Segmentation Aided Multiclass Classification of Lung Disease in Chest X-ray Images using Graph Neural Networks

Iftekharul Islam Shovon[a], Ijaz Ahmad[b], and Seokjoo Shin[a]
[a]Dept. of Computer Engineering, Chosun University, Gwangju, Korea
[b]Dept. of Electrical and Computer Engineering, Korea University, Seoul, Korea
shovon@chosun.kr, ijaz@korea.ac.kr, sjshin@chosun.ac.kr (*Corresponding author*)

***Abstract*—Deep learning (DL)-based medical image classification has become a pivotal research area in computer vision, significantly enhancing the diagnostic process across the medical field. DL-based image classification can be aided by supplementary data, further improving its performance. One prominent example of such data is gaze-points, which involve eye-tracking techniques to document radiologists' interaction with images. Recently, gaze-point data has been used as a pre-processing method to map an image into a graph, enabling the use of graph neural networks (GNNs) in computer vision domain. Such techniques have demonstrated significant improvements compared to conventional convolution neural networks (CNNs); however, their reliance on human involvement makes them labor-intensive, thus limiting their applications. To address this, we propose to leverage DL-based automatic segmentation mask generation to prepare image data as input for the GNN. This technique uses the segmentation mask as the attention information to guide the classifier. The results demonstrate that the proposed segmentation-aided classification model surpasses conventional CNN models and delivers the same performance as the existing supplementary data-aided techniques while reducing manual labor.**

***Index Terms*—GNN, medical image, classification**

I. Introduction

The recent advancements of artificial intelligence in various applications such as speech recognition, image classification etc. can be attributed to the progress made in the deep learning (DL) over the years [1]. DL, a subset of ML, is distinguished by its use of multi-layered neural networks, where the core principle is to abstract features from raw data, with these features progressively increasing in complexity and specificity at higher layers. This process reflects aspects of human cognitive processing, particularly how humans perceive and interpret vast amounts of information through hierarchical processing. With the advent of powerful computational resources and extensive datasets, DL has achieved remarkable success across various domains, including natural language processing, computer vision, and medical imaging, profoundly impacting academic research and industry applications.

In recent years, the use of DL models for image classification has significantly grown in popularity. Advanced classification algorithms have been widely adopted, such as CNNs, SVMs, DNNs, Transformers, RNNs, and Graph Neural Networks (GNNs) [2] for natural image classification. However, medical image classification poses distinct challenges [3]. For instance, X-ray images often display minimal contrast in soft tissues and contain intricate anatomical details that overlap within a two-dimensional space [4]. In these images, organs and blood vessels frequently exhibit similar intensities, complicating the classification process. To mitigate these issues, recent studies have introduced supplementary data such as eye-gaze points, which are collected via eye-tracking technologies during radiologist screenings. These gaze points generate patterns that assist in locating abnormalities in X-rays. The methodologies employing this data can be categorized into three main types: Attention Consistency Architecture [5], Two-Stream Architecture [6], and Gaze Data Input [7]. However, these approaches rely heavily on human intervention, requiring radiologists to use eye-tracking devices, thereby introducing a possibility for human error.

In this paper, we introduce a data-aided method that uses image segmentation mask in preparing an image as an input for a graph neural network (GNN)-based chest X-ray (CXR) image classification. A CXR image is initially processed to generate a masked image where some pixel intensity values are set to zero, while others remain non-zero [8]. This method eliminates human intervention in creating data for processing an image as a graph by using a lightweight U-Net segmentation mask [9], which is traditionally required. In our proposed system, the first step is patch embedding, in which the CXR images are segmented into patches, which are then processed by a transformer model to extract features from them. Alongside the patch embedding, positional embedding and mask embedding are utilized to construct a graph. This graph is subsequently fed into a GNN, which utilizes the entire graph in what is known as a graph-level task to make predictions. For comparison, we evaluate the performance of our model using metrics such as accuracy, AUC, precision, F1 score, and recall.

II. Proposed Method

This section outlines the framework of our proposed segmentation aided classification model for disease diagnosis in chest X-ray (CXR) images using GNNs. Fig. 1 depicts the overall framework of our proposed system that consists of two main modules: the graph generation and GNN-based classification modules. In the first module, a graph representation for

979-8-3315-0694-0/25/$31.00 ©2025 IEEE

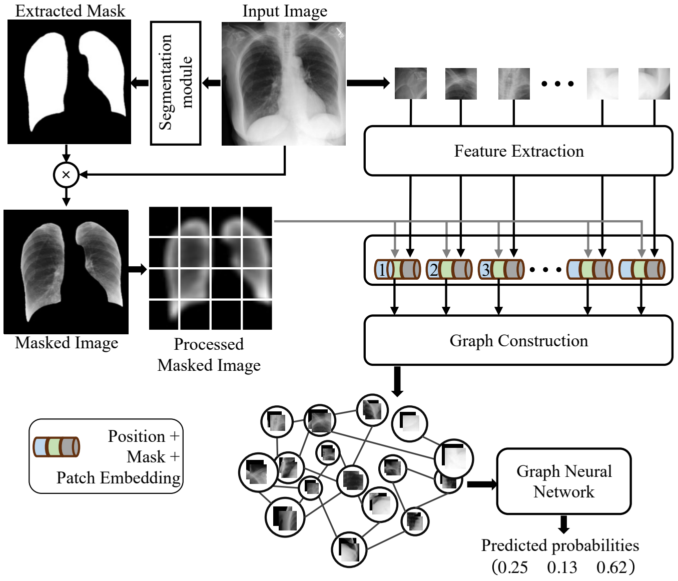


Fig. 1: A schematic representation of proposed method. The process involves feature extraction from X-ray images and their corresponding masks, followed by patch and position embeddings for graph construction, and the application of a graph neural network for disease prediction.

each image is generated by using their actual data and their corresponding masks. After this step, a GNN is utilized for classification purposes. Initially, the GNN updates and aggregates data from the nodes to create a comprehensive feature representation of the entire graph. Predictions are then made based on these graphs, known as graph-level classification. A detailed explanation for each step is given below.

A. Graph Representation

For graph representation, the proposed method takes two types of data as input: the CXR image and its corresponding mask. The mask data is differentiated from the regular grid structure of an image by containing only isolated lungs area. Also, this mask acts as the attention information.To construct a graph, we embed the mask and image data into feature vectors by integrating them using the following embedding system.

1) Patch Embedding: In an image $I_{H,W}$, the number of rows and columns can be given as the product of two integers such as $H = P \times Q$ rows and $W = R \times Q$ columns. Therefore, the image can be partitioned into $N = P \times R$ square blocks, each consisting of Q^2 pixels [10]. The image $I \in \mathbb{R}^{W \times H}$ is composed of N patches, $B = \{b_1, b_2, \ldots, b_N\}$, where each patch $b_j \in \mathbb{R}^{Q \times Q}$ for $j = 1, 2, \ldots, N$. To encode local image information, a feature vector $y_j^I \in \mathbb{R}^D$ can be extracted from each patch b_j as [11]:

$$y_j^I = G(b_j) \tag{1}$$

where $G(\cdot)$ a function that extracts features from image patches as proposed in [11]. For computational efficiency, we treat each patch as a graph node instead of individual pixels.

2) Mask Embedding: This section discusses the mask creation and embedding techniques utilized in our model. In medical image analysis, segmentation is the process of differentiating pixels representing lesions or organs from the background pixels [12]. This work implements U-Net [9] architecture, a lightweight DL model for automatic segmentation, that generates a segmentation mask to isolate the lung area in the CXR images. This isolated region is a focal point for identifying the area of interest, ensuring that only the relevant lung region is highlighted and eliminating potential errors where the model might incorrectly identify regions of interest. Furthermore, this method eliminates human intervention by allowing the original image to pass through a machine-learning model to produce the mask.

Similar to the input image, the dimensions of the masked image are $W \times H$, subdivided into patches of $Q \times Q$. Let $m_{(s_j,t_j)}$ be a pixel value at position (s_j, t_j) in the masked image, we process each patch B_i according to (2) given in [7] to represent its attention features. Therefore, the processed mask y_j^T is obtained as:

$$y_j^T = \sum_{(s_j,t_j) \in B_i} m(s_j, t_j) \tag{2}$$

3) Position Embedding: GNN treats features as unordered nodes during graph processing; therefore, we implement position embedding technique proposed in [13] to preserve the positional information of the original images. The positional embedding method consists of two steps. First, we add a learnable absolute positional encoding vector, $e_i \in \mathbb{R}^D$, to the feature vector $(y_j^l + y_j^T)$. Second, we calculate the relative positional distance between nodes as $e_i^T e_j$, which is then used as a distance metric in *k*-nearest neighbor algorithm to find the adjacent nodes of a given node for graph construction.

B. Graph Construction

To construct the graph $G = \{V, E\}$, with V vertices and (E) edges. V consist of mask embedding y_t, position embedding y_l and graph feature vector v_i

$$v_i = y_i^I + y_i^T + e_i. \tag{3}$$

To define the edge of the graph we use k-nearest neighbors,

$$E = \{(v_i, v_j) \mid v_j \in K(v_i)\}, \tag{4}$$

where $K(x)$ represents the neighbors of v_i. Using the vertices V and edges E we construct the graph $G = \{V, E\}$

C. Graph Neural Network

Our model consists of L graph processing blocks which were inspired by [13], featuring an average pooling layer as well as a graph classification head, supplemented with multiple fully connected (FC) layers and a graph convolution layer (GCN) [18]. If a graph is represented as N, with D-dimensional feature vectors, and the input of the graph at block t is $V^t = [v_1^t, v_2^t, \ldots, v_N^t] \in \mathbb{R}^{N \times D}$, the graph processing block outputs $Z^t \in \mathbb{R}^{N \times D}$ as:

$$U^t = \Omega_2(\Theta(\Omega_1(V^t))) + V^t \tag{5}$$

TABLE I: Performance analysis of proposed segmentation-aided classification model with existing deep learning techniques on the same dataset using different evaluation metrics.

Methods	Data Type	Accuracy	AUC				Precision	Recall	F1-Score	Human Intervention
			Normal	CHF	Pneumonia	Average				
Temporal Model [14]	-	-	0.890	0.850	0.680	0.810	-	-	-	Yes
Cooperative Learning [15]		80.00%	0.953	0.927	0.894	0.925	-	-	-	Yes
U-Net+Gaze [14]	Eye-gaze data	-	0.910	0.890	0.790	0.870	-	-	-	Yes
Densenet121+Gaze [16]		-	-	-	-	0.836	-	-	0.270	Yes
GazeMTL [17]		78.50%	0.915	0.913	0.833	0.887	0.786	0.781	0.779	Yes
IAA [6]		78.50%	0.915	0.913	0.875	0.900	0.780	0.781	0.779	Yes
EffNet+GG-CAM [5]		77.57%	0.906	0.914	0.843	0.888	0.770	0.773	0.770	Yes
GazeGNN [7]		79.76%	0.938	0.916	0.914	0.923	0.839	0.821	0.823	Yes
Proposed Method	Segmentation mask	80.40%	0.905	0.916	0.862	0.892	0.802	0.783	0.787	No

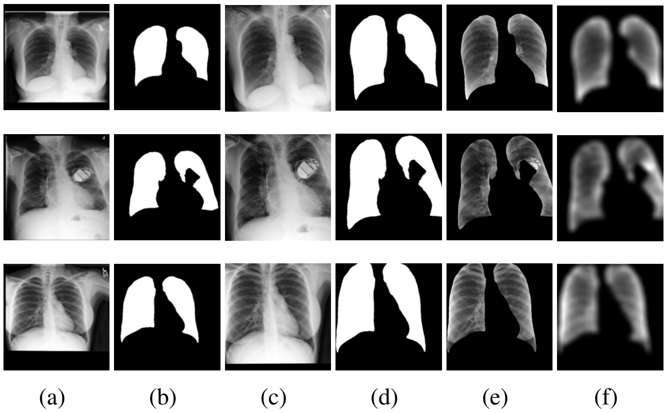


Fig. 2: Example images for each label from the dataset with their preprocessing results. The original images are in (a) and their corresponding masks are in (b), (c) and (d) are transformed images and masks, (e) shows the segmented region in the original image while (f) is the mask embedding. The images and masks are for Normal class in the first, CHF class in the second and Pneumonia in the third rows.

$$Z^t = \Omega_4(\Omega_3(U^t)) + U^t \tag{6}$$

Here, Θ denotes the graph convolution and Ω denotes FC layers. U^t represents the intermediate output after the first shortcut connection, and $M^t = \Omega_1(V^t)$ is the input for the graph convolution layer. Hence, the graph convolution $S^t = \Theta(M^t)$ is constructed as:

$$f_i^t = \mathbf{W} \cdot \max(\{m_i^t - m_j^t \mid j \in K(m_i^t)\}) \tag{7}$$

$\mathbf{W}$ is the learnable weight matrix for updating the feature of the node. The aggregation used here is the max function, which aggregates the maximum features from the i-th node's neighbors. Thus, the graph convolution aggregates neighbors' feature information into the node feature, and finally, the classification head, which is a series of FC layers with a softmax function, predicts the probability of each category.

III. Simulation Results

A. Experimental Setup

Using PyTorch, the experiment was conducted on a Windows PC equipped with Intel i5 CPU and an NVIDIA RTX 3060 GPU. The AdamW optimizer [19] was selected for the experiments. For model hyperparameters, we followed the training setup of [7]. Also, during training, the model with the highest accuracy was saved as our best model. For performance evaluation, we considered methods proposed in [5]–[7], [14]–[17] as baselines.

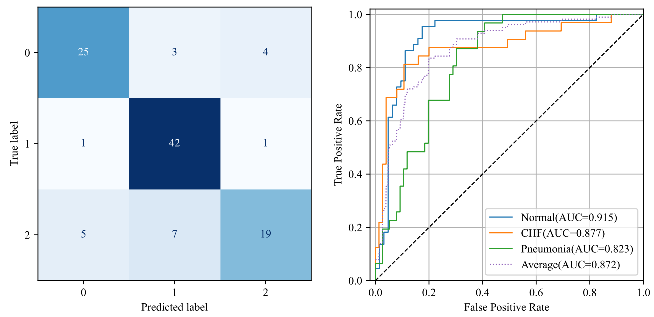
B. Dataset

The experiments were conducted on a publicly available CXR image dataset [20], which comprises 1083 samples divided into three classes: Normal, Congestive Heart Failure (CHF), and Pneumonia. The original images are of size 3000×3000, all in grayscale, which were resized to 224×224 using random center crop before partitioning them into patches. Besides random cropping, we applied random flip and rotation on the training set as our data augmentation technique. Fig.3. shows example images from our dataset for each class along with their corresponding masks and preprocessing to obtain mask embedding. The corresponding mask of each image in the dataset was obtained using a pre-trained U-Net model.

C. Performance Analysis

Table I summarizes a detailed performance of the proposed model in terms of accuracy, AUC, precision, recall and F1-score. For the baseline, we considered conventional methods [5]–[7], [14], [15], [17] implemented on the same dataset. These methods are divided into two groups based on their dependency on supplementary data: techniques [14], [15] does not require additional data, and techniques [5]–[7], [17] use eye-gaze data to guide their classification model. These metric scores are directly reported from [7], except for [7], which we got by running their available open-source code. Following the model architecture of [7], our proposed model employs a transformer model to learn patch embedding, and a GNN model to process graph representation of an image for the classification. [7]'s technique relies on eye-gaze data for graph generation, which requires human intervention thus making it labor intensive. On the other hand, proposed model leverages DL-based automatic segmentation mask generation to prepare image data as an input for the GNN.

From Table I, it is evident that our method surpasses models that do not require supplementary data in terms of accuracy. Even though [15] requires human intervention, it falls short of our model in terms of accuracy. Compared to data aided techniques, our proposed model achieved better classification performance across all metrics. Although GazeGNN outperforms



(a)

(b)

Fig. 3: Performance analysis of proposed model on each class using confusion matrix and Receiver-Operating Characteristic (ROC) curves. (a) shows the confusion matrix of proposed model showing the classification performance across three categories. The matrix highlights the model's effectiveness in predicting the correct labels, with relatively low classification rates. (b) plots the (ROC) curves comparing the True Positive Rate against the False Positive Rate for different classes. The model demonstrates strong discrimination ability, as evidenced by the high AUC values across all categories.

the proposed method on most evaluation metrics, except for accuracy, the advantage of our method is that it eliminates human intervention while still maintaining higher accuracy. Fig. 2(a) displays the confusion matrix of our proposed method, showing that most instances are correctly classified with strong results in class 1. Fig. 2(b) presents the Receiver-Operating Characteristic (ROC) curves, comparing the performance of different AUC values for various labels, including Normal (0.905), CHF (0.916), and Pneumonia (0.862). The average AUC of our proposed model is 0.892, further demonstrating its robustness in classification performance.

IV. Conclusion

This paper proposed a novel segmentation-aided medical image classification framework leveraging Graph Neural Networks (GNN). Our approach utilized chest X-ray images and corresponding masked images to construct a graph, which the GNN processes for disease classification. Proposed method dealt with the fundamental limitation of existing data-aided techniques for example, they rely on human intervention to collect the necessary data to construct a graph, by implementing a DL-based automatic technique. Results showed that proposed model is effective, outperforming classical and several existing data-aided techniques in terms of accuracy, precision, recall, F1 score, and average AUC scores.

In the future, we are interested to implement a more efficient technique to process the mask data for graph representation.

Acknowledgment

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government. (MSIT) (RS-2023-00278294).

References

[1] Y. Bengio, Y. Lecun, and G. Hinton, "Deep learning for ai," *Commun. ACM*, vol. 64, p. 58–65, June 2021.

[2] C. Zhang, C. Liu, X. Zhang, and G. Almpanidis, "An up-to-date comparison of state-of-the-art classification algorithms," *Expert Systems with Applications*, vol. 82, pp. 128–150, 2017.

[3] L. S. Chow and R. Paramesran, "Review of medical image quality assessment," *Biomedical signal processing and control*, vol. 27, pp. 145–154, 2016.

[4] M. Périard and P. Chaloner, "Diagnostic x-ray imaging quality assurance: an overview," *Canadian Journal of Medical Radiation Technology*, vol. 27, pp. 171–177, 1996.

[5] H. Zhu, S. Salcudean, and R. Rohling, "Gaze-guided class activation mapping: Leverage human visual attention for network attention in chest x-rays classification," in *Proceedings of 15th International Symposium on Visual Information Communication and Interaction*, pp. 1–8, 2022.

[6] Y. Gao, T. S. Sun, L. Zhao, and S. R. Hong, "Aligning eyes between humans and deep neural network through interactive attention alignment," *Proceedings of the ACM on Human-Computer Interaction*, vol. 6, no. CSCW2, pp. 1–28, 2022.

[7] B. Wang, H. Pan, A. Aboah, Z. Zhang, E. Keles, D. Torigian, B. Turkbey, E. Krupinski, J. Udupa, and U. Bagci, "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.

[8] I. Kakogeorgiou, S. Gidaris, B. Psomas, Y. Avrithis, A. Bursuc, K. Karantzalos, and N. Komodakis, "What to hide from your students: Attention-guided masked image modeling," in *Computer Vision – ECCV 2022*, pp. 300–318, Springer Nature Switzerland, 2022.

[9] M. Yahyatabar, P. Jouvet, and F. Cheriet, "Dense-unet: a light model for lung fields segmentation in chest x-ray images," in *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pp. 1242–1245, IEEE, 2020.

[10] I. Ahmad, W. Choi, and S. Shin, "Comprehensive analysis of compressible perceptual encryption methods—compression and encryption perspectives," *Sensors*, vol. 23, no. 8, 2023.

[11] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "Pvt v2: Improved baselines with pyramid vision transformer," *Computational Visual Media*, vol. 8, pp. 415–424, Sep 2022.

[12] M. H. Hesamian, W. Jia, X. He, and P. Kennedy, "Deep learning techniques for medical image segmentation: achievements and challenges," *Journal of digital imaging*, vol. 32, pp. 582–596, 2019.

[13] K. Han, Y. Wang, J. Guo, Y. Tang, and E. Wu, "Vision gnn: an image is worth graph of nodes," in *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, (Red Hook, NY, USA), Curran Associates Inc., 2024.

[14] A. Karargyris, S. Kashyap, I. Lourentzou, J. T. Wu, A. Sharma, M. Tong, S. Abedin, D. Beymer, V. Mukherjee, E. A. Krupinski, *et al.*, "Creation and validation of a chest x-ray dataset with eye-tracking and report dictation for ai development," *Scientific data*, vol. 8, no. 1, p. 92, 2021.

[15] Z. Qiu, H. Rivaz, and Y. Xiao, "Joint chest x-ray diagnosis and clinical visual attention prediction with multi-stage cooperative learning: enhancing interpretability," *arXiv preprint arXiv:2403.16970*, 2024.

[16] C. Ma, L. Zhao, Y. Chen, S. Wang, L. Guo, T. Zhang, D. Shen, X. Jiang, and T. Liu, "Eye-gaze-guided vision transformer for rectifying shortcut learning," *IEEE Transactions on Medical Imaging*, vol. 42, no. 11, pp. 3384–3394, 2023.

[17] K. Saab, S. M. Hooper, N. S. Sohoni, J. Parmar, B. Pogatchnik, S. Wu, J. A. Dunnmon, H. R. Zhang, D. Rubin, and C. Ré, "Observational supervision for medical image classification using gaze data," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, Sep 27–Oct 1, 2021, Proceedings, Part II 24*, pp. 603–614, Springer, 2021.

[18] G. Li, M. Muller, A. Thabet, and B. Ghanem, "Deepgcns: Can gcns go as deep as cnns?," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9267–9276, 2019.

[19] I. Loshchilov, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

[20] A. Karargyris, S. Kashyap, I. Lourentzou, J. T. Wu, A. Sharma, M. Tong, S. Abedin, D. Beymer, V. Mukherjee, E. A. Krupinski, *et al.*, "Creation and validation of a chest x-ray dataset with eye-tracking and report dictation for ai development," *Scientific data*, vol. 8, no. 1, p. 92, 2021.