing a structured grid based graph whose nodes are image patches. Eye-gaze data, represented by scatter points, indicates radiologists' attention locations. The text data, represented by sentences, conveys the radiologists' descriptions of the areas they focused on. We construct graphs by combining X-ray image patches, VAMs from gaze data, and text embeddings to represent each node. The feature vector for patch-based node i, $\mathbf{x}_i$, is defined as follows:

$$\mathbf{x}_i = \mathbf{x}_i^{(T)} + \mathbf{x}_i^{(G)} + \mathbf{x}_i^{(I)}, \tag{1}$$

where $\mathbf{x}_i^{(T)}$ is the text embedding, $\mathbf{x}_i^{(G)}$ is the gaze embedding, and $\mathbf{x}_i^{(I)}$ is the patch embedding of image. These features collectively represent the N nodes $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$.

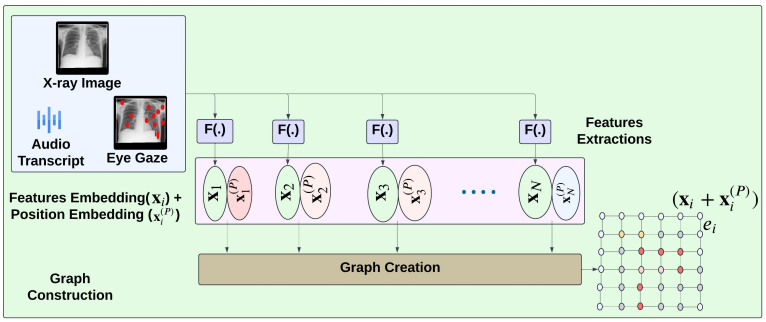

Fig. 3: An overview of our proposed graph construction using the features from image, eye gaze, text, and position embeddings.

Image Patch Embedding The image is divided into N numbers of 15×15 patches and each patch is considered as an individual node. Image patch is used as graph node because it assembles local context and is more computational-efficient than individual pixels. We use the overlapping patch embedding method to enhance feature extraction from image patches. This technique enlarges the patch window so adjacent patches overlap by half, using a convolutional operation with zero padding to preserve spatial relationships and local features. Implemented in the Pyramid Vision Transformer (PVT) v2, this method significantly improves image classification, object detection, and segmentation by providing richer feature representations compared to non-overlapping patch methods [38].

Eye Gaze Embedding Eye-gaze data provides attention locations and the duration of focus, adding temporal information. As shown in Fig. 2, we extract gaze coordinates and time spent at each point, creating a matrix g whose values represent gaze duration. The longer a radiologist looks at a location, the more red-saturated the color: $g[u, v] = t$. Here, t is the time duration of eye gaze at (u, v) location. Patch-level gaze data is generated by summing gaze values within patches, and normalizing gaze intensity. To create VAMs, we aggregate fixation times within each patch:

$$x_i^{(G)} = \sum_{(u_l, v_l) \in patch_i} g[u_l, v_l], \tag{2}$$

where $i \in [1, N]$ (total patches), and $l \in [1, L]$ (number of gazes). This scalar $x_i^{(G)}$ is then replicated to a vector $\mathbf{x}_i^{(G)} \in \mathbb{R}^D$ for feature fusion with image and

text, where D denotes the feature dimension. $\mathbf{x}_i^G$ represents the patch's attention feature that captures detailed attention patterns, allowing the model to focus on important regions.

Text Embedding For text embedding, we first identify the start and end times of gaze fixation from the time-series gaze events, as shown in Fig. 2 (Eye Gaze Time). We filter out the stop words such as, 'the', 'a', 'patient' and so on, concatenate, and align transcript phrases within these gaze times to the corresponding patches via the gaze fixation points. To propagate textual information to nearby patches, we check each patch for phrases, fill empty patches with phrases from their neighbors, and update the non-empty ones with their neighbors' textual information. This process is summarized by the following equation:

$$\text{Phrases}_i = \text{Phrases}_i \cup \bigcup_{j \in \text{NN}(i)} \text{Phrases}_j. \tag{3}$$

Here, $\text{NN}(i)$ represents the neighboring nodes of patch i. When patch i has no phrases ($\text{Phrases}_i = \emptyset$), it inherits the union of phrases from its neighboring patches. If patch i already contains phrases ($\text{Phrases}_i \neq \emptyset$), it is enriched by adding unique phrases from its neighbors. This ensures robust propagation of textual information, covering approximately 70% of initially empty patches. We utilized BioClinicalBERT [2], a BERT variant fine-tuned on medical texts, to capture biomedical semantics for the text encoder.

In short, our method leverages gaze data (fixation points and their start and end times) to initialize the alignment between speech-to-transcript texts and patches, and then propagates the textual information to a local neighborhood, thereby enriching the semantic representation of image patches.

Position Embedding To maintain the original positional information in the GNN after patch formation, we adopt a position embedding method from [12]. This involves adding a learnable absolute positional encoding vector $\mathbf{x}_i^{(P)}$ to the feature vector $\mathbf{x}_i$ in Fig. 2 and calculating the relative positional distance between nodes as $(\mathbf{x}_i^{(P)})^T \mathbf{x}_j^{(P)}$ to determine neighbors in the k-nearest neighbors algorithm for graph construction.

Link Creation To construct the links E of the graph, we use the k-nearest neighbors algorithm to define the connections between nodes [12]:

$$E = \{(\mathbf{x}_i, \mathbf{x}_j) \mid \mathbf{x}_j \in K(\mathbf{x}_i)\}, \tag{4}$$

where $K(x_i)$ represents the k-nearest neighbors of $\mathbf{x}_i$. This method ensures that graph $G = (V, E)$ captures the relational information between different parts of the image, integrating visual, gaze and text data effectively for GNN to process.

3.2 Graph Neural Network

As illustrated in Fig. 4, our GNN model comprises B graph processing blocks, an average pooling layer, a graph classification head, and an attention head. Each graph processing block includes multiple fully connected (FC) layers, a graph convolutional layer [25], and a graph attention network (GAT) layer [7]. Our graph contains N nodes corresponding to N image patches and each node

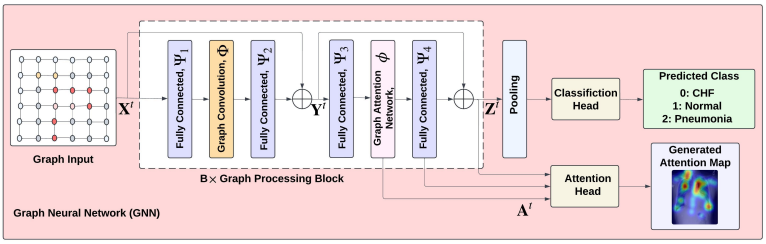

Fig. 4: The architecture of the proposed Graph Neural Network (GNN).

has a D-dimensional feature vector, $\mathbf{x}_i$ along with its position embedding $\mathbf{x}_i^{(P)}$. The input to a specific layer of the graph processing module at iteration t is expressed as: $\mathbf{X}^t = [\mathbf{x}_n^t, \mathbf{x}_n^P]_{n=[1,\dots N]}$. This module transforms the input into the output defined by:

$$\mathbf{Y}^t = \Psi_2(\Phi(\Psi_1(\mathbf{X}^t))) + \mathbf{X}^t, \tag{5}$$

$$\mathbf{Z}^t = \Psi_4(\phi(\Psi_3(\mathbf{Y}^t))) + \mathbf{Y}^t, \tag{6}$$

where Φ signifies the operation of graph convolution, ϕ represents the operation of the graph attention network layer, and Ψ indicates the fully connected (FC) layer. The FC layers incorporate activation and batch normalization. At the end of the process, a fully connected layer with a softmax function acts as the classification head, delivering probability predictions for each class. The attention head after the B graph processing blocks takes the input of $\mathbf{A}^t = (\Psi_3(\mathbf{Y}^t)) + \Psi_4(\phi(\Psi_3(\mathbf{Y}^t))) + \mathbf{Z}^t$ and feeds them to Grad-CAM [32] to generate the VAMs from our GNN model. This design ensures that the GNN effectively assembles relational data across various segments of the graph, facilitating a comprehensive representation for disease classification.

Graph Attention Network Our attention mechanism in the graph processing block is computed by applying a shared attention mechanism to neighboring nodes. Specifically, given a node i and its neighbor j, the attention coefficient e_{ij} is calculated using the following equation:

$$e_{ij} = \text{LeakyReLU}\left(\mathbf{a}^\top [\mathbf{W}\mathbf{x}_i \| \mathbf{W}\mathbf{x}_j]\right), \tag{7}$$

where $\mathbf{a}$ is a learnable weight vector, $\mathbf{W}$ is a weight matrix, $\mathbf{x}_i$ and $\mathbf{x}_j$ are the feature vectors of nodes i and j, respectively, and $\|$ denotes concatenation. The LeakyReLU activation function is applied to introduce non-linearity. The attention coefficients are then normalized using the softmax function. These normalized attention coefficients are used to compute a weighted sum of the neighbors' features.

$$\mathbf{x}_i' = \sigma\left(\sum_{j \in \mathcal{N}(i)} \alpha_{ij} \mathbf{W} \mathbf{x}_j\right), \tag{8}$$

where σ is a non-linear activation function, ReLU and α_{ij} is normalized attention coefficient. The updated feature vector $\mathbf{x}_i'$ of node i incorporates information from its most relevant neighbors.

3.3 Loss Function

To train our GNN model, we utilize two different types of loss functions: one for the classification task and another for attention alignment. Combining these loss functions, the overall loss for training the model is:

$$\mathcal{L} = \mathcal{L}_{\text{classification}} + \lambda \mathcal{L}_{\text{attention}}, \tag{9}$$

where λ is a hyperparameter that balances the contribution of the attention loss to the total loss. This combined loss function ensures that the model not only classifies accurately but also aligns its attention mechanisms with those of human experts, leading to more interpretable and reliable predictions.

Classification Task Loss We employed the cross-entropy loss function for the classification task. This loss function is defined as:

$$\mathcal{L}_{\text{classification}} = -\sum_{m=1}^{\mathcal{M}} y_m \log(\hat{y}_m), \tag{10}$$

where y_m is the true label and $\hat{y}_m$ is the predicted probability, calculated over all training samples, M. This loss function measures the difference between the true labels and the predicted probabilities, guiding the model to improve its classification accuracy.

Attention Loss For the attention alignment, we used the Intersection over Union (IoU) as the attention loss function. This loss measures the overlap between the generated attention maps from the model and the VAMs derived from the radiologist's eye gaze data. The IoU loss is defined as:

$$\mathcal{L}_{\text{attention}} = 1 - \frac{\sum_m (A_m \cap V_m)}{\sum_m (A_m \cup V_m)}, \tag{11}$$

where A_m represents the attention map generated by RET-GNN, and V_m denotes the VAM from the radiologist's eye gaze data. The intersection ($A_m \cap V_m$ measures the common activated areas between the two maps, while the union ($A_m \cup V_m$) measures the total activated areas. Maximizing this overlap ensures that the model's attention aligns closely with human expert attention patterns.

4 Experiments

4.1 Implementation

We train our model using the AdamW optimizer with a learning rate of 0.0001 and a batch size of 32. 20% of the training dataset is allocated for validation, and we preserve the top validation accuracy checkpoint including the optimal hyper parameters. Our system uses an NVIDIA Tesla V100-PCIE-32GB GPU with PyTorch. Each training iteration takes about 3 minutes, spanning 350 epochs. These implementation details apply across different experiments.

4.2 Dataset Preparation

Our study utilizes a public chest X-ray dataset [21], comprising 1083 cases from the MIMIC-CXR dataset [20], eye gaze data, audio transcripts, and classification ground labels. Additionally, we employ the REFLACX dataset [24], which includes eye-tracking data and timestamped report transcriptions for 2,616 chest X-rays collected from five radiologists. This dataset comprises 3,032 sets of synchronized gaze data, dictation reports, and classification labels. Both datasets contain grayscale X-ray images (approx. 3000×3000 pixels) categorized as Normal, CHF, or Pneumonia. Model performance is assessed using metrics such as accuracy and AUC. Data augmentation methods involve random resizing, cropping to 224×224 pixels, horizontal flipping, and rotation up to 5 degrees. For qualitative comparison, static visual attention maps (VAMs) are generated as ground truth from the eye-gaze data, employing the post-processing method outlined in [21].

4.3 Quantitative Comparison

We conducted a comprehensive comparison of our model against several latest state-of-the-art methods to evaluate its performance. The methods compared include the U-Net+Gaze model [21], the DenseNet121-based model [33], and GazeGNN [36]. Each model uses the official training and test datasets from [21], allowing for a fair evaluation by directly including their reported results.

The U-Net+Gaze model [21] combines U-Net model with gaze data for improved medical image segmentation. Similarly, DenseNet121-based model [33] uses a densely connected CNN to enhance classification performance. GazeGNN [36] integrates gaze data within a GNN framework, capturing contextual relationships. We also compared our model with other gaze-guided approaches

Table 1: Classification results on the Chest X-Ray dataset MIMIC-CXR [21]. All models compared are trained using images, gazes and texts. In practice, only image is available during testing. Since some methods report testing results using additional modalities beyond images, we include those scenarios for comparison purposes, though they are not practical for real-world inferences. Different testing scenarios are color-coded and the best results are denoted by bold.

Method	Modality in Testing			Accuracy	AUC
	X-ray Image	Gaze	Text		
U-Net+Gaze [21]	✓	×	×	-	0.857
	✓	✓	×	-	0.870
BioViL [6]	✓	×	✓	82.20%	0.891
DenseNet121 [33]	✓	✓	×	-	0.836
GazeMTL [30]	✓	×	×	76.85%	0.858
	✓	✓	×	78.50%	0.887
GazeGNN [36]	✓	×	×	80.38%	0.893
	✓	✓	×	83.18%	0.923
Ours (RET-GNN)	✓	×	×	**85.23%**	**0.901**
	✓	✓	×	**86.05%**	**0.913**
	✓	×	✓	**84.91%**	**0.926**
	✓	✓	✓	**87.82%**	**0.938**

using the same dataset, notably GazeMTL [30], which uses multi-task learning. We retrained GazeMTL with the same splits as our model for fair comparison, ensuring unbiased performance evaluation. The quantitative results are shown in Tables 1 and 2.

The RET-GNN model leverages radiologists' eye gaze and text data, using only X-ray images during inference for disease classification. X-ray images serve as the central anchor, allowing the seamless use of other modalities. VAMs capture critical visual cues, enhancing pattern recognition, as shown in Fig. 5. Text data describing abnormal regions, combined with gaze data, significantly boosts performance. Since other methods also report testing results using gazes and texts beyond images, we include these testing scenarios for comparisons, though we think it is impractical in real-world inference (we cannot ask radiologists to provide their eye gaze and speech report on an image to diagnose this image). To accommodate different testing scenarios, we set the tensors of the unused modality to zero while the model is trained based on tri-modality of X-ray images, eye gaze, and text. For instance, while the model is trained with all modalities, we set eye-gaze feature vectors to zero during testing to evaluate the model's performance in the testing scenario of using X-ray images and text.

The RET-GNN model demonstrates superior performance across all testing scenarios on the MIMIC-CXR dataset [21]. When using only X-ray images, it achieves the highest accuracy of 85.23% and an AUC of 0.901, outperforming

Table 2: Classification results on REFLACX dataset [24].

Method	Modality in Testing			Accuracy	AUC
	X-ray Image	Gaze	Text		
SimCLR [6]	✓	×	×	78.80%	0.849
GloRIA [16]	✓	×	✓	79.00%	0.886
Ours (RET-GNN)	✓	×	×	**81.69%**	**0.900**
	✓	✓	×	82.00%	0.862
	✓	×	✓	**83.02%**	**0.920**
	✓	✓	✓	**84.01%**	**0.914**

all other methods [30, 36] in this real-world scenario. Combining X-ray images and gaze data, the model achieves an even higher accuracy of 86.05% and an AUC of 0.913, showcasing the added value of gaze data. Incorporating text data with X-ray images results in an accuracy of 84.91% and the highest AUC of 0.926 in this category, indicating significant improvement from text integration. When all three modalities (X-ray images, gaze, and text) are used, RET-GNN achieves the best performance with an accuracy of 87.82% and an AUC of 0.938. These results illustrate the model's robustness and effectiveness in leveraging multi-modal data for chest X-ray classification. Similar results can be seen for the REFLACX dataset [24] across multiple testing scenarios. The higher AUC values in our results indicate better model performance in distinguishing between positive and negative classes.

4.4 Ablation Study

To study the effectiveness of eye gaze and transcript information during the model training, we have conducted an ablation study. First, we remove the gaze and text embeddings from the GNN models (we convert the tensors of the unused modality to zero), using only the patch embedding from the X-ray images and position embedding during training. Then, we remove the text data while retaining gaze and position embeddings to observe the impact of text data. Similarly, we remove gaze data while keeping other modalities to assess the importance of gaze information. Finally, we compare these results with the model using all modalities to evaluate the contribution of each data type. The ablation study focuses on the effect of each modality during training for image classification, with only X-ray images being used as input in the testing stage.

The comparison in Table 3 reveals several key insights. When combining gaze data with X-ray images, the model achieves higher accuracy compared to using X-ray images alone. We theorize that eye gaze helps the model focus on abnormal regions, aiding pattern recognition for disease classification. The model also shows higher accuracy when combining text data with X-ray images, though the AUC value is slightly lower than using X-ray images alone, likely due to

Table 3: Ablation study on the effect of the training for image classification. Only X-ray image is used as input in the testing stage for the ablation study.

Modality in Testing	Modality in Training			Accuracy	AUC
	X-ray Image	Gaze	Text		
X-Ray Image	✓	×	×	77.36%	0.869
	✓	✓	×	80.83%	0.866
	✓	×	✓	84.43%	0.864
	✓	✓	✓	85.23%	0.901

the diversity of spoken reports. Text data alone may result in low-confidence classifications, but when combined with gaze data, it significantly boosts AUC. Additional information from the text about that region significantly helps the model in finding the radiologists' pattern for disease classification with greater confidence, and gaze data makes the model focus on the abnormal region. This highlights the complementary nature of gaze and text modalities, enhancing the model's robustness and accuracy. This comprehensive evaluation underscores the value of incorporating multiple data types to improve the performance and reliability of medical image analysis models. The model gains additional domain knowledge from radiologists' eye gazes and text reports, beyond the X-ray images. The text data, represented by sentences, reflects the radiologists' descriptions of the attention locations in images.

4.5 Qualitative Analysis

Figure 5 provides qualitative examples, showing heat maps that highlight areas of higher focus by radiologists ((b) and (e)) and relevance to the model's predictions ((c) and (f)). Hotter colors (red, yellow) indicate regions used to determine specific conditions. Correct predictions (green box) demonstrate strong correlation between the highlighted areas (5(c)) and known disease markers (5(b)), reflecting consistency between the model's attention maps and diagnostic markers. Incorrect predictions (red box, 5(f)) show where the model's focus may not accurately represent diagnostic markers, suggesting areas for improvement in model training and architecture.

Overall, this qualitative analysis emphasizes the model's strengths in identifying critical regions for correct predictions while also highlighting future work for further refinement to reduce incorrect classifications. We identify some possible reasons for the misclassification of the X-ray images along with incorrectly generated heatmaps. Due to overfitting to training data, our model may have learned the patterns and noise specific to the training dataset instead of generalizing. The heatmaps generated from that might highlight features that are overly specific to the training set but irrelevant to other data, leading to incorrect regions of interest. Additionally, if certain classes are underrepresented or over-represented ("Normal"), the model may not learn to recognize them effectively. This can lead to misclassifications, particularly for underrepresented

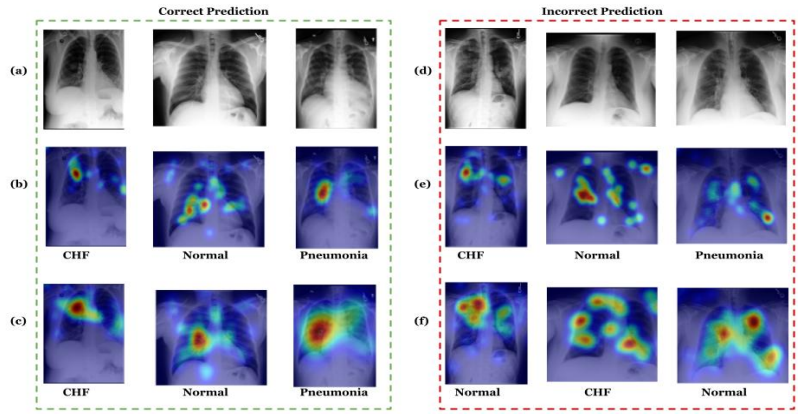

Fig. 5: Visual comparison between the original eye gaze points and the generated attention map from the GNN model, (a) The original chest X-ray image, (b) VAMs generated from the time aggregation of the ground truth eye gaze data, and (c) The generated attention map from our GNN.

classes, as the model's learned features do not adequately distinguish between different conditions. Moreover, the generated heatmaps may be inaccurate if the model cannot fully capture the spatial and contextual relationships in the data, leading to incorrect areas being highlighted. Since we are using the radiologist's eye gaze and text to aid the GNN model, the noise created by the human can also mislead the model for the unseen test datasets.

5 Conclusion

In this paper, we present a novel framework that integrates data from multiple domains (image, eye gaze, and text report) using X-ray images as the central anchor. This approach combines various types of medical data to enhance the overall understanding and analysis of medical images. We transform eye-gaze data into VAMs using gaze points and time durations, capturing critical visual cues from radiologists. Our model also employs a text embedding method, spreading textual information through a neighborhood propagation technique to capture the global context within medical images. Eye gaze focuses the model on abnormal regions, enhancing pattern recognition for disease classification. Combined with text data, which describes abnormal regions, gaze data significantly boosts the performance of our RET-GNN model. This method enhances the model's understanding of relationships between different image regions, resulting in more accurate and robust disease classification, addressing challenges in aligning heterogeneous data and leveraging expert knowledge effectively.

Acknowledgement. The authors have been supported by NSF grants: ECCS-2026357 and ECCS-2025929.

References

1. Albers, J., Wagner, W.L., Fiedler, M.O., Rothermel, A., Wünnemann, F., Di Lillo, F., Dreossi, D., Sodini, N., Baratella, E., Confalonieri, M., et al.: High resolution propagation-based lung imaging at clinically relevant x-ray dose levels. Scientific Reports **13**(1), 4788 (2023)
2. Alsentzer, E., Murphy, J.R., Boag, W., Weng, W.H., Jin, D., Naumann, T., McDermott, M.: Publicly available clinical bert embeddings. arXiv preprint arXiv:1904.03323 (2019)
3. Ashraf, H., Sodergren, M.H., Merali, N., Mylonas, G., Singh, H., Darzi, A.: Eye-tracking technology in medical education: A systematic review. Medical teacher **40**(1), 62–69 (2018)
4. Baltruschat, I.M., Nickisch, H., Grass, M., Knopp, T., Saalbach, A.: Comparison of deep learning approaches for multi-label chest x-ray classification. Scientific reports **9**(1), 6381 (2019)
5. Bhattacharya, M., Jain, S., Prasanna, P.: Radiotransformer: a cascaded global-focal transformer for visual attention–guided disease classification. In: European Conference on Computer Vision. pp. 679–698. Springer (2022)
6. Boecking, B., Usuyama, N., Bannur, S., Castro, D.C., Schwaighofer, A., Hyland, S., Wetscherek, M., Naumann, T., Nori, A., Alvarez-Valle, J., et al.: Making the most of text semantics to improve biomedical vision–language processing. In: European conference on computer vision. pp. 1–21. Springer (2022)
7. Brody, S., Alon, U., Yahav, E.: How attentive are graph attention networks? arXiv preprint arXiv:2105.14491 (2021)
8. Bustos, A., Pertusa, A., Salinas, J.M., De La Iglesia-Vaya, M.: Padchest: A large chest x-ray image dataset with multi-label annotated reports. Medical image analysis **66**, 101797 (2020)
9. Celard, P., Iglesias, E.L., Sorribes-Fdez, J.M., Romero, R., Vieira, A.S., Borrajo, L.: A survey on deep learning applied to medical images: from simple artificial neural networks to generative models. Neural Computing and Applications **35**(3), 2291–2323 (2023)
10. Demner-Fushman, D., Kohli, M.D., Rosenman, M.B., Shooshan, S.E., Rodriguez, L., Antani, S., Thoma, G.R., McDonald, C.J.: Preparing a collection of radiology examinations for distribution and retrieval. Journal of the American Medical Informatics Association **23**(2), 304–310 (2016)
11. Dhar, T., Dey, N., Borra, S., Sherratt, R.S.: Challenges of deep learning in medical image analysis—improving explainability and trust. IEEE Transactions on Technology and Society **4**(1), 68–75 (2023)
12. Han, K., Wang, Y., Guo, J., Tang, Y., Wu, E.: Vision gnn: An image is worth graph of nodes. Advances in neural information processing systems **35**, 8291–8303 (2022)
13. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
14. Hsieh, J.: Spatial and temporal motion characterization for x-ray ct. Medical Physics (2024)
15. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4700–4708 (2017)

16. Huang, S.C., Shen, L., Lungren, M.P., Yeung, S.: Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3942–3951 (2021)
17. Huang, Z., Bianchi, F., Yuksekgonul, M., Montine, T.J., Zou, J.: A visual–language foundation model for pathology image analysis using medical twitter. Nature medicine **29**(9), 2307–2316 (2023)
18. Huff, D.T., Weisman, A.J., Jeraj, R.: Interpretation and visualization techniques for deep learning models in medical imaging. Physics in Medicine & Biology **66**(4), 04TR01 (2021)
19. Iqbal, S., Qureshi, A.N., Alhussein, M., Aurangzeb, K., Choudhry, I.A., Anwar, M.S.: Hybrid deep spatial and statistical feature fusion for accurate mri brain tumor classification. Frontiers in Computational Neuroscience **18**, 1423051 (2024)
20. Johnson, A.E., Pollard, T.J., Berkowitz, S.J., Greenbaum, N.R., Lungren, M.P., Deng, C.y., Mark, R.G., Horng, S.: Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. Scientific data **6**(1), 317 (2019)
21. Karargyris, A., Kashyap, S., Lourentzou, I., Wu, J.T., Sharma, A., Tong, M., Abedin, S., Beymer, D., Mukherjee, V., Krupinski, E.A., et al.: Creation and validation of a chest x-ray dataset with eye-tracking and report dictation for ai development. Scientific data **8**(1), 92 (2021)
22. Kaushal, S., Sun, Y., Zukerman, R., Chen, R.W., Thakoor, K.A.: Detecting eye disease using vision transformers informed by ophthalmology resident gaze data. In: 2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). pp. 1–4. IEEE (2023)
23. Kundur, N.C., Anil, B.C., Dhulavvagol, P.M., Ganiger, R., Ramadoss, B.: Pneumonia detection in chest x-rays using transfer learning and tpus. Engineering, Technology & Applied Science Research **13**(5), 11878–11883 (2023)
24. Lanfredi, R.B., Zhang, M., Auffermann, W., Chan, J., Duong, P.A., Srikumar, V., Drew, T., Schroeder, J., Tasdizen, T.: Reflacx: Reports and eye-tracking data for localization of abnormalities in chest x-rays (2021)
25. Li, G., Muller, M., Thabet, A., Ghanem, B.: Deepgcns: Can gcns go as deep as cnns? In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9267–9276 (2019)
26. Neves, J., Hsieh, C., Nobre, I.B., Sousa, S.C., Ouyang, C., Maciel, A., Duchowski, A., Jorge, J., Moreira, C.: Shedding light on ai in radiology: A systematic review and taxonomy of eye gaze-driven interpretability in deep learning. European Journal of Radiology p. 111341 (2024)
27. Noda, M., Yoshimura, H., Okubo, T., Koshu, R., Uchiyama, Y., Nomura, A., Ito, M., Takumi, Y., et al.: Feasibility of multimodal artificial intelligence using gpt-4 vision for the classification of middle ear disease: Qualitative study and validation. JMIR AI **3**(1), e58342 (2024)
28. Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D., Bagul, A., Langlotz, C., Shpanskaya, K., et al.: Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. arXiv preprint arXiv:1711.05225 (2017)
29. Rubin, J., Sanghavi, D., Zhao, C., Lee, K., Qadir, A., Xu-Wilson, M.: Large scale automated reading of frontal and lateral chest x-rays using dual convolutional neural networks. arXiv preprint arXiv:1804.07839 (2018)
30. Saab, K., Hooper, S.M., Sohoni, N.S., Parmar, J., Pogatchnik, B., Wu, S., Dunnmon, J.A., Zhang, H.R., Rubin, D., Ré, C.: Observational supervision for medical

image classification using gaze data. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part II 24. pp. 603–614. Springer (2021)
31. Sánchez-Oro, R., Nuez, J.T., Martínez-Sanz, G.: Radiological findings for diagnosis of sars-cov-2 pneumonia (covid-19). Medicina Clínica (English Edition) **155**(1), 36–40 (2020)
32. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision. pp. 618–626 (2017)
33. van Sonsbeek, T., Zhen, X., Mahapatra, D., Worring, M.: Probabilistic integration of object level annotations in chest x-ray classification. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 3630–3640 (2023)
34. Tahri Sqalli, M., Aslonov, B., Gafurov, M., Mukhammadiev, N., Sqalli Houssaini, Y.: Eye tracking technology in medical practice: a perspective on its diverse applications. Frontiers in Medical Technology **5**, 1253001 (2023)
35. Wang, B., Aboah, A., Zhang, Z., Bagci, U.: Gazesam: What you see is what you segment. arXiv preprint arXiv:2304.13844 (2023)
36. Wang, B., Pan, H., Aboah, A., Zhang, Z., Keles, E., Torigian, D., Turkbey, B., Krupinski, E., Udupa, J., Bagci, U.: Gazegnn: A gaze-guided graph neural network for chest x-ray classification. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 2194–2203 (2024)
37. Wang, S., Ouyang, X., Liu, T., Wang, Q., Shen, D.: Follow my eye: Using gaze to supervise computer-aided diagnosis. IEEE Transactions on Medical Imaging **41**(7), 1688–1698 (2022)
38. Wang, W., Xie, E., Li, X., Fan, D.P., Song, K., Liang, D., Lu, T., Luo, P., Shao, L.: Pvt v2: Improved baselines with pyramid vision transformer. Computational Visual Media **8**(3), 415–424 (2022)
39. Wang, Z., Wu, Z., Agarwal, D., Sun, J.: Medclip: Contrastive learning from unpaired medical images and text. arXiv preprint arXiv:2210.10163 (2022)
40. Watanabe, A., Ketabi, S., Namdar, K., Khalvati, F.: Improving disease classification performance and explainability of deep learning models in radiology with heatmap generators. Frontiers in radiology **2**, 991683 (2022)
41. Xie, Y., Yang, B., Guan, Q., Zhang, J., Wu, Q., Xia, Y.: Attention mechanisms in medical image segmentation: A survey. arXiv preprint arXiv:2305.17937 (2023)
42. You, K., Gu, J., Ham, J., Park, B., Kim, J., Hong, E.K., Baek, W., Roh, B.: Cxr-clip: Toward large scale chest x-ray language-image pre-training. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 101–111. Springer (2023)
43. Zhang, J., Huang, J., Jin, S., Lu, S.: Vision-language models for vision tasks: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence (2024)