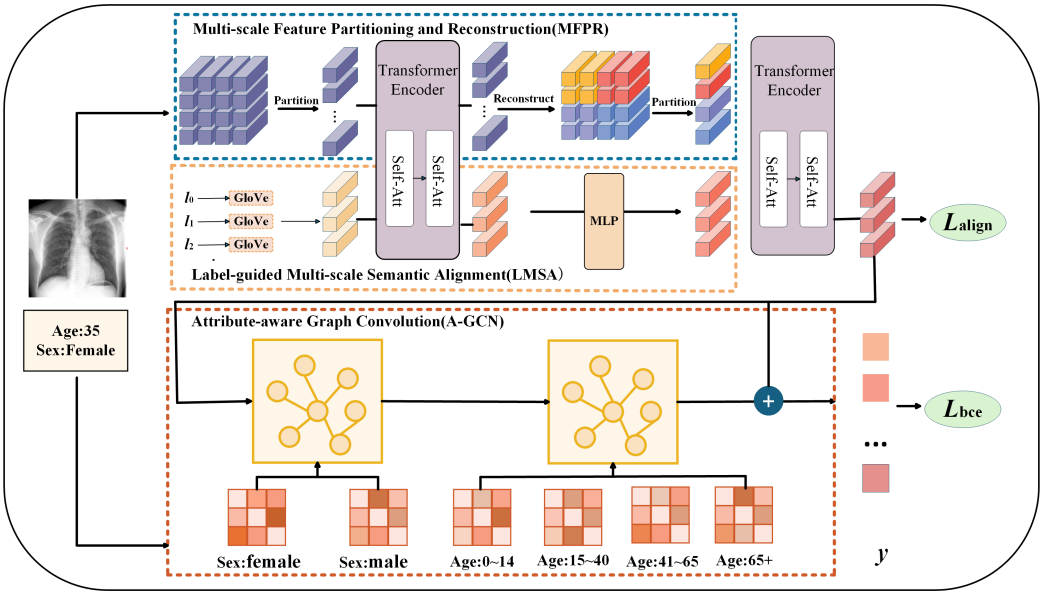
Graphical Abstract

A Multi-label Chest X-Ray Image Classification Algorithm Based on Multi-Scale and Attribute-Aware Semantic Graph

Qian Wang, Zhijuan Wu, Jiyu Gao, Hongnian Yu, Yongqiang Cheng



Highlights

A Multi-label Chest X-Ray Image Classification Algorithm Based on Multi-Scale and Attribute-Aware Semantic Graph

Qian Wang, Zhijuan Wu, Jiyu Gao, Hongnian Yu, Yongqiang Cheng

- A multi-scale feature partitioning method is proposed to improve lesion recognition.
- A label-guided alignment method is designed to enhance visual-semantic consistency.
- An attribute-aware graph method is proposed to better capture label dependencies.

A Multi-label Chest X-Ray Image Classification Algorithm Based on Multi-Scale and Attribute-Aware Semantic Graph

Qian Wang[a,*], Zhijuan Wu[a], Jiyu Gao[a], Hongnian Yu[b], Yongqiang Cheng[c]

[a]*School of Information Science and Engineering, Yanshan University, Qinhuangdao, 066004, Hebei, China*
[b]*School of Computing Engineering and the Built Environment, Edinburgh Napier University, Edinburgh, EH10 5DT, UK*
[c]*School of Computer Science, University of Sunderland, Sunderland, SR1 3SD, UK*

Abstract

Multi-label Chest X-Ray classification is crucial for intelligent diagnosis, yet existing algorithms usually ignore lesion-scale heterogeneity and attribute-conditioned label dependencies, limiting their clinical generalizability. To address these issues, this paper proposes MSASG, a multi-label Chest X-Ray image classification algorithm based on Multi-Scale and Attribute-aware Semantic Graph which enhances discriminative power and semantic consistency. Firstly, a Multi-scale Feature Partitioning and Reconstruction method is proposed to capture lesion patterns at different scales. Secondly, a Label-guided Multi-scale Semantic Alignment method is proposed to improve visual–semantic alignment by integrating label embeddings into feature extraction and using a Transformer to model high-order cross-modal dependencies. Finally, an Attribute-aware Graph Convolutional Network method is proposed to construct attribute-specific label co-occurrence matrices and dynamically select relevant structures during inference, enabling personalized characterization of label dependencies. Experiments on ChestX-ray14 and CheXpert show that MSASG outperforms state-of-the-art methods in recognizing complex lesion co-occurrence and adapting to heterogeneous popula-

*Corresponding author

Email addresses: wangqian@ysu.edu.cn (Q. Wang), wzj0911@stumail.ysu.edu.cn (Z. Wu), gjy9184@stumail.ysu.edu.cn (J. Gao), H.Yu@napier.ac.uk (H. Yu), yongqiang.cheng@sunderland.ac.uk (Y. Cheng)

tions.

Keywords:
Medical image analysis, Chest X-ray, Multi-label classification, Graph convolutional network, Transformer

1. Introduction

Chest X-Ray (CXR) is a widely used medical imaging technique due to its efficiency, low cost, and minimal radiation exposure. It plays a crucial role in the early detection and diagnosis of various thoracic diseases, such as nodules, cardiomegaly, and edema. The frequent co-occurrence of multiple abnormalities in CXR images has rendered multi-label classification one of the core tasks in intelligent CXR image interpretation. In recent years, deep learning has achieved remarkable progress in this field. A notable milestone is CheXNet (Rajpurkar et al., 2017), a DenseNet121-based algorithm that achieved radiologist-level pneumonia detection performance, significantly promoting the widespread application of convolutional neural networks in CXR multi-label classification (Huang et al.; He et al., 2017; 2016). Subsequently, researchers have persistently optimized network architectures. For instance, Chowdary et al. (Chowdary and Kanhangad, 2022) proposed a dual-branch structure that integrates global and local features to facilitate richer semantic representations. Similarly, Xu et al. (Xu and Duan, 2024) incorporated both image-level and lesion-level attention mechanisms to improve the detection of small-scale lesions. Meanwhile, the Transformer architecture, with its global representation capabilities, has been progressively incorporated into this task to address the limitations of convolutional neural networks in constructing long-range dependencies (Jiang et al., 2024). Furthermore, to capture semantic dependencies among disease labels, Lee et al. (Lee et al., 2022) introduced a label graph structure powered by graph neural networks. Although the classification performance has been improved, most existing algorithms still rely heavily on visual features, which limit s their generalizability across varying lesion scales and diverse patient populations.

In clinical practice, thoracic diseases exhibit substantial variation in lesion scale, diverse semantic characteristics, and label co-occurrence patterns shaped by patient-specific attributes such as age and gender. These factors pose significant challenges for automated diagnosis, as they affect both visual

representation learning and label prediction. While existing algorithms have made progress in visual backbone design and in capturing label-level correlations, they often fall short of addressing these complex, interdependent factors in the context of real-world chest X-ray interpretation.

Notably, current multi-label CXR image classification still faces three key challenges: (1) Insufficient representation of spatial-scale heterogeneity in lesion regions. Thoracic abnormalities exhibit considerable variation in spatial scale, ranging from small nodules to extensive diffuse opacities. Traditional Convolutional Neural Networks(CNNs) often fail to capture such large-scale variations, thereby limiting their effectiveness in lesion localization and recognition (Wang et al., 2021). (2) Inadequate cross-modal alignment between visual and semantic information. Label semantics are typically introduced only at the prediction stage, offering limited guidance during feature extraction. The weak associations under multi-scale fusion ultimately compromise the discriminative effectiveness of the classification system (Zhao et al., 2021). (3)Weak personalized learning of label semantic dependencies. The co-occurrence structure among labels is strongly influenced by attributes such as age and gender. However, existing static graph methods fail to capture such population heterogeneity and lack mechanisms for attribute-aware adaptation, ultimately limiting their generalization in complex clinical scenarios (Chen et al., 2021a).

In response to the key challenges, a multi-label CXR image classification algorithm based on Multi-Scale and Attribute-aware Semantic Graph(MSASG) is proposed. Firstly, a Multi-scale Feature Partitioning and Reconstruction (MFPR) method is designed to integrate fine-grained local details and global contextual information through local partitioning and spatial reconstruction, thereby enhancing the ability to capture lesion scale heterogeneity. Secondly, the Label-guided Multi-scale Semantic Alignment (LMSA) method is introduced, in which label embeddings are incorporated into a Transformer-based structure as semantic priors to improve visual–semantic alignment. Finally, an Attribute-aware Graph Convolutional Network (A-GCN) is proposed to construct a conditional label co-occurrence matrix based on individual attributes, thereby improving generalization across heterogeneous populations. While the proposed algorithm is designed for chest X-ray interpretation, it possesses strong generalizability and can be effectively applied to other multi-label medical imaging tasks, such as fundus images for detecting diabetic retinopathy and glaucoma, and skin lesion images for identifying co-existing dermatological conditions like melanoma and seborrheic keratosis. The prin-

cipal contributions are summarized as follows:

- Considering the spatial-scale heterogeneity of lesions, a multi-scale feature partitioning and reconstruction method is introduced to extract local features at multiple scales, thereby enhancing lesion recognition across diverse spatial patterns.
- Considering the importance of aligning label semantics with visual features, a label-guided multi-scale semantic alignment method is designed, where label embeddings are incorporated into the visual feature extraction process to improve visual-semantic alignment.
- Considering the population heterogeneity in semantic dependencies, an attribute-aware graph convolutional network is constructed, where attributes are integrated into both label co-occurrence construction and the graph propagation process to better capture label dependencies.

2. Related Work

Multi-label CXR image classification has become a pivotal task in intelligent medical image analysis, garnering increasing attention due to its significance in screening, diagnosis, and clinical decision-making. The emergence of large-scale public datasets such as ChestX-ray14 (Wang et al., 2017) and CheXpert (Irvin et al., 2019) has greatly accelerated the adoption of deep neural networks in this domain, laying a solid foundation for automated lesion recognition and multi-label prediction.

However, lesion regions in CXR images exhibit substantial heterogeneity in scale, morphology, and location, ranging from small pulmonary nodules to extensive cardiomegaly. Owing to the fixed receptive fields, traditional CNNs (Jin et al.; Wang et al., 2023; 2024) struggle to simultaneously capture both local details and global semantic context, thereby limiting their effectiveness in recognizing lesions across diverse scales. To enhance attention to critical regions, Chen et al.(Chen et al., 2019a) introduced lesion location information to provide spatial guidance, while Kamal et al.(Kamal et al., 2022) integrated anatomical structure priors to guide the feature extraction process toward potential lesion areas. Nonetheless, such methods still primarily rely on local perception mechanisms, making it difficult to capture long-range dependencies that are spatially distant but semantically related. The widespread adoption of the Transformer architecture (Vaswani et al., 2017) in

computer vision tasks (Dosovitskiy et al.; Yuan et al., 2020; 2021) has introduced a new paradigm for CXR multi-label classification by effectively capturing long-range dependencies. Wang et al.(Wang et al., 2021) introduced the Vision Transformer (ViT) into this task, achieving notable improvements in capturing complex structural patterns. Building upon this, PCAN(Zhu et al., 2022) enhanced lesion-related semantic representations through channel attention, while STERN (Rocha et al., 2024) employed multi-scale spatial attention to emphasize discriminative regions. Ding et al. (Ding et al., 2023) proposed a detection framework that integrates convolutional block attention module(CBAM), deformable convolution network(DCN), and multiscale feature encoding(MSFE) to address the need for multi-scale target modeling in medical images, providing a new perspective for scale adaptation. In addition, Zhang et al. (Zhang et al., 2023) combined attention mechanisms with contrastive learning to further improve region-level recognition accuracy. These algorithms substantially improve the perception of complex lesion structures from multiple perspectives, such as channel, spatial, and contrastive dimensions, thereby enhancing multi-label classification performance. Despite these advances, most Transformer-based algorithms adopt fixed-size patch partitioning, which limits the capacity to address structural heterogeneity and fuzzy lesion boundaries in medical images, and often fail to preserve spatial continuity.

On the other hand, multi-label classification requires not only image-level discriminative capability but also precise alignment between semantic labels and corresponding image regions. Early methods (Chen et al.; Pham et al., 2019b; 2021) perform classification as a separate step after feature extraction, without establishing explicit links between semantic labels and spatial regions. This limits their capacity for region-level semantic interpretation. To enhance semantic alignment, Wang et al. (Wang et al., 2018) proposed the TieNet algorithm, which performs cross-modal representation learning by integrating image and text embeddings, thereby improving the alignment between image regions and semantic labels. Ding et al. (Ding et al., 2025b) employed modality-invariant representation and progressive registration to achieve cross-modal alignment, suggesting that modality-invariant feature spaces can improve semantic stability. However, most existing algorithms still lack semantic guidance during the early stages of visual feature encoding. In particular, the correspondence between visual regions and semantic concepts remains ambiguous during multi-scale feature extraction. This limitation weakens the discriminative consistency and generalization capability

in complex multi-label classification scenarios.

Moreover, in multi-label CXR image classification, disease labels typically exhibit complex semantic co-occurrence patterns and hierarchical dependencies. Effectively constructing these label dependencies is crucial for improving prediction accuracy. The introduction of graph convolutional networks (GCNs) (Kipf and Welling, 2016) has established a robust paradigm for capturing label dependencies. Chen et al. (Chen et al., 2021a) constructed a static label co-occurrence graph based on empirical co-occurrence statistics and utilized GCNs to enable collaborative label reasoning, thereby pioneering the integration of graph-based methods in CXR multi-label classification. Subsequent advancements have extended this foundation. Zhao et al. (Mao et al., 2022) proposed ImageGCN, which incorporated cross-instance graph structures to enhance the semantic richness of label representations. Zhou et al. (Yang et al., 2021) employed graph gating and attention mechanisms to adaptively update the label graph structure. To further strengthen the expressiveness and reliability of label graph construction, anatomical priors (Lian et al., 2021) and hierarchical label taxonomies (Chen et al., 2021b) have been integrated, promoting structural validity and label consistency. Ding et al. (Ding et al., 2025a) integrated high-frequency prior information with local class-regional label guidance in diffusion models to enhance cross-modality image translation, highlighting the importance of structural priors and label constraints in complex scene modeling. Despite these progresses, many existing algorithms still rely on global co-occurrence patterns and static relationships, without accounting for patient-specific factors that modulate label dependencies. Although several studies have investigated dynamic graph construction (Zhu et al.; Hu et al., 2023; 2023), the influence of individual attributes such as age and gender on disease co-occurrence remains largely underexplored. This omission restricts the adaptability and generalization capacity of current algorithms in clinically diverse and demographically heterogeneous populations.

In summary, existing algorithms still face limitations in scale-aware recognition, semantic alignment, and personalized label reasoning. To address these challenges, a multi-label CXR image classification algorithm based on multi-scale and attribute-aware semantic graph is proposed, incorporating multi-scale perception, semantic alignment, and attribute-aware dependency learning.

Table 1: List of notations used in the paper

Notation	Description
x_n	The n-th input CXR image
y_n	The label of x_n
$\mathbf{a}_n$	The attribute vector of x_n
$\mathbf{F}_i$	The spatial feature map at scale i
$\mathbf{T}_i$	The visual token sequence after splitting at scale i
$\mathbf{E}_i$	The semantic embedding at scale i
$\mathbf{H}^l$	The label embedding at the l-th GCN layer
R	The set of attributes
$V^{(r)}$	The value set of attribute r
$D^{(r)}(v)$	The subset of samples where attribute r takes the value v
$\mathbf{N}^{(r)}(v)$	The co-occurrence frequency matrix
$\mathbf{A}^{(r)}(v)$	The attribute-aware label co-occurrence matrix

3. The Proposed Algorithm

To enhance disease recognition in multi-label CXR image classification, this paper proposes an algorithm based on the multi-scale and attribute-aware semantic graph. As illustrated in Fig. 1, MSASG consists of four main components. Firstly, in the multi-scale feature partitioning and reconstruction method, multi-scale visual features are extracted through iterative spatial partitioning and reconstruction. Secondly, in the label-guided multi-scale semantic alignment method, label embeddings are integrated as semantic priors to explicitly align visual and semantic features across scales. Then, in the attribute-aware semantic graph method, label co-occurrence dependencies are constructed by incorporating patient-specific attributes. Finally, to support multi-label prediction, the entire algorithm is trained end-to-end using a joint loss function. The notations used are summarized in Table 1.

3.1. Multi-scale Feature Partitioning and Reconstruction

To enable multi-scale visual representation learning, a MFPR method is introduced. It adopts a recursive strategy of feature partitioning and spatial reconstruction to progressively extract and integrate image representations at varying scales. By capturing both local structures and global contextual information, MFPR facilitates accurate perception of lesion regions with diverse spatial distributions.

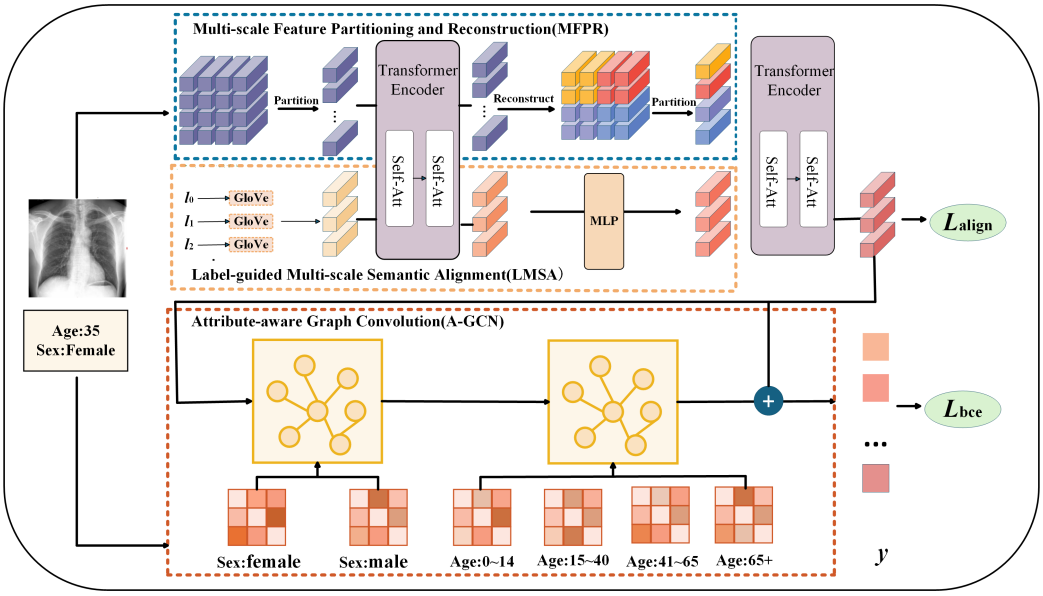


Figure 1: Overview of the proposed MSASG algorithm

Feature Partitioning. Given an input spatial feature map $\mathbf{F}_i \in \mathbb{R}^{h\times w\times c}$, where h, w, and c denote the height, width, and number of channels respectively, a soft partitioning strategy is adopted to partition $\mathbf{F}_i$ into multiple overlapping local patches. Each patch is of size $k \times k$, extracted with stride s and zero-padding p to preserve spatial coverage. The total number of local patches, denoted as l_i, is calculated as:

$$l_i = \lfloor (h + 2p - k)/(k - s) + 1 \rfloor \times \lfloor (w + 2p - k)/(k - s) + 1 \rfloor \tag{1}$$

After flattening, each local patch is linearly projected into a d-dimensional feature vector, forming a visual token sequence $\mathbf{T}_i \in \mathbb{R}^{l_i\times d}$.

Feature Reconstruction. After interaction with label semantic embeddings, the token sequence $\mathbf{T}_i$ is updated to $\mathbf{T}'_i$. To restore the spatial structure and preserve local positional information, a reconstruction operation is performed to map the sequence back into a spatial feature map:

$$\mathbf{F}_{i+1} = \text{Reshape}(\mathbf{T}'_i) \tag{2}$$

where $\mathbf{T}'_i \in \mathbb{R}^{l_i\times d}$ represents the updated token sequence, and the reconstructed feature map is $\mathbf{F}_{i+1} \in \mathbb{R}^{h'\times w'\times d}$, with $l_i = h' \times w'$. Here, h', w',

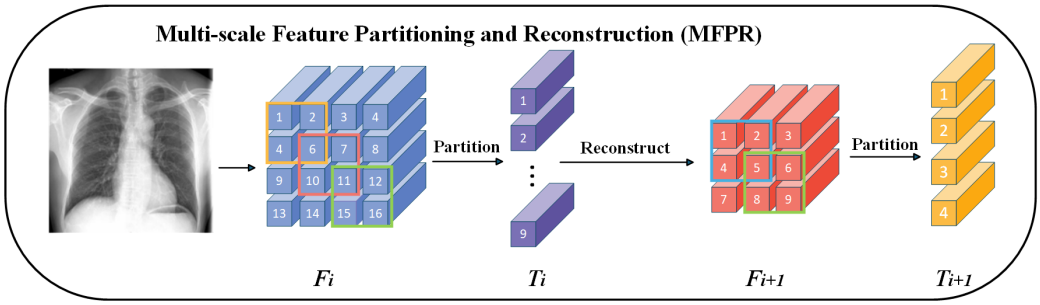


Figure 2: Illustration of Multi-scale Feature Partitioning and Spatial Reconstruction

and d denote the height, width, and number of channels in the reconstructed feature map, respectively.

This partitioning and reconstruction process can be recursively applied to form the core pathway of the MFPR method. As illustrated in Figure 2, MFPR initially partitions the input feature map $\mathbf{F}_1$ into a token sequence $\mathbf{T}_1$, which is subsequently fused with label semantics and reconstructed into $\mathbf{F}_2$. In each following iteration, the feature map $\mathbf{F}_i$ is partitioned into a token sequence $\mathbf{T}_i$, updated through semantic interaction, and then reconstructed into $\mathbf{F}_{i+1}$. Through this recursive "partitioning $\rightarrow$ interaction $\rightarrow$ reconstruction" mechanism, MFPR progressively extracts and integrates multi-scale semantic information, thereby improving the accuracy and structural integrity of lesion region representation.

3.2. Label-guided Multi-scale Semantic Alignment

To enhance the semantic perception of lesion regions and capture intricate cross-modal dependencies between visual features and label semantics, a LMSA method is introduced. Built upon the multi-scale visual representations extracted by the MFPR module, LMSA integrates label semantic embeddings to establish a unified cross-modal representational space. Furthermore, explicit alignment between visual features and label semantics is achieved at multiple scales through a Transformer encoder, enabling more accurate and comprehensive understanding of the lesion regions.

Specifically, let the multi-label set be denoted as $L = \{l_1, l_2, \ldots, l_C\}$, where C represents the total number of categories. A pretrained word embedding algorithm (e.g., GloVe) is utilized to encode each label into a semantic vector, resulting in the initial label embedding matrix:

$$\mathbf{E} = \{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_C\}, \quad \mathbf{E} \in \mathbb{R}^{C \times d} \tag{3}$$

where $\mathbf{e}_j \in \mathbb{R}^d$ denotes the embedding vector of the j-th label, and d is the embedding dimension.

At the i-th scale level, a sequence of visual tokens $\mathbf{T}_i \in \mathbb{R}^{l_i \times d}$ are extracted by MFPR method. To enable cross-modal semantic interaction, the visual token sequence is concatenated with the corresponding label embedding matrix $\mathbf{E}_i \in \mathbb{R}^{C \times d}$, where $\mathbf{E}_1$ is initialized as $\mathbf{E}$ at the initial scale level, yielding the cross-modal input sequence:

$$\mathbf{Z}_i = [\mathbf{T}_i; \mathbf{E}_i], \quad \mathbf{Z}_i \in \mathbb{R}^{(l_i+C) \times d} \tag{4}$$

Then, a Transformer encoder is adopted to model the cross-modal sequence $\mathbf{Z}_i$. Its multi-head self-attention mechanism jointly captures high-order interactions among visual tokens, label embeddings, and their semantic associations. To this end, the input sequence $\mathbf{Z}_i$ is first linearly projected to query, key, and value representations, and the scaled dot-product attention is computed as:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\mathbf{Q}\mathbf{K}^\top / \sqrt{d}\right)\mathbf{V} \tag{5}$$

where $\mathbf{Q} = \mathbf{Z}_i\mathbf{W}_Q$, $\mathbf{K} = \mathbf{Z}_i\mathbf{W}_K$, and $\mathbf{V} = \mathbf{Z}_i\mathbf{W}_V$, with $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V \in \mathbb{R}^{d \times d}$ denoting the learnable weight matrices.

By aggregating outputs from multi-heads, the updated visual and semantic representations are obtained:

$$[\mathbf{T}'_i \in \mathbb{R}^{l_i \times d},\ \mathbf{E}'_i \in \mathbb{R}^{C \times d}] = \text{Attention}(\mathbf{Z}_i) \tag{6}$$

here, $\mathbf{T}'_i$ denotes the visual features enriched with semantic context, while $\mathbf{E}'_i$ represents the label embeddings enhanced with visual information. To enable dynamic semantic propagation across scales, the $\mathbf{E}'_i$ is linearly transformed to generate the label embeddings for the next scale:

$$\mathbf{E}_{i+1} = \mathbf{E}'_i\mathbf{W}_e \tag{7}$$

By recursively applying the above process across multiple scales, the LMSA method facilitates the fusion of semantic information within the visual feature space, ultimately generating a global label representation $\mathbf{E}^{final}$. To further guide the learning process, a multi-label alignment loss $\mathcal{L}_{\text{align}}$ is introduced, providing explicit supervision for learning semantically consistent and discriminative label embeddings. As a result, the performance of multi-label recognition in complex scenes is improved.

3.3. Attribute-aware Graph Convolutional Network

To improve the capture of multi-label semantic dependencies, an A-GCN is introduced. A-GCN integrates attribute-conditioned mechanisms into both the construction of label dependencies and the graph propagation process. Specifically, it constructs attribute-conditioned label co-occurrence matrices and dynamically selects graph structures in accordance with instance-specific attributes. The overall process is shown in Algorithm 1.

3.3.1. Construction of Attribute-conditioned Label Co-occurrence Matrices

Label co-occurrence is a common phenomenon in CXR images. However, such co-occurrence is heavily influenced by individual attributes (e.g., gender, age), leading to structural variations across different subgroups. To capture these attribute-driven differences, an attribute-conditioned construction of label co-occurrence matrices is proposed, which explicitly constructs label dependencies within distinct attribute contexts.

The attribute space is composed of m discrete dimensions, denoted as $R = \{r_1, r_2, \ldots, r_m\}$. Each attribute r is associated with a set of discrete values $V^{(r)} = \{v_1^{(r)}, v_2^{(r)}, \ldots\}$. For example, gender corresponds to $\{\text{male}, \text{female}\}$, and age is represented as $\{\text{child}, \text{young}, \text{middle-aged}, \text{elderly}\}$. For any attribute-value pair (r, v), the training set is partitioned into subsets $D^{(r)}(v)$ where attribute r takes the value v . Within each subset, the co-occurrence frequency of label pairs (y_i, y_j) is computed as:

$$\mathbf{N}_{ij}^{(r)}(v) = \text{count}(y_i = 1 \wedge y_j = 1 \mid r = v) \tag{8}$$

then the co-occurrence frequency matrix $\mathbf{N}_{ij}^{(r)}(v)$ is normalized to obtain the attribute-conditioned label co-occurrence matrix :

$$\mathbf{A}_{ij}^{(r)}(v) = \mathbf{N}_{ij}^{(r)}(v) / \mathbf{N}_{i}^{(r)}(v) \tag{9}$$

where $\mathbf{N}_{i}^{(r)}(v) = \sum_j \mathbf{N}_{ij}^{(r)}(v)$. Specifically, $\mathbf{A}_{ij}^{(r)}(v)$ denotes the probability of label y_j co-occurring with label y_i, conditioned on attribute r taking the value v.

To reduce noise from low-frequency co-occurrence, a confidence threshold τ is used to sparsify entries with probabilities below τ:

$$\mathbf{A}_{ij}^{(r,v)} = \begin{cases} \mathbf{N}_{ij}^{(r)}(v) / \mathbf{N}_{i}^{(r)}(v), & \text{if } \mathbf{N}_{ij}^{(r)}(v) / \mathbf{N}_{i}^{(r)}(v) \geq \tau \\ 0, & \text{otherwise} \end{cases} \tag{10}$$

By filtering out non-significant label relationships, this strategy enhances both the structural integrity and robustness of the co-occurrence matrix. For each attribute r, a set of conditional co-occurrence matrices $\{\mathbf{A}^{(r)}(v) \mid v \in V^{(r)}\}$ is constructed. During the graph convolution process, these matrices are dynamically selected based on the attribute values of each sample, enabling personalized semantic propagation and improving the context adaptability of label embeddings.

3.3.2. Attribute-Aware Graph Convolution

Graph Convolutional Network (GCN) is a classical graph neural network that aggregates and propagates node features based on the adjacency structure. It has been widely adopted in structured semantic tasks due to its effectiveness in capturing topological dependencies through iterative message passing. Let C denote the number of labels, $\mathbf{A} \in \mathbb{R}^{C\times C}$ the adjacency matrix, and I the identity matrix. To incorporate self-loops, the adjacency matrix is augmented as $\tilde{\mathbf{A}} = \mathbf{A} + \mathbf{I}$. The corresponding degree matrix is defined as $\tilde{\mathbf{D}}_{ii} = \sum_j \tilde{\mathbf{A}}_{ij}$. The normalized adjacency matrix is then given by:

$$\hat{\mathbf{A}} = \tilde{\mathbf{D}}^{-1/2}\tilde{\mathbf{A}}\tilde{\mathbf{D}}^{-1/2} \tag{11}$$

The node features at the l-th layer are denoted as $\mathbf{H}^{(l)} \in \mathbb{R}^{C\times d_l}$, where C is the number of nodes and d_l is the feature dimension at layer l. The corresponding trainable weight matrix is represented by $\mathbf{W}^{(l)} \in \mathbb{R}^{d_l\times d_{l+1}}$. The standard forward propagation rule in GCN is formulated as:

$$\mathbf{H}^{(l+1)} = \sigma(\hat{\mathbf{A}}\mathbf{H}^{(l)}\mathbf{W}^{(l)}) \tag{12}$$

However, in multi-label CXR image classification, using a fixed graph structure for feature propagation cannot fully capture the semantic differences of label relationships in different attribute contexts. To address this issue, this paper proposes an A-GCN method that dynamically updates label embeddings by incorporating sample-level attribute information. Each graph convolutional layer in A-GCN is conditioned on a specific attribute dimension r. For a given sample with attribute value $v \in V^{(r)}$, the corresponding attribute-specific adjacency matrix $\mathbf{A}^{(r)}(v)$ is selected to guide feature propagation. The output of the LMSA method serves as the initial label representation, denoted by $\mathbf{H}^{(0)} = \mathbf{E}^{final}$. The layer-wise propagation at layer l is defined as:

$$\mathbf{H}^{(l+1)} = \sigma(\hat{\mathbf{A}}^{(r)}(v)\mathbf{H}^{(l)}\mathbf{W}^{(l)}) \tag{13}$$

where $\hat{\mathbf{A}}^{(r)}(v)$ is the normalized form of $\mathbf{A}^{(r)}(v)$, and $\sigma(\cdot)$ denotes a non-linear activation function.

By stacking multiple layers of graph convolutions based on different attribute dimensions, the MSASG algorithm can integrate the dynamic correlation structure between labels in the multi-attribute context layer by layer, enabling personalized label representation in the attribute space.

3.4. Multi-Label Classification Loss Function

In multi-label CXR image classification, each image is associated with a binary vector $\mathbf{y} = [y_1, y_2, \ldots, y_C]$,where $y_i = 1$ indicates the presence of the i-th disease label. To capture spatial and semantic dependencies, the algorithm integrates the semantic representation $\mathbf{E}^{final}$ with the attribute-enriched embedding $\mathbf{H}^{(L)}$. The fused vector $\mathbf{Z}$ is then processed by a sigmoid function to produce the final prediction probabilities:

$$\hat{\mathbf{y}} = \sigma\left(\mathbf{Z}\right) \tag{14}$$

Owing to the inherent class imbalance, where the majority of disease labels are negative, a class-aware weighted binary cross-entropy loss is adopted in both branches to improve recognition of minority classes. Specifically, for each label, If $y_i = 1$, the positive loss term is scaled by a factor of N_n/N_p, where N_p and N_n denote the numbers of positive and negative samples for each label within the current batch, respectively. The weighted binary cross-entropy loss is defined as:

$$L_{bce} = -\sum_{y_i=1}(N_n/N_p)\log(\hat{y}_i) - \sum_{y_i=0}\log(1-\hat{y}_i) \tag{15}$$

the overall objective integrates both classification and alignment losses:

$$L_{total} = L_{cls} + \lambda L_{align} \tag{16}$$

where $\lambda \in [0, 1]$ is a trade-off coefficient that balances the contribution of each component.

Algorithm 1 Attribute-aware Graph Convolution

Input: Multi-label dataset $D = \{(x_n, y_n, \mathbf{a}_n)\}_{n=1}^{N}$; Attribute dimensions $R = \{r_1, r_2, \ldots, r_m\}$; Discrete values $V^{(r)}$ for each $r \in \mathbb{R}$; Threshold τ; Number of GCN layers L.
Output: Final label embedding $\mathbf{H}^{(L)}$ for each sample.

Phase 1: Constructing Attribute-aware Co-occurrence Matrices
for all $r \in \mathbb{R}$ **do**
    for all $v \in V^{(r)}$ **do**
        $D^{(r)}(v) \leftarrow \{(x, y) \in D \mid a_r = v\}$
        Initialize $\mathbf{N}_{ij}^{(r)}(v) \in \mathbb{R}^{C \times C}$
        for all $(x, y) \in D^{(r,v)}$ **do**
            for all (i, j) such that $y_i = y_j = 1$ **do**
                $\mathbf{N}_{ij}^{(r)}(v) \leftarrow \mathbf{N}_{ij}^{(r)}(v) + 1$
            end for
        end for
        $\mathbf{A}_{ij}^{(r)}(v) \leftarrow \mathbf{N}_{ij}^{(r)}(v) / \sum_j \mathbf{N}_{ij}^{(r)}(v)$
        if $\mathbf{A}_{ij}^{(r)}(v) < \tau$ **then**
            $\mathbf{A}_{ij}^{(r)}(v) \leftarrow 0$
        end if
    end for
end for

Phase 2: Performing Attribute-aware Graph Convolution
$\mathbf{H}^{(0)} \leftarrow \mathbf{E}^{final}$
for $l = 0$ to $L - 1$ **do**
    Select attribute r_l for layer l
    Determine attribute value v_l for current sample
    $\hat{\mathbf{A}}^{(r_l)}(v_l) \leftarrow$ normalized $\mathbf{A}^{(r_l)}(v_l)$
    $\mathbf{H}^{(l+1)} \leftarrow \sigma\left(\hat{\mathbf{A}}^{(r_l)}(v_l)\mathbf{H}^{(l)}\mathbf{W}^{(l)} + \mathbf{b}^{(l)}\right)$
end for
return $\mathbf{H}^{(L)}$

3.5. Theoretical Justification

To theoretically validate the proposed multi-scale semantic alignment, a geometric analysis of the discriminative structure of label embeddings is conducted, accompanied by a formal proof of the enhanced discriminative capacity achieved through multi-scale fusion.

Proposition. Let $\mathbf{E} = \{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_C\} \subset \mathbb{R}^d$ be the initial label embedding, and define the average inter-class center distance as:

$$D_{\text{inter}}(\mathbf{E}) = 2/[C(C-1)] \sum_{1 \le m < n \le C} \|\mathbf{e}_m - \mathbf{e}_n\|_2 \tag{17}$$

Denote $\mathbf{E}^{(i)}$ as the label embedding after interacting with the visual features from the i-th scale, and let $\mathbf{E}^{final}$ represent the embedding after full multi-scale fusion. Then, the inequality holds:

$$D_{\text{inter}}(\mathbf{E}^{final}) > \max_{1 \le i \le I} D_{\text{inter}}(\mathbf{E}^{(i)}) \tag{18}$$

Proof. For each scale i, the label embedding is updated via interaction with the visual features $\mathbf{T}_i$ as follows:

$$\mathbf{E}'_i = \text{Attention}([\mathbf{T}_i; \mathbf{E}_i])_{[C]} = \mathbf{E}_i + \Delta_i \tag{19}$$

where Δ_i denotes the semantic perturbation induced by the visual context T_i, with directionality governed by local features, reflecting class distinctions at the corresponding scale. Single-scale interaction introduces only a directional perturbation Δ_i, resulting in the updated inter-class distance:

$$D_{\text{inter}}(\mathbf{E}^{(i)}) = D_{\text{inter}}(\mathbf{E}_i + \Delta_i) \tag{20}$$

Since perturbations from different scales are complementary in the semantic space, the relation holds:

$$\text{Span}(\{\Delta_{final}\}) \supsetneq \text{Span}(\Delta_i), \ \forall i \tag{21}$$

Therefore, the geometric tension of the fused embedding structure exceeds that of any single scale:

$$\|\mathbf{e}_m^{final} - \mathbf{e}_n^{final}\|_2 > \|\mathbf{e}_m^{(i)} - \mathbf{e}_n^{(i)}\|_2, \ \forall m \neq n, \ \forall i \tag{22}$$

Consequently, it follows that:

$$D_{\text{inter}}(\mathbf{E}^{final}) > \max_i D_{\text{inter}}(\mathbf{E}^{(i)}) \tag{23}$$

This indicates that the multi-scale alignment mechanism theoretically leads to a label embedding structure with stronger inter-class separability.

4. Experiments

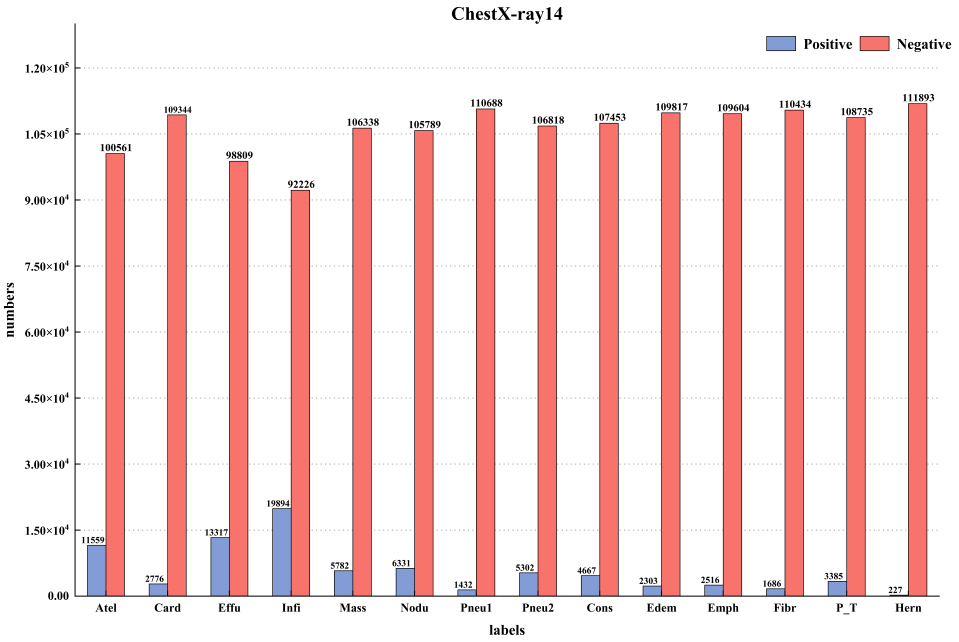
A systematic evaluation is conducted to assess the effectiveness, robustness, and generalization of the proposed MSASG algorithm. Two large-scale public datasets are used, accompanied by detailed descriptions of their annotation protocols, data distributions, and demographic profiles. To ensure reproducibility and statistical rigor, experimental configurations and evaluation metrics are clearly specified. Based on this setup, a series of analyses is performed, including algorithm analysis, comparative experiments, and ablation studies. Additionally, qualitative visualizations provide intuitive evidence supporting the feasibility of attribute representation and its effectiveness in multi-label prediction.

4.1. Datasets

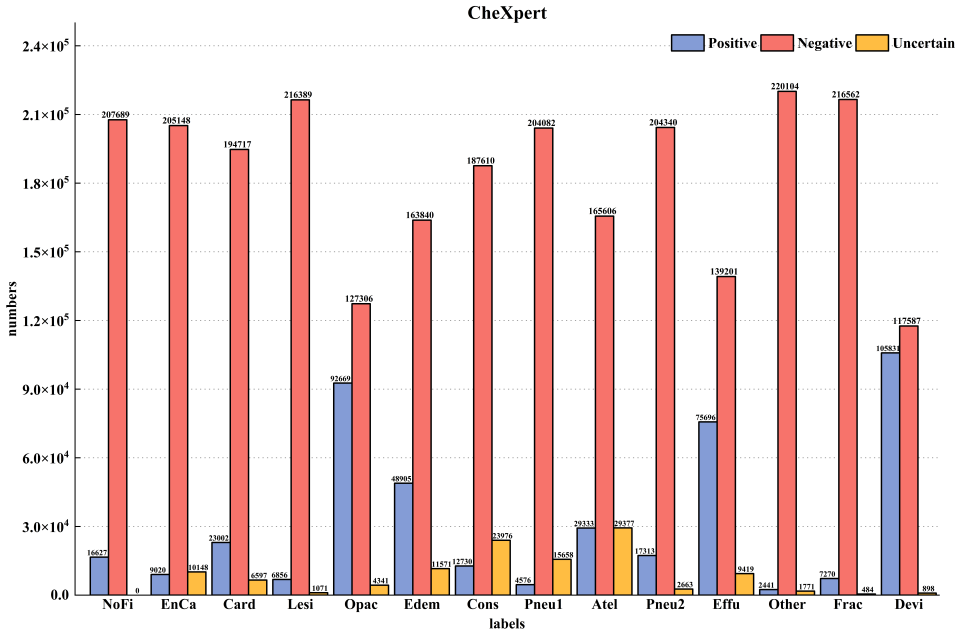
ChestX-ray14 and CheXpert are two large-scale public CXR datasets widely used in automated diagnosis. Both offer multi-label annotations and demographic attributes, enabling attribute-aware representation of disease co-occurrence patterns. As shown in Figure 3, their label distributions exhibit notable imbalance in the proportions of positive, negative, and uncertain samples, which may influence algorithm training and evaluation. To mitigate this, each dataset is randomly divided into training (70%), validation (10%), and test (20%) subsets, with consistent demographic distributions maintained across splits to ensure experimental stability and fairness.

(1) ChestX-ray14. Released by the National Institutes of Health (NIH), this dataset contains 112,120 frontal chest X-ray images from 30,805 patients, accompanied by image-level annotations for 14 common thoracic diseases. The labels were automatically extracted from radiology reports using natural language processing, achieving annotation accuracy exceeding 90%. Among these images, 51,708 present at least one abnormality, while the remainder are labeled as normal.

(2) CheXpert. Published by the Stanford University School of Medicine, this dataset contains 224,316 CXR images from 65,240 patients, annotated with 14 categories of thoracic diseases. A key distinction lies in its uncertainty-aware labeling scheme: each label takes a value of 1 (positive), 0 (negative), or -1 (uncertain), capturing diagnostic ambiguity and linguistic variability in clinical reports. Following (Irvin et al., 2019), two label conversion strategies are applied: CheXpert_1s, treating all uncertain labels as positive, and



(a) Distribution of ChestX-ray14 dataset



(b) Distribution of CheXpert dataset

Figure 3: Distribution of positive, uncertain, and negative samples for each label

CheXpert_0s, treating them as negative. This enables controlled evaluation of algorithm robustness under label uncertainty.

4.2. Experimental Settings and Evaluation Metrics

All experiments are conducted on a computing platform equipped with NVIDIA RTX GPUs using the PyTorch framework. All input images are center-cropped, resized to 224×224 pixels, and normalized based on ImageNet statistics (mean $\mu = [0.485, 0.456, 0.406]$, standard deviation $\sigma = [0.229, 0.224, 0.225]$). To ensure reproducibility and avoid bias, no data augmentation is applied during training. The AdamW (Loshchilov and Hutter, 2017) optimizer is employed for algorithm training, with an initial learning rate of 1×10^{-4} and a weight decay of 1×10^{-5}. A learning rate scheduler dynamically adjusts the learning rate during training, which is conducted over 50 epochs with a batch size of 64. The Transformer adopts 12 attention heads to capture global dependencies within the input features. In the MFPR method, three-level feature partitioning is performed. The parameters for the first, second, and third partitions are set as $(k = 7, s = 4, p = 2)$, $(k = 3, s = 2, p = 1)$, and $(k = 3, s = 2, p = 1)$, respectively, to progressively build multi-scale spatial–semantic representations.

To comprehensively evaluate the performance of the proposed MSASG algorithm on multi-label classification tasks, the Area Under the Curve (AUC) and Receiver Operating Characteristic (ROC) curve are adopted as the primary evaluation metrics, which are particularly suitable for imbalanced label distributions in medical image analysis. The ROC curve characterizes the relationship between the false positive rate (FPR) and true positive rate (TPR) across different decision thresholds, defined as:

$$TPR = TP/(TP + FN) \tag{24}$$

$$FPR = FP/(FP + TN) \tag{25}$$

where TP, FP, TN, and FN denote the number of true positive, false positive, true negative, and false negative samples, respectively. AUC represents the area under the ROC curve; a higher value (closer to 1) indicates stronger discriminative capability. AUC is computed independently for each label, and the macro-average is reported to reflect the overall algorithm performance.

4.3. Algorithm Analysis

The impact of key architectural configurations is analyzed to evaluate their influence on algorithm performance, with a particular focus on feature

partitioning, partitioning parameter combinations, age discretization strategies, and the confidence threshold τ.

4.3.1. Impact of the Number of Feature Partitioning Iterations

To evaluate the impact of iterative feature partitioning, experiments are conducted on the ChestX-ray14, CheXpert_1s, and CheXpert_0s datasets, with the number of partitioning iterations ranging from 1 to 4. As shown in Figure 4, the performance of the MSASG algorithm consistently improves as the number of partitioning iterations increases, reaching its peak with a three-iteration feature partitioning configuration, where average AUCs of 84.2%, 83.3%, and 83.8% are achieved on the respective datasets. These results confirm that multi-scale partitioning enhances semantic feature extraction and facilitates effective fusion of local and global information, which is crucial for addressing lesion size variability. However, further increasing the number of iterations leads to performance saturation or slight degradation, indicating that excessive partitioning may introduce redundancy and reduce discriminative capability. Therefore, a three-iteration feature partitioning configuration is adopted as the default to strike a balance between performance and computational complexity.

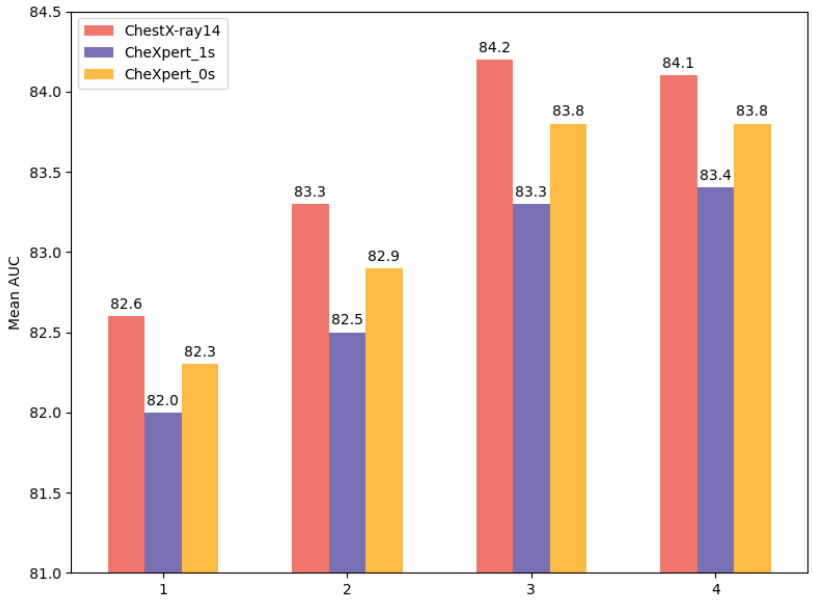


Figure 4: Mean AUC under different numbers of feature partitioning iterations

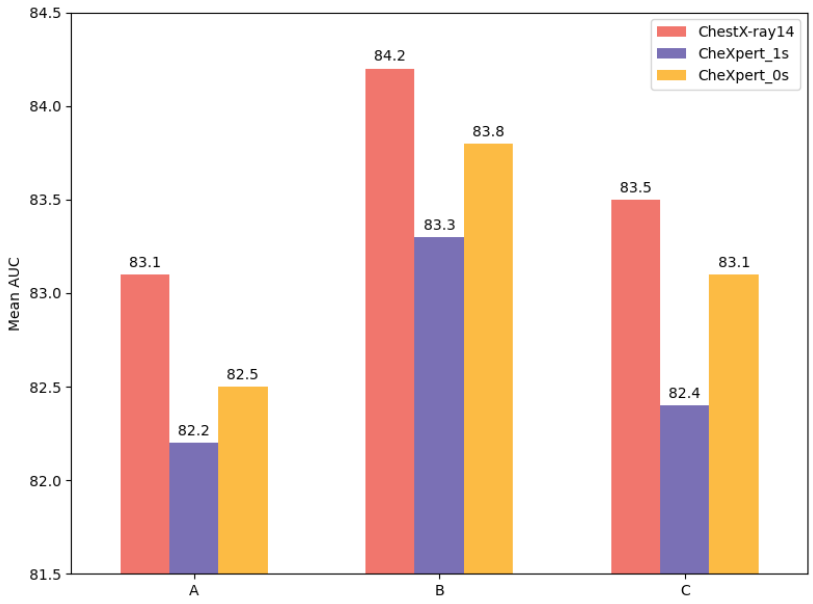


Figure 5: Mean AUC under different feature partitioning parameter configurations

4.3.2. Impact of Feature Partitioning Parameter Combinations

The MFPR method adopts a three-iteration feature partitioning configuration, where each iteration applies patch-wise operations parameterized by kernel size, stride, and padding, denoted as $(k/s/p)$. To assess the influence of different parameter choices, three representative configurations are evaluated: A ($9/5/2 \rightarrow 5/3/1 \rightarrow 3/2/1$), B ($7/4/2 \rightarrow 3/2/1 \rightarrow 3/2/1$), and C ($5/3/1 \rightarrow 3/2/1 \rightarrow 3/2/1$). Experiments on the ChestX-ray14, CheXpert_1s, and CheXpert_0s datasets present the results shown in Figure 5. Configuration B consistently achieves the highest average AUCs (84.2%, 83.3%, and 83.8%, respectively), reflecting a better balance between capturing global contextual information and preserving fine-grained lesion features. Configuration A, while benefiting from a larger receptive field in earlier iterations, shows limited ability in representing subtle details. In contrast, Configuration C, with enhanced focus on local structures, struggles to retain broader anatomical information due to its reduced spatial coverage. Overall, Configuration B provides the most favorable balance between global context awareness and local detail sensitivity, and thus serves as the default configuration.

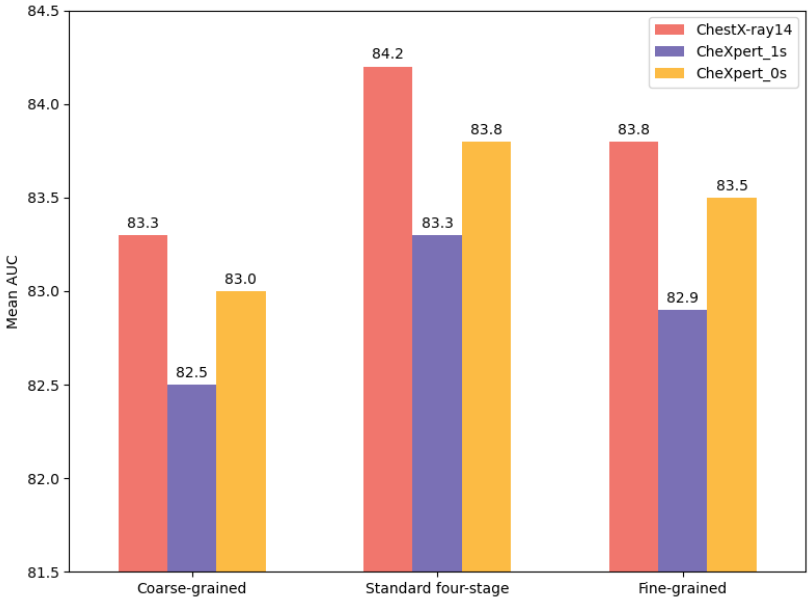


Figure 6: Mean AUC under different age discretization strategies

4.3.3. Impact of Age Discretization Strategies

To evaluate the impact of age discretization strategies on graph construction, three schemes are investigated: a coarse division with a 50-year threshold, a four-stage division (child, youth, middle-aged, elderly), and a fine-grained division using 10-year intervals. Comparative experiments are conducted on the ChestX-ray14, CheXpert_1s, and CheXpert_0s datasets, with all other graph construction components held constant. The results are presented in Figure 6. The four-stage division consistently achieves the highest average AUCs across datasets (84.2%, 83.3%, and 83.8%, respectively), indicating a favorable balance among semantic clarity, sample distribution, and structural stability. In contrast, the coarse strategy lacks age-specific representation, while the fine-grained strategy suffers from data sparsity and fragmented graph connectivity. Considering both accuracy and structural robustness, the four-stage discretization is adopted as the default age discretization strategy.

4.3.4. Impact of Confidence Threshold τ

To investigate the impact of the confidence threshold τ on attribute-aware graph construction, experiments are conducted on the ChestX-ray14, CheX-

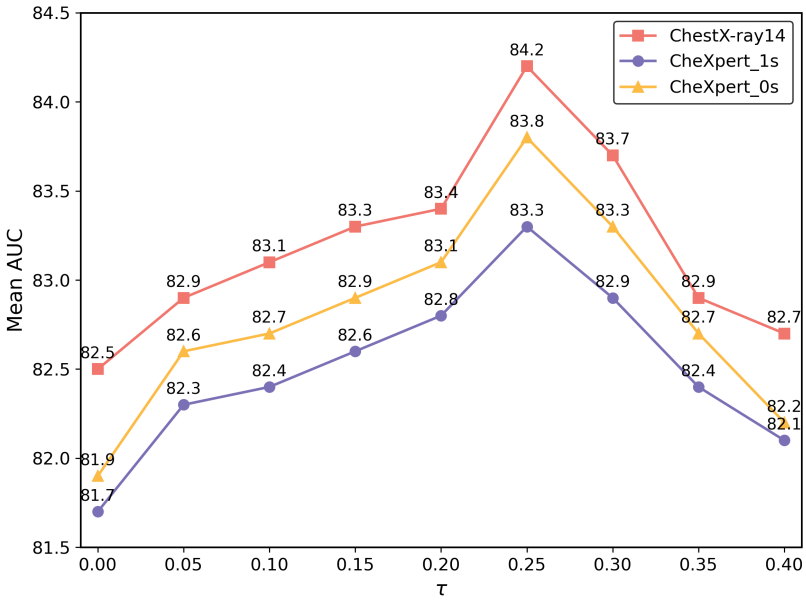


Figure 7: Mean AUC under different confidence thresholds τ

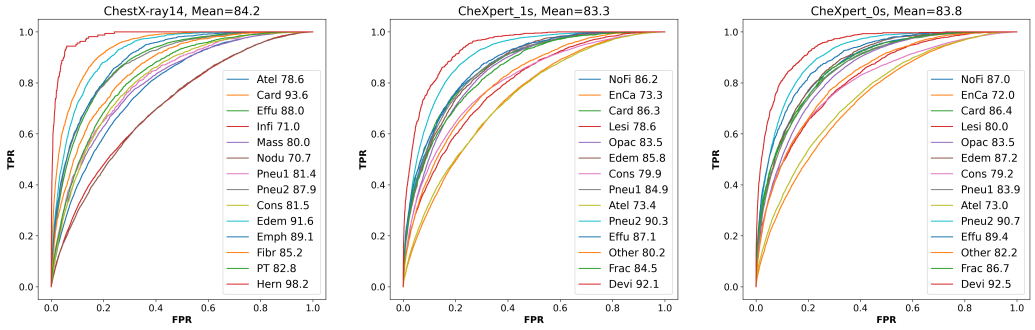
pert_1s, and CheXpert_0s datasets, with τ ranging from 0 to 0.4. As shown in Figure 7, moderate sparsification consistently enhances performance: the mean AUC increases with higher τ values and peaks around $\tau = 0.25$, achieving 84.2%, 83.3%, and 83.8% on the three datasets, respectively. This suggests that pruning low-confidence label relations enhances the discriminative capacity of the graph structure and facilitates more effective message propagation. However, when τ exceeds 0.35, a noticeable performance drop is observed, indicating that excessive sparsification may compromise semantic connectivity among labels and hinder information flow. Based on these observations, $\tau = 0.25$ is selected as the default configuration to ensure a favorable balance between structural clarity and information retention.

Based on the above experimental results, three-iteration feature partitioning (7/4/2→3/2/1→3/2/1) and the standard four-stage age discretization consistently yield the best performance across all datasets, confirming their effectiveness in balancing representational capacity and structural robustness. To further evaluate overall discriminative performance across datasets, Figure 8 illustrates ROC curves for different disease categories on the ChestX-ray14, CheXpert_1s, and CheXpert_0s datasets using the default configura-

tion.

4.4. Comparative Experiments

To comprehensively evaluate the effectiveness and generalization capability of the proposed MSASG algorithm for multi-label classification, extensive comparative experiments are conducted against representative state-of-the-art algorithms. The baseline algorithms span three major architectural paradigms in current research: CNNs, Transformer-based methods, and GCNs, thereby covering the dominant methodological families commonly adopted in multi-label CXR image classification.



(a) ChestX-ray14 (b) CheXpert_1s (c) CheXpert_0s

Figure 8: ROC curves and AUC for each disease label on all datasets

(1) Evaluation on ChestX-ray14. As shown in Table 2, the proposed MSASG algorithm achieves the highest average AUC of 84.2% across 14 thoracic disease categories on the ChestX-ray14 dataset, consistently outperforming all baseline methods. Compared with the CNN-based CheXNet (80.7%), MSASG shows a 3.5 percentage point gain, highlighting its strength in capturing complex radiographic patterns. Against Transformer-based methods like LT-ViT (82.8%), MSASG further improves global context understanding by combining multi-scale feature partitioning with semantic alignment, enabling more accurate lesion localization and classification. For graph-based methods, MSASG outperforms CRAL (81.6%), CheXGCN (82.6%), and ImageGCN (83.2%), with AUC gains of 2.6%, 1.6%, and 1.0%, respectively. These results confirm the advantage of attribute-guided graph construction in capturing semantic dependencies and adapting to disease-specific topologies. At the category level, MSASG achieves notably higher AUCs for diseases with overlapping semantics and close spatial distribution, such as Effusion

(88.0%), Consolidation (81.5%), and Cardiomegaly (93.6%), consistently outperforming all baselines. In summary, the results highlight the reliability and effectiveness of MSASG for automated thoracic disease diagnosis.

(2) Evaluation on CheXpert. To evaluate the effectiveness of MSASG under label uncertainty, experiments are conducted on the CheXpert dataset using two widely recognized uncertainty-handling protocols: CheXpert_1s and CheXpert_0s. As reported in Tables 3 and 4, MSASG achieves the highest overall AUCs under both configurations, reaching 83.3% and 83.8%, respectively, consistently outperforming all baseline methods. Notably, the CheXpert_0s configuration demonstrates superior performance. These findings highlight the strong adaptability and generalization capability of MSASG in clinically realistic scenarios characterized by uncertain or noisy supervision.

Table 2: AUC for each label on ChestX-ray14(%)

Algorithm	Atel	Card	Effu	Infi	Mass	Nodu	P1	P2	Cons	Edem	Emph	Fibr	PT	Hern	**Mean**
U-DCNN (Wang et al., 2017)	71.6	80.7	78.4	60.9	70.6	67.1	63.3	80.6	70.8	83.5	81.5	76.9	70.8	76.7	73.8
CheXNet (Rajpurkar et al., 2017)	76.9	88.5	82.5	69.4	82.4	75.9	71.5	85.2	74.5	84.2	90.6	82.1	76.6	90.1	80.7
CRAL (Guan and Huang, 2020)	78.1	88.3	83.1	69.7	83.0	76.4	72.5	86.6	75.8	85.3	91.1	82.6	78.0	91.8	81.6
GWSA-LCD (Xu et al., 2024)	77.0	87.7	82.7	70.1	82.1	79.0	73.2	87.0	74.6	84.7	92.4	83.9	78.2	92.1	81.8
PCAN (Zhu et al., 2022)	78.5	89.9	83.7	70.6	83.4	78.6	73.0	87.1	76.3	85.4	92.1	81.7	79.1	94.3	82.4
LLAGnet (Chen et al., 2019a)	78.3	88.5	83.4	70.3	84.1	79.0	72.9	87.7	75.4	85.1	93.9	83.2	79.8	91.6	82.4
CheXGAT (Lee et al., 2022)	78.6	87.9	83.7	69.9	83.9	79.3	74.1	87.9	75.4	85.1	**94.4**	84.2	79.4	93.1	82.6
CheXGCN (Chen et al., 2020)	78.6	89.3	83.2	69.9	84.0	80.0	73.9	87.6	75.1	85.0	94.4	83.4	79.5	92.9	82.6
LT-ViT (Marikkar et al., 2023)	**80.7**	90.6	85.3	**72.3**	81.1	73.4	74.1	85.6	81.2	88.4	93.0	82.9	78.8	92.8	82.8
TransDD (Jiang et al., 2024)	79.1	88.5	84.2	71.5	83.7	**80.3**	74.5	88.5	75.3	85.9	94.4	84.9	80.3	92.4	83.1
ImageGCN (Mao et al., 2022)	80.2	89.4	87.4	70.2	**84.3**	76.8	71.5	**90.0**	79.6	88.3	91.5	82.5	79.1	94.3	83.2
Ours	78.6	**93.6**	**88.0**	71.0	80.0	70.7	**81.4**	87.9	**81.5**	**91.6**	89.1	**85.2**	**82.8**	**98.2**	**84.2**

Table 3: AUC for each label on CheXpert under the CheXpert_1s setting(%)

Algorithm	NoFi	EnCa	Card	Lesi	Opac	Edem	Cons	P1	Atel	P2	Effu	Other	Frac	Devi	**Mean**
U_Ones (Irvin et al., 2019)	87.5	67.6	87.3	76.4	79.5	88.0	73.5	79.4	72.2	89.8	90.1	80.5	79.1	89.6	81.5
CheXNet (Rajpurkar et al., 2017)	87.6	67.5	87.2	76.5	79.6	87.9	73.6	79.5	72.3	89.7	90.2	80.6	79.5	89.5	81.6
PCAN (Zhu et al., 2022)	88.0	68.0	86.8	76.5	80.2	87.4	73.4	79.9	72.2	89.1	89.5	81.0	79.6	89.7	81.7
GWSA-LCD (Xu et al., 2024)	88.0	68.0	86.9	76.6	80.0	88.0	73.7	79.7	72.3	89.1	89.6	81.2	80.1	89.8	81.7
LT-ViT (Marikkar et al., 2023)	87.9	67.9	86.9	76.8	79.9	88.6	73.9	79.6	72.3	89.1	89.7	81.3	80.8	89.9	81.9
CRAL (Guan and Huang, 2020)	87.9	67.9	86.5	77.1	80.3	88.2	73.7	79.7	73.2	89.7	89.8	81.3	80.2	90.3	82.2
CheXGAT (Lee et al., 2022)	**88.2**	68.1	86.7	76.9	79.5	87.9	74.1	80.3	73.2	89.0	**90.3**	81.1	80.0	88.7	82.2
TransDD (Jiang et al., 2024)	87.9	68.2	86.6	77.1	79.8	88.2	73.8	80.2	73.0	89.1	90.2	80.8	79.8	88.9	82.3
ImageGCN (Mao et al., 2022)	88.1	**68.4**	86.6	**77.3**	80.2	**88.6**	73.5	80.1	72.9	89.2	90.2	80.6	79.6	89.2	82.4
LLAGnet (Chen et al., 2019a)	87.7	68.1	87.2	76.5	82.1	87.8	74.3	80.2	73.1	**91.0**	90.1	82.1	81.1	89.9	82.4
CheXGCN (Chen et al., 2020)	87.9	68.2	**87.6**	76.8	82.1	88.4	74.5	80.5	73.1	91.3	90.6	**82.3**	**84.2**	90.2	82.7
Ours	86.2	73.3	86.3	78.6	**83.5**	85.8	**79.9**	**84.9**	**73.4**	90.3	87.1	80.2	84.5	**92.1**	**83.3**

Table 4: AUC for each label on CheXpert under the CheXpert_0s setting(%)

Algorithm	NoFi	EnCa	Card	Lesi	Opac	Edem	Cons	P1	Atel	P2	Effu	Other	Frac	Devi	**Mean**
PCAN (Zhu et al., 2022)	87.7	69.5	87.5	75.3	80.6	88.6	77.7	80.6	73.6	90.4	90.4	81.5	81.7	89.2	81.9
CRAL (Guan and Huang, 2020)	**88.4**	68.4	88.4	76.4	80.9	88.7	77.3	80.8	73.6	90.4	90.5	80.5	82.7	89.2	82.1
GWSA-LCD (Xu et al., 2024)	88.2	68.5	88.1	76.4	81.0	88.5	78.3	80.7	72.7	90.7	89.8	81.4	82.6	89.2	82.1
LT-ViT (Marikkar et al., 2023)	88.0	68.7	87.8	76.5	81.2	88.4	77.4	80.6	73.8	91.0	89.1	82.4	82.6	89.3	82.2
U_Zeros (Irvin et al., 2019)	87.7	69.1	87.6	76.6	80.7	88.1	77.5	80.4	73.0	90.2	90.2	81.5	80.7	89.3	82.3
CheXGAT (Lee et al., 2022)	87.8	69.8	87.8	76.7	80.7	88.1	78.2	81.1	73.2	90.4	**91.2**	82.0	80.8	90.1	82.4
ImageGCN (Mao et al., 2022)	88.2	**70.2**	**88.9**	75.1	81.3	**89.0**	78.0	80.9	73.2	91.0	91.4	80.7	81.6	89.3	82.6
CheXNet (Mao et al., 2022)	87.9	69.4	87.5	76.4	81.2	88.0	77.5	80.3	73.3	91.1	90.3	81.9	82.7	89.8	82.7
TransDD (Jiang et al., 2024)	87.8	69.4	87.6	76.3	81.6	88.2	77.8	81.8	72.2	91.3	90.6	82.5	82.9	89.7	82.9
LLAGnet (Chen et al., 2019a)	87.6	69.4	87.8	76.3	82.0	88.4	78.2	81.2	73.1	91.5	90.9	83.1	83.1	89.6	83.0
CheXGCN (Chen et al., 2020)	87.9	69.7	87.7	76.8	82.2	88.6	78.4	81.0	73.6	**91.7**	90.7	**83.5**	83.3	89.9	83.2
Ours	87.0	72.0	86.4	**80.0**	**83.5**	87.2	**79.2**	**83.9**	**73.0**	90.7	89.4	82.2	**86.7**	**92.5**	**83.8**

4.5. Ablation Study

To systematically assess the effectiveness of the key components and strategies within the MSASG algorithm, two types of ablation experiments are conducted on the ChestX-ray14 dataset. The first type investigates the individual contribution of each method, aiming to quantify its necessity and functional impact within the overall architecture. The second type focuses on the influence of different attribute categories on the construction of the label co-occurrence, thereby exploring the role of attribute-specific information in shaping label dependency structures. To ensure the rigor and comparability of the evaluation, all experiments adopt the mean AUC as the primary evaluation metric.

4.5.1. Analysis of Key Components

To systematically evaluate the individual and combined contributions of the key components in the MSASG algorithm, a series of ablation studies are conducted. Specifically, four experimental variants are designed by selectively disabling the MFPR method, the LMSA method, and A-GCN method. The experiment are assessed on the ChestX-ray14 dataset, and the results are summarized in Table 5. The complete MSASG algorithm yields the highest mean AUC of 84.2%, demonstrating the benefit of integrating all three method. The exclusion of MFPR results in the most significant performance drop, with AUC decreasing to 82.6%, underscoring its essential role in capturing lesion structure and promoting spatial generalization through multi-scale fusion. When A-GCN is removed, the AUC drops to 83.0%, indicating that structured semantic representation is vital for modeling inter-label dependencies under complex label distributions. Similarly,

the omission of LMSA results in a moderate decline to 83.6%, suggesting that semantic alignment enhances attention to discriminative pathological regions and supports semantic consistency. In summary, MFPR, LMSA, and A-GCN jointly enhance MSASG performance by addressing spatial representation, semantic alignment, and label relationship learning.

Table 5: AUC for each label under different key component configurations(%)

	Atel	Card	Effu	Infi	Mass	Nodu	Pneu1	Pneu2	Cons	Edem	Emph	Fibr	PT	Hern	Mean
w/o MFPR	78.5	88.3	83.5	70.1	78.1	70.0	80.2	85.6	80.3	90.5	89.0	84.6	81.7	96.1	82.6
w/o A-GCN	78.3	88.6	84.4	**71.2**	79.9	70.7	80.4	86.3	80.7	90.1	88.5	84.7	82.1	96.7	83.0
w/o LMSA	**78.7**	92.3	85.9	70.5	79.4	70.6	81.2	86.7	81.1	91.3	88.7	84.4	82.3	97.3	83.6
MSASG	78.6	**93.6**	**88.0**	71.0	**80.0**	**70.7**	**81.4**	**87.9**	**81.5**	**91.6**	**89.1**	**85.2**	**82.8**	**98.2**	**84.2**

4.5.2. Analysis of Attribute-aware Label Dependency Construction

To assess the influence of attribute information on multi-label CXR image classification, label co-occurrence matrices are constructed under four configurations: (1) No Relation, (2) Age-based, (3) Gender-based, and (4) Age + Gender. These configurations aim to reveal how demographic attributes independently and jointly influence the construction of the label dependency graph. As shown in Table 6, the baseline without attribute integration achieves a mean AUC of 83.0%. When age information is incorporated, the AUC increases to 83.5%, suggesting that age reflects disease progression trends across patient populations and serves as a valuable prior. Similarly, when gender information is used, the AUC reaches 83.7%, indicating that physiological differences between genders affect label dependencies and semantic co-occurrence patterns. The highest AUC of 84.2% is obtained when both age and gender are integrated, demonstrating that combining multiple demographic cues enables more expressive and personalized label relationships, thereby enhancing both structural representation and classification performance. These results underscore the value of attribute-aware graph construction in capturing latent inter-label correlations. In particular, the fusion of age and gender offers complementary semantic information, thereby enhancing the generalization of algorithm and enabling better adaptation to patient-specific variability in real-world clinical practice.

4.6. Discussion

This section discusses several critical factors that affect the performance, reliability, and practical applicability of the proposed algorithm, including la-

Table 6: AUC for each label under different attribute-aware configurations(%)

	Atel	Card	Effu	Infi	Mass	Nodu	Pneu1	Pneu2	Cons	Edem	Emph	Fibr	PT	Hern	Mean
No Relation	78.3	88.6	84.4	**71.2**	79.9	70.7	80.4	86.3	80.7	90.1	88.5	84.7	82.1	96.7	83.0
Age	78.4	91.1	85.6	71.0	79.8	70.8	80.7	87.2	80.7	91.1	88.9	85.1	82.3	97.3	83.5
Sex	78.5	92.2	86.3	71.1	79.9	**70.9**	81.1	87.1	80.9	90.4	89.0	84.8	82.2	97.5	83.7
Age+Sex	**78.6**	**93.6**	**88.0**	71.0	**80.0**	70.7	**81.4**	**87.9**	**81.5**	**91.6**	**89.1**	**85.2**	**82.8**	**98.2**	**84.2**

bel embeddings, loss functions, computational efficiency, attribute flexibility, and robustness.

4.6.1. Discussion on Label Embedding

To evaluate the sensitivity of the algorithm to the choice of initial label embeddings, general-domain GloVe embeddings are compared with domain-specific BioBERT embeddings. As shown in Table 7, both options yield highly comparable AUC scores, with most label-wise differences within 0.3%. The largest observed gap is 0.8%, favoring GloVe on the "Edema" label. This consistency suggests the specific choice of embedding initialization has limited impact on overall performance. The algorithm demonstrates a strong capacity to refine semantic representations through multi-scale alignment, regardless of the initial embedding space. Given its simplicity and lower computational overhead, GloVe is adopted as the default label embedding throughout the experiments.

Table 7: AUC for each label with GloVe and BioBERT embeddings(%)

	Atel	Card	Effu	Infi	Mass	Nodu	Pneu1	Pneu2	Cons	Edem	Emph	Fibr	PT	Hern	Mean
BioBERT	78.4	**93.8**	**88.1**	71.0	79.6	70.4	**81.7**	**88.1**	**81.6**	90.8	89.1	84.7	**82.9**	97.7	84.1
GloVe	**78.6**	93.6	88.0	**71.0**	**80.0**	**70.7**	81.4	87.9	81.5	**91.6**	**89.1**	**85.2**	82.8	**98.2**	**84.2**

4.6.2. Discussion on Loss Functions

To investigate the impact of loss function selection on multi-label medical image classification, four widely used loss functions are evaluated on the ChestX-ray14 dataset: Binary Cross-Entropy (BCE), Class-aware Weighted BCE (CBCE), Focal Loss (FL), and Asymmetric Loss (ASL). As shown in Table 8, the choice of loss function notably affects performance, particularly under class imbalance and label sparsity. CBCE achieves the highest overall AUC by applying class-specific weights that mitigate the influence of frequent labels. FL enhances sensitivity to minority and hard samples but introduces

higher variability across disease categories. ASL slightly improves predictions for rare classes by reducing negative sample interference, though with less consistent performance. Among all candidates, CBCE demonstrates the most robust and balanced results, and is thus adopted as the default loss function.

Table 8: AUC for each label under different loss functions(%)

	Atel	Card	Effu	Infi	Mass	Nodu	Pneu1	Pneu2	Cons	Edem	Emph	Fibr	PT	Hern	Mean
BCE	78.3	91.8	87.5	70.7	79.6	70.3	80.3	87.1	80.8	91.1	87.9	84.3	81.5	96.2	83.3
FL	78.6	92.2	87.6	70.7	79.7	70.4	80.9	87.2	81.4	91.3	88.0	84.6	81.7	97.9	83.7
ASL	**78.7**	92.8	87.7	70.8	79.9	**70.9**	**81.5**	87.1	81.2	91.4	88.0	84.7	81.9	**98.3**	83.9
CBCE	78.6	**93.6**	**88.0**	**71.0**	**80.0**	70.7	81.4	**87.9**	**81.5**	**91.6**	**89.1**	**85.2**	**82.8**	98.2	**84.2**

4.6.3. Discussion on Computational Efficiency

The complexity of an algorithm is a crucial factor when considering its integration into clinical workflows. In this paper, computational efficiency is evaluated from three perspectives: (1) The number of trainable parameters; (2) Training efficiency, measured by average training time per image; and (3) Testing efficiency, measured by average testing time per image. All evaluations are conducted on the ChestX-ray14 dataset under a consistent computational environment to ensure fairness. As summarized in Table 9, the proposed algorithm contains the fewest parameters (251M) and exhibits the fastest training time (45.27 ms/image), while maintaining competitive testing speed (40.90 ms/image). Compared to existing algorithms such as U-DCNN, CheXNet, and LLAGnet, the algorithm demonstrates a favorable trade-off between efficiency and predictive performance, highlighting its suitability for time-sensitive clinical applications.

Table 9: Comparison of parameters, training and testing time

Algorithm	Parameters (M)	Training Time (ms)	Testing Time (ms)
U-DCNN (Wang et al., 2017)	298	56.40	43.40
CheXNet (Rajpurkar et al., 2017)	341	57.40	45.93
LLAGnet (Chen et al., 2019a)	370	71.86	41.71
Ours	251	45.27	40.90

4.6.4. Discussion on Additional Attributes

Although only age and gender are utilized as attribute dimensions in the current implementation, the A-GCN method is inherently flexible and readily extendable to incorporate additional clinical information when available. Attributes such as smoking status or medication history can be discretized or embedded and subsequently integrated into the attribute space $\mathcal{R} = \{r_1, r_2, \ldots, r_m\}$. Both the construction of attribute-conditioned label co-occurrence matrices and the dynamic graph selection process can be naturally adapted to accommodate these extended attributes. This design flexibility enables the algorithm to better capture the complexity of real-world clinical configurations, where richer patient metadata are often accessible. Incorporating a broader set of attributes may further enhance the modeling of label dependencies and improve overall predictive performance.

4.6.5. Discussion on Algorithm robustness

In clinical applications, chest X-ray images are frequently compromised by quality issues such as noise, artifacts, and resolution inconsistencies. To evaluate the robustness of the proposed algorithm under such conditions, three representative perturbations are applied: Gaussian noise simulates acquisition-related interference; Angle deviation reflects geometric misalignments typically caused by improper device angle or patient positioning; and Block occlusion artifacts emulate localized visual obstructions resulting from external objects. All experiments are conducted on the ChestX-ray14 dataset under consistent configurations. Figure 9 illustrates visual examples of each perturbation. As shown in Table 10, although performance slightly degrades under these adverse conditions, the overall results remain stable, indicating that the algorithm maintains strong robustness and is suitable for clinical deployment. Furthermore, to assess sensitivity to resolution variation, the algorithm is evaluated on inputs downsampled to a range of resolutions (128–512 pixels). As illustrated in Figure 10, the AUC progressively improves with increasing resolution and reaches a plateau beyond 224 pixels, reflecting high adaptability to diverse imaging standards.

Table 10: Mean AUC under different perturbations(%)

Perturbations	Original image	Gaussian noise	Angle deviation	Block artifacts
AUC	84.2	84.0	83.7	83.4

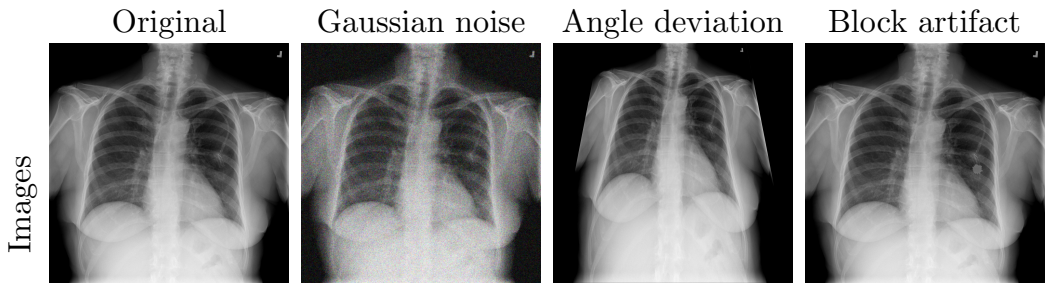


Figure 9: Visual example of image perturbations

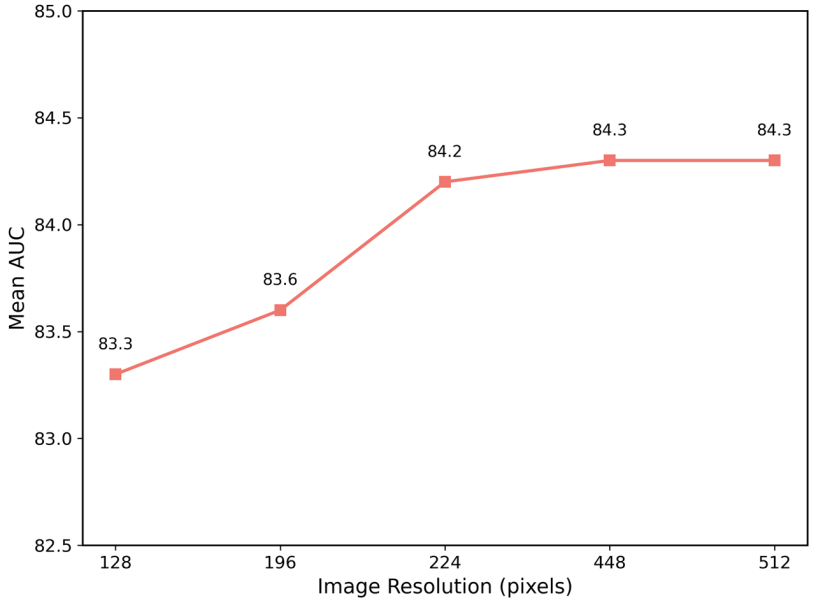


Figure 10: Mean AUC under different image resolutions

4.7. Qualitative Results and Visualization Analysis

Qualitative evaluations are conducted to assess the effectiveness of MSASG. At the structural level, Figures 13 and 14 illustrate that demographic factors such as gender and age significantly influence label co-occurrence graphs. This confirms the capability of the attribute-aware graph construction mechanism to capture clinically meaningful, group-sensitive disease relationships. To further support the theoretical justification presented in Section 3.5, empirical analysis is conducted on the label embedding space. Specifically, the average inter-class center distance D_{inter} is calculated across multiple scales and the fused representation. As shown in Table 11, the fused embedding consistently achieves the largest D_{inter} (1.67), validating inequality (19) and confirming that multi-scale interaction enhances class separability.

Table 11: Inter-class Distance at Different Scales

	$E^{(1)}$	$E^{(2)}$	$E^{(3)}$	E^{final}
D_{inter}	1.22	1.28	1.19	1.67

Beyond structural validation, image-based prediction results further demonstrate the utility of the proposed algorithm. As shown in Figure 11, predicted labels from MSASG align closely with pathological regions, maintaining high semantic and spatial consistency with radiological abnormalities. Additionally, Figure 12 provides qualitative comparisons between MSASG and its ablated variant (MFPR+LMSA), which excludes the A-GCN component. Heatmaps generated by MSASG display more accurate and concentrated localization of disease regions, better aligning with ground-truth annotations. These visualizations highlight the discriminative power of MSASG and its strength in learning structure-aware and clinically interpretable representations. These results confirm the effectiveness of MSASG in capturing attribute-aware label dependencies and accurately localizing disease regions, highlighting its practical value for multi-label medical image analysis.

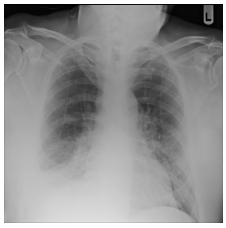
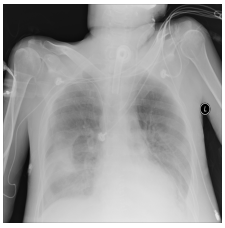
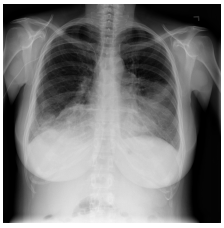
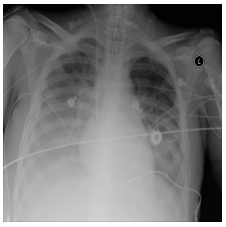
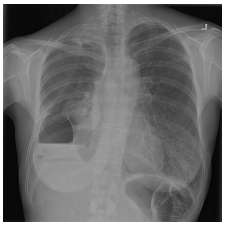
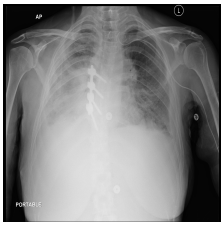
Sample				
Prediction	Effusion: 0.923 Nodule: 0.811 Atelectasis: 0.447 Infiltration: 0.341 Mass: 0.337	Infiltration: 0.854 Effusion: 0.817 Atelectasis: 0.388 Pneumothorax: 0.256 Edema: 0.248	Effusion: 0.925 Infiltration: 0.921 Pneumothorax: 0.489 Atelectasis: 0.331 Consolidation: 0.318	Mass: 0.852 Pneumothorax: 0.812 Effusion: 0.515 Atelectasis: 0.413 Infiltration: 0.393
Sample				
Prediction	Effusion: 0.914 Atelectasis: 0.847 Infiltration: 0.811 Nodule: 0.711 Emphysema: 0.462	Effusion: 0.875 Infiltration: 0.755 Consolidation: 0.492 Atelectasis: 0.360 Edema: 0.342	Effusion: 0.835 Mass: 0.817 Pleural Thickening: 0.758 Pneumothorax: 0.500 Infiltration: 0.306	Effusion: 0.835 Fibrosis: 0.723 Infiltration: 0.365 Atelectasis: 0.322 Nodule: 0.271

Figure 11: Top-5 predicted labels with probabilities for CXR images (Ground-truth labels are shown in red).

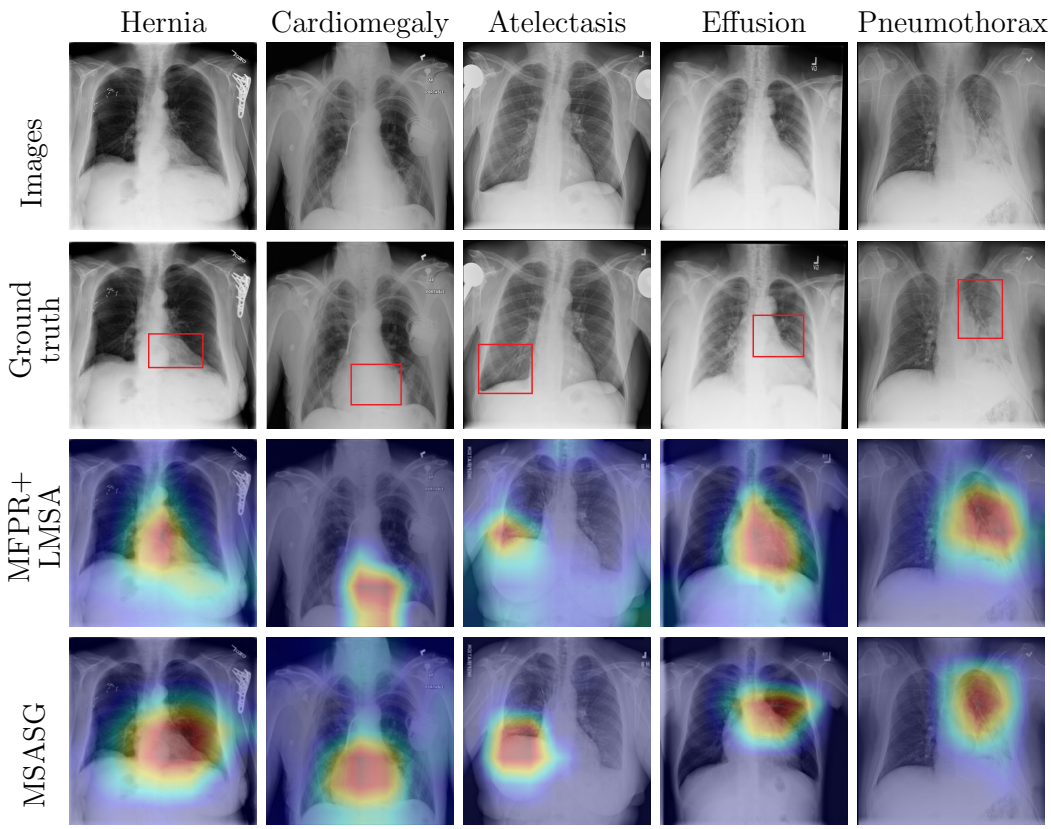
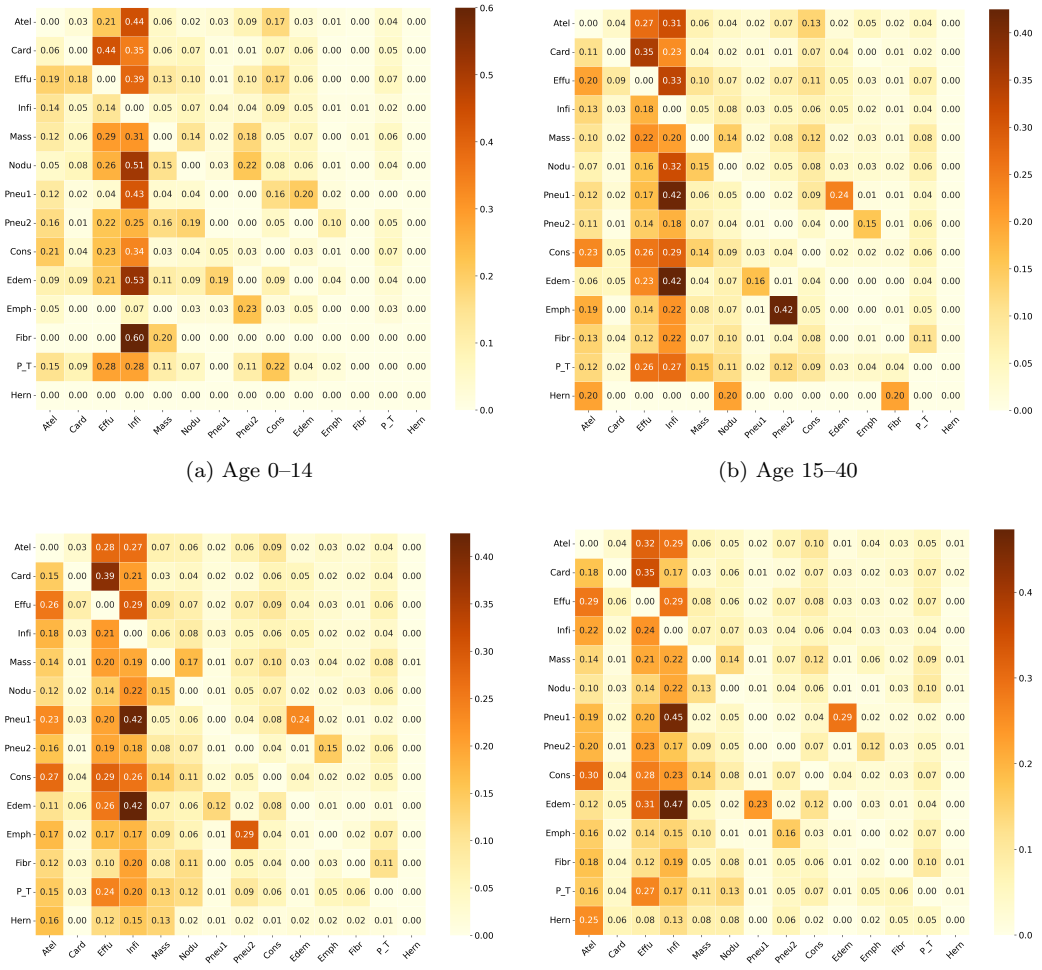


Figure 12: Illustration of lesion locations obtained by the Grad-CAM method



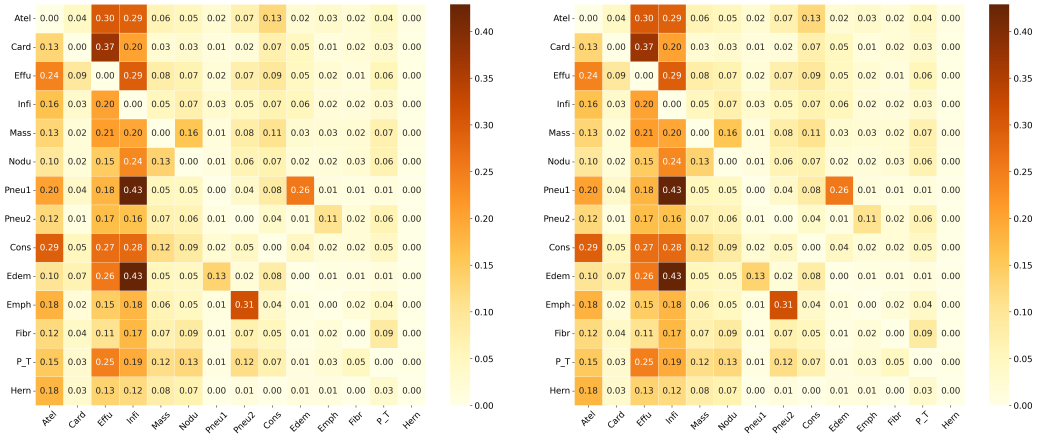
(a) Age 0–14

(b) Age 15–40

(c) Age 41–65

(d) Age>65

Figure 13: Label co-occurrence under different age conditions



(a) Female

(b) Male

Figure 14: Label co-occurrence under different gender conditions

5. Conclusion and Future Work

To address the challenges of lesion scale heterogeneity and the limited generalizability of label dependency representations in multi-label CXR classification, this paper proposes a novel algorithm based on multi-scale and attribute-aware semantic graph, where MFPR, LMSA, and A-GCN are collaboratively employed to extract multi-scale visual features and construct attribute-conditioned label dependency. To evaluate its effectiveness, extensive experiments are conducted on two widely used chest X-ray datasets: ChestX-ray14 and CheXpert. Compared with existing state-of-the-art algorithms, MSASG consistently achieves higher AUC, demonstrating robust performance in multi-label CXR image classification. Ablation studies further confirm that each component of the algorithm contributes significantly to overall performance. Visualizations of label co-occurrence matrices under different attribute configurations highlight the importance of incorporating attribute information. Notably, the highest AUC is achieved when gender and age are jointly integrated. Finally, the prediction results provide strong evidence that the proposed algorithm achieves accurate identification of disease patterns in clinical practice.

Future work will explore deeper integration of personalized attributes and assess the generalizability of the algorithm to other medical imaging modalities. These efforts aim to support its deployment in intelligent clinical decision systems and enhance its utility in real-world healthcare scenarios.

Acknowledgments

This work is supported by the National Natural Science Foundation of China under Grant Nos. 62376240 and 62302428, the S&T Program of Hebei under Grant No. 236Z0304G, the Hebei Natural Science Foundation under Grant No. F2024203085, and the Science Research Project of Hebei Education Department under Grant No. BJ2025012. The authors are grateful to valuable comments and suggestions of the reviewers.

References

Chen, B., Li, J., Lu, G., Yu, H., and Zhang, D. (2020). Label co-occurrence learning with graph convolutional networks for multi-label chest x-ray image classification. *IEEE journal of biomedical and health informatics*, 24(8):2292–2302. https://doi.org/10.1109/JBHI.2020.2967084.

Chen, B., Li, J., Lu, G., and Zhang, D. (2019a). Lesion location attention guided network for multi-label thoracic disease classification in chest x-rays. *IEEE journal of biomedical and health informatics*, 24(7):2016–2027. https://doi.org/10.1109/JBHI.2019.2952597.

Chen, B., Zhang, Z., Li, Y., Lu, G., and Zhang, D. (2021a). Multi-label chest x-ray image classification via semantic similarity graph embedding. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(4):2455–2468. https://doi.org/10.1109/TCSVT.2021.3079900.

Chen, H., Miao, S., Xu, D., Hager, G. D., and Harrison, A. P. (2019b). Deep hierarchical multi-label classification of chest x-ray images. In *International conference on medical imaging with deep learning*, pages 109–120. PMLR. https://doi.org/10.48550/arXiv.1907.07849.

Chen, Z.-M., Wei, X.-S., Wang, P., and Guo, Y. (2021b). Learning graph convolutional networks for multi-label recognition and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(6):6969–6983. https://doi.org/10.1109/TPAMI.2021.3063496.

Chowdary, G. J. and Kanhangad, V. (2022). A dual-branch network for diagnosis of thorax diseases from chest x-rays. *IEEE Journal of Biomedical and Health Informatics*, 26(12):6081–6092. https://doi.org/10.1109/JBHI.2022.3215694.

Ding, J., Du, Y., Li, W., Pei, L., and Cui, N. (2025a). Lg-diff: Learning to follow local class-regional guidance for nearshore image cross-modality high-quality translation. *Information Fusion*, 117:102870. https://doi.org/10.1016/j.inffus.2024.102870.

Ding, J., Ye, C., Wang, H., Huyan, J., Yang, M., and Li, W. (2023). Foreign bodies detector based on detr for high-resolution x-ray images of textiles. *IEEE Transactions on Instrumentation and Measurement*, 72:1–10. https://doi.org/10.1109/TIM.2023.3246510.

Ding, J., Zhao, Y., Pei, L., Shan, Y., Du, Y., and Li, W. (2025b). Modal-invariant progressive representation for multimodal image registration. *Information Fusion*, 117:102903. https://doi.org/10.1016/j.inffus.2024.102903.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*. https://doi.org/10.48550/arXiv.2010.11929.

Guan, Q. and Huang, Y. (2020). Multi-label chest x-ray image classification via category-wise residual attention learning. *Pattern Recognition Letters*, 130:259–266. https://doi.org/10.1016/j.patrec.2018.10.027.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778. https://doi.org/10.1109/CVPR.2016.90.

Hu, Y., Fang, X., Kang, P., Chen, Y., Fang, Y., and Xie, S. (2023). Dual noise elimination and dynamic label correlation guided partial multi-label learning. *IEEE Transactions on Multimedia*, 26:5641–5656. https://doi.org/10.1109/TMM.2023.3338080.

Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. (2017). Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708. https://doi.org/10.1109/CVPR.2017.243.

Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K., et al. (2019). Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 590–597. https://doi.org/10.1609/aaai.v33i01.3301590.

Jiang, X., Zhu, Y., Liu, Y., Cai, G., and Fang, H. (2024). Transdd: A transformer-based dual-path decoder for improving the performance of thoracic diseases classification using chest x-ray. *Biomedical Signal Processing and Control*, 91:105937. https://doi.org/10.1016/j.bspc.2023.105937.

Jin, Y., Lu, H., Zhu, W., and Huo, W. (2023). Deep learning based classification of multi-label chest x-ray images via dual-weighted metric loss. *Computers in biology and medicine*, 157:106683. https://doi.org/10.1016/j.compbiomed.2023.106683.

Kamal, U., Zunaed, M., Nizam, N. B., and Hasan, T. (2022). Anatomy-xnet: An anatomy aware convolutional neural network for thoracic disease classification in chest x-rays. *IEEE Journal of Biomedical and Health Informatics*, 26(11):5518–5528. https://doi.org/10.1109/JBHI.2022.3199594.

Kipf, T. N. and Welling, M. (2016). Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*. https://doi.org/10.48550/arXiv.1609.02907.

Lee, Y.-W., Huang, S.-K., and Chang, R.-F. (2022). Chexgat: A disease correlation-aware network for thorax disease diagnosis from chest x-ray images. *Artificial Intelligence in Medicine*, 132:102382. https://doi.org/10.1016/j.artmed.2022.102382.

Lian, J., Liu, J., Zhang, S., Gao, K., Liu, X., Zhang, D., and Yu, Y. (2021). A structure-aware relation network for thoracic diseases detection and segmentation. *IEEE Transactions on Medical Imaging*, 40(8):2042–2052. https://doi.org/10.1109/TMI.2021.3070847.

Loshchilov, I. and Hutter, F. (2017). Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*. https://doi.org/10.48550/arXiv.1711.05101.

Mao, C., Yao, L., and Luo, Y. (2022). Imagegcn: Multi-relational image graph convolutional networks for disease identification with chest x-rays. *IEEE transactions on medical imaging*, 41(8):1990–2003. https://doi.org/10.1109/TMI.2022.3153322.

Marikkar, U., Atito, S., Awais, M., and Mahdi, A. (2023). Lt-vit: A vision transformer for multi-label chest x-ray classification. In *2023 IEEE International Conference on Image Processing (ICIP)*, pages 2565–2569. IEEE. https://doi.org/10.1109/ICIP49359.2023.10222175.

Pham, H. H., Le, T. T., Tran, D. Q., Ngo, D. T., and Nguyen, H. Q. (2021). Interpreting chest x-rays via cnns that exploit hierarchical disease dependencies and uncertainty labels. *Neurocomputing*, 437:186–194. https://doi.org/10.1016/j.neucom.2020.03.127.

Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D., Bagul, A., Langlotz, C., Shpanskaya, K., et al. (2017). Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711.05225*. https://doi.org/10.48550/arXiv.1711.05225.

Rocha, J., Pereira, S. C., Pedrosa, J., Campilho, A., and Mendonça, A. M. (2024). Stern: Attention-driven spatial transformer network for abnormality detection in chest x-ray images. *Artificial Intelligence in Medicine*, 147:102737. https://doi.org/10.1016/j.artmed.2023.102737.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30. https://doi.org/10.48550/arXiv.1706.03762.

Wang, H., Wang, S., Qin, Z., Zhang, Y., Li, R., and Xia, Y. (2021). Triple attention learning for classification of 14 thoracic diseases using chest radiography. *Medical Image Analysis*, 67:101846. https://doi.org/10.1016/j.media.2020.101846.

Wang, Q., Wang, X., Liu, H., Wang, Y., Ren, J., and Zhang, B. (2024). A domain adaptive iot intrusion detection algorithm based on gwr-gcn

feature extraction and conditional domain adversary. *IEEE Internet of Things Journal*. https://doi.org/10.1109/JIOT.2024.3457894.

Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., and Summers, R. M. (2017). Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2097–2106. https://doi.org/10.1109/CVPR.2017.369.

Wang, X., Peng, Y., Lu, L., Lu, Z., and Summers, R. M. (2018). Tienet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9049–9058. https://doi.org/10.1109/CVPR.2018.00943.

Xu, Q. and Duan, W. (2024). Dualattnet: Synergistic fusion of image-level and fine-grained disease attention for multi-label lesion detection in chest x-rays. *Computers in Biology and Medicine*, 168:107742. https://doi.org/10.1016/j.compbiomed.2023.107742.

Xu, Y., Lam, H.-K., Bao, X., and Wang, Y. (2024). Learning group-wise spatial attention and label dependencies for multi-task thoracic disease classification. *Neurocomputing*, 573:127228. https://doi.org/10.1016/j.neucom.2023.127228.

Yang, B., Kang, Y., Zhang, L., and Li, H. (2021). Ggac: Multi-relational image gated gcn with attention convolutional binary neural tree for identifying disease with chest x-rays. *Pattern Recognition*, 120:108113. https://doi.org/10.1016/j.patcog.2021.108113.

Yuan, L., Chen, Y., Wang, T., Yu, W., Shi, Y., Jiang, Z.-H., Tay, F. E., Feng, J., and Yan, S. (2021). Tokens-to-token vit: Training vision transformers from scratch on imagenet. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 558–567. https://doi.org/10.1109/ICCV48922.2021.00060.

Zhang, Y., Luo, L., Dou, Q., and Heng, P.-A. (2023). Triplet attention and dual-pool contrastive learning for clinic-driven multi-label medical

image classification. *Medical image analysis*, 86:102772. https://doi.org/10.1016/j.media.2023.102772.

Zhao, G., Fang, C., Li, G., Jiao, L., and Yu, Y. (2021). Contralaterally enhanced networks for thoracic disease detection. *IEEE Transactions on Medical Imaging*, 40(9):2428–2438. https://doi.org/10.1109/TMI.2021.3077913.

Zhu, X., Liu, J., Liu, W., Ge, J., Liu, B., and Cao, J. (2023). Scene-aware label graph learning for multi-label image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1473–1482. https://doi.org/10.1109/ICCV48922.2023.00150.

Zhu, X., Pang, S., Zhang, X., Huang, J., Zhao, L., Tang, K., and Feng, Q. (2022). Pcan: Pixel-wise classification and attention network for thoracic disease classification and weakly supervised localization. *Computerized Medical Imaging and Graphics*, 102:102137. https://doi.org/10.1016/j.compmedimag.2022.102137.