

CheXNet: Combing Transformer and CNN for Thorax Disease Diagnosis from Chest X-ray Images^{*}

Xin Wu¹, Yue Feng^{1(✉)}, Hong Xu^{1,2}, Zhuosheng Lin¹, Shengke Li¹, Shihan Qiu¹, QiChao Liu¹, and Yuangang Ma¹

¹ Faculty of Intelligent Manufacturing, Wuyi University, Jiangmen 529020, Guangdong, China

² Institute for Sustainable Industries and Liveable Cities, Victoria University, Melbourne 8001, Australia

✉Corresponding author: J002443@wyu.edu.cn

Abstract. Multi-label chest X-ray (CXR) image classification aims to perform multiple disease label prediction tasks. This concept is more challenging than single-label classification problems. For instance, convolutional neural networks (CNNs) often struggle to capture the statistical dependencies between labels. Furthermore, the drawback of concatenating CNN and Transformer is the lack of direct interaction and information exchange between the two models. To address these issues, we propose a hybrid deep learning network named CheXNet. It consists of three main parts in the CNN and Transformer branches: Label Embedding and Multi-Scale Pooling module (MEMSP), Inner Branch module (IB), and Information Interaction module (IIM). Firstly, we employ label embedding to automatically capture label dependencies. Secondly, we utilize Multi-Scale Pooling (MSP) to fuse features from different scales and an IB to incorporate local detailed features. Additionally, we introduce a parallel structure that allows interaction between the CNN and the Transformer through the IIM. CNN can provide richer inputs to the Transformer through bottom-up feature extraction, whilst the Transformer can guide feature extraction in the CNN using top-down attention mechanisms. The effectiveness of the proposed method has been validated through qualitative and quantitative experiments on two large-scale multi-label CXR datasets with average AUCs of 82.56% and 76.80% for CXR11 and CXR14, respectively.

Keywords: Hybrid network · Multi-label · Chest X-ray image.

^{*} This work is supported by the Basic Research and Applied Basic Research Key Project in General Colleges and Universities of Guangdong Province, China (2021ZDZX1032); the Special Project of Guangdong Province, China (2020A1313030021); and the Scientific Research Project of Wuyi University (2018TP023, 2018GR003).

1 Introduction

Chest X-ray (CXR), as a painless examination method, plays an important role in auxiliary clinical diagnosis. Meanwhile, it is one of the most common radiology tests used to screen for and diagnose a variety of lung conditions. However, achieving highly reliable diagnostic results for thoracic diseases using CXRs remains challenging due to the dependence on the expertise of radiologists.

In the past few years, CNNs in particular have shown remarkable performance in the diagnosis of various thoracic diseases [11]. Pesce et al. [12] utilized a CNN to extract features and input them into a classifier and a locator for detecting lung lesions. Sahlol et al. [13] employed a pre-trained MobileNet to extract features from CXR images. Baltruschat et al. [2] evaluated the performance of various methods for classifying 14 disease labels using an extended ResNet50 architecture and text data.

The great success of Transformers [16] has inspired researchers [3,8,15] to try to introduce Transformers into the field of computer vision. Furthermore, some studies have utilized Transformers to capture multi-label information in images and improve classification performance. Taslimi et al. [14] introduced a Swin Transformer backbone, which predicts each label by sharing components across models. Xiao et al. [17] utilized Masked Auto encoders to pre-train vision Transformers (ViT), reconstructing missing pixel images from a small portion of separate X-ray image.

However, there are still some challenges that need to be addressed in the classification of multi-label CXR images. Firstly, there may exist interdependencies between different labels in multi-label CXR images, such as certain lung abnormalities being related to cardiac abnormalities. The second is the imbalance of labels. Thirdly, there can be prominent local lesion features and scattered global features in the images.

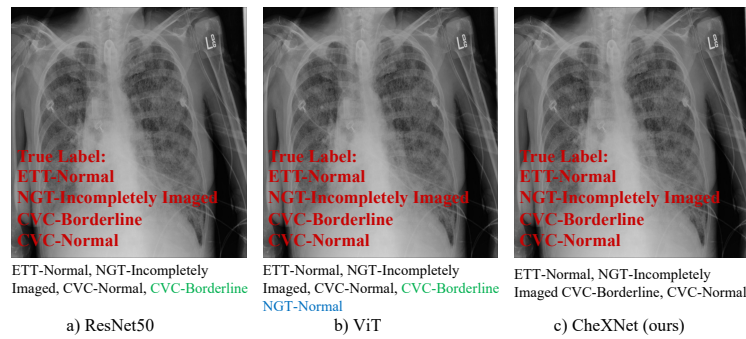


Fig. 1. Examples of recognition results of the CNN (ResNet50), the Transformer (ViT) and our proposed CheXNet. The true labels are in red font, the incorrectly identified labels are in green font, and the correctly predicted labels with small probability are in blue font.

In this paper, we propose CheXNet to address the aforementioned challenges. Firstly, we introduce self-attention operations on label embedding. This approach adaptively captures the correlations among labels without relying on manually predefined label relationships. Secondly, we employ cross-attention between image features and label features, allowing the model to weigh the image features based on the importance of each label. Additionally, we introduce the MSP block in the Transformer branch to extract features at different levels and then fuse them. Finally, we utilize a parallel structure that allows interaction between the CNN and the Transformer. To demonstrate the superiority of our approach, we visualize ResNet50, ViT and the proposed CheXNet, as shown in Fig. 1.

The main contributions of this study are summarized as follows:

- (1) We propose a CheXNet model for multi-label CXR image classification, which captures both short local features and global representations.
- (2) For the Transformer branch, we introduce the MSP block to perform multi-scale pooling, aiming to enhance the richness and diversity of feature representations. Additionally, the label embedding, using self-attention, adaptively captures the correlations between labels. For the CNN branch, an embedded residual structure is employed to learn more detailed information. The IIM supports cross-branch communication and helps to explore implicit correlations between labels.
- (3) We evaluate the CheXNet on two publicly available datasets, CXR11 and CXR14. The experimental results demonstrate that the CheXNet outperforms existing models on both datasets in terms of performance.

2 Related Work

In this section, we discuss label dependency and balance issues observed in multi-label classification methods, as well as multi-label CXR image classification methods for a wide range of lesion locations.

2.1 Label Dependency and Imbalance

In multi-label CXR image classification, the challenges of label dependency and label imbalance are common. These issues significantly impact accurate classification and model performance evaluation. To address these challenges, researchers have employed various methods, including weighted loss functions, hierarchical classification, and transfer learning. Allaouzi et al. [1] proposed a method that combines a CNN model with convolutional filters capable of detecting local patterns in images. By learning the dependencies between features and labels in multi-label classification tasks, they enhanced the accuracy and reliability of disease diagnosis. Lee et al. [7] introduced a hybrid deep learning model. The model consists of two main modules: image representation learning and graph representation learning. Yang et al. [19] utilized a triple network ensemble learning framework consisting of three CNNs and a classifier. The framework aimed to learn combined features and address issues such as class imbalance and network ensemble.

2.2 Extensive Lesion Location

In CXR images, there can be multiple lesion locations, with each lesion corresponding to a different pathology. Some researchers have adopted different methods, such as region localization, and attention mechanism. Ma et al. [10] proposed a cross attention network approach for the automated classification of thoracic diseases in CXR images. This method efficiently extracts more meaningful representations from the data and improves performance through cross-attention, requiring only image-level annotations. Guan et al. [4] introduced category residual attention learning to address the problem of pathologies interfering with the recognition of target-related pathologies. This approach enables the prediction of multiple pathologies' presence in the attention view of specific categories. The goal of this method is to suppress interference from irrelevant categories by assigning higher weights to relevant features, thereby achieving automatic classification of thoracic diseases.

3 Approaches

The multi-label CXR image classification method of CheXNet consists of three main stages, as shown in Fig. 2. The first stage involves extracting initial features, using the Stem module. These features are then split into two branches, one sent to the Transformer branch and the other to the CNN branch. In the second stage, the Transformer branch utilizes MEMSP, while the CNN branch utilizes a nested IB. The stacking of MEMSP modules and IB modules corresponds to the number of layers in a vanilla Transformer, which is set to 12. Additionally, the IIM consists of the CNN branch to the Transformer branch (C2T) and the Transformer branch to the CNN branch (T2C) components, which progressively interactively fuse feature maps. Finally, after obtaining the features T and features C of the two branches separately, we directly sum up the branch fusions.

3.1 Label Embedding and MSP Block

For an input image x , the Stem module first extracts feature. Then, in the Transformer branch, the label embedding is used to query the MEMSP. However, most existing works primarily focus on regression from inputs to binary labels, while overlooking the relationship between visual features and the semantic vectors of labels. Specifically, we obtain the features extracted by the convolution as the K and V inputs and the label embeddings as the query $Q_i \in \mathbb{R}^{C \times d}$, with d denoting dimensionality, cross-noting the desired K , V and Q . We use a Transformer-like architecture, which includes a self-attention module, a cross-attention block, an MSP block, and an FFN block. When using the self-attention block, label embedding is the conversion of labels into vector representations so that the computer can better understand and process them. By incorporating label embeddings into the MEMSP module, it can effectively and automatically capture the semantics

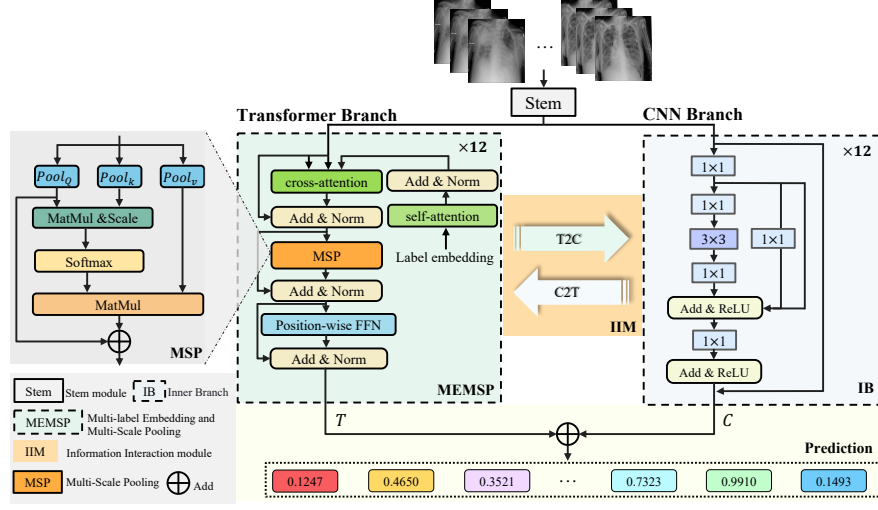


Fig. 2. An overview diagram of the proposed CheXNet framework, where the Transformer branch uses the MEMSP block (green part), the CNN branch uses the IB block (blue part), and the IIM uses C2T and T2C (orange part). T and C denote the final output characteristics of MEMSP and IB, respectively.

of the labels and make more accurate predictions for multiple labels associated with an input sample. K , V , and Q are label features, all denoted as Q_{i-1} . The specific formula is as follows:

$$\text{self-attention} : Q_i^{(1)} = \text{MultiHead}(Q_{i-1}, Q_{i-1}, Q_{i-1}) \quad (1)$$

$$\text{coss-attention} : Q_i^{(2)} = \text{MultiHead}(Q_i^{(1)}, K, V) \quad (2)$$

$\text{MultiHead}(Q, K, V)$ and $Q_i = \text{FNN}(x)$ have the same decoder definition as the standard Transformer [16]. We did not use masked multi-head attention, but instead used self-attention, as autoregressive prediction is not required in multi-label image classification. In the MSP block, multiple pooling operations (2×2 , 3×3 , 4×4) are applied separately to Q , K , and V , allowing each pooling size to extract features at different levels. These pooled features are then fused together. Specifically, given the input feature F , three copies of F are created to obtain pools, denoted as $Pool_i$, where i represents q, k, v . The $Pool_q$ and $Pool_k$ of different scales are multiplied element-wise, resulting in $Pool'_q$ and $Pool'_k$. The $Pool_v$ of different scales are summed element-wise, resulting in $Pool'_v$. Then, the Softmax function is applied to normalize the values, ensuring that the sum of all elements is equal to 1. The normalized values are used as weights to linearly combine the value vectors in $Pool'_v$, and the residual of $Pool'_q$ is added to the

weighted sum. The final output is the sum of these two terms.

$$Pool'_q = \text{SUM}(\text{MatMul}(Pool^i_q)) \quad (3)$$

$$Pool'_k = \text{SUM}(\text{MatMul}(Pool^i_k)) \quad (4)$$

$$Pool'_v = \text{SUM}(Pool^i_v) \quad (5)$$

$$MSP = \text{Softmax} \left(\frac{Pool'_q Pool'_k}{\sqrt{d_k}} \right) Pool'_v + Pool'_q \quad (6)$$

Where $i = 2, 3, 4$, d_k represents the vector dimension of q and k .

3.2 Inner Branch

The CNN branch adopts a nested structure with multiple residual branches, where the resolution of feature maps decreases with the depth of the network. Firstly, we employ the basic bottleneck block of ResNet [5], typically consisting of three convolutional layers. The first convolutional layer uses a 1×1 projection convolution to reduce the dimensionality of the feature maps. The second convolutional layer utilizes a larger 3×3 spatial convolution to extract features. The third convolutional layer again employs a smaller 1×1 projection convolution to further reduce the dimensionality of the feature maps. It also includes a residual connection between the input and output. The IB modifies the residual block by introducing an additional nested branch, replacing the main 3×3 convolution with a residual block. The CNN branch can continuously contribute localized feature details to the Transformer branch through the C2T module within the IIM module. The IB module significantly enhances the model's feature representation capacity, particularly in the context of multi-label CXR classification tasks.

3.3 C2T and T2C in IIM

For the CNN branch, mapping features to the Transformer branch is a crucial problem. Similarly, for the Transformer branch, embedding patch representations into the CNN branch is also important. CNN features are represented as $[B, C, H, W]$, where B denotes the batch size, C denotes the number of channels, H denotes the image height, and W denotes the image width. On the other hand, Transformer features are represented as $[B, _, C]$, where $_$ represents the sum of the number of image patches and the number of class tokens, usually equal to $H \times W + 1$. To address this issue, we propose the C2T and T2C to progressively integrate the feature maps in an interactive manner, as shown in Fig. 3.

The C2T method involves dimensionality transformation of the feature maps using 1×1 convolutions. Additionally, we combine features from different channels to enhance the expressive power of the features. We utilize average pooling to downsample the feature maps, reducing their spatial dimensions while preserving the essential information. The GELU activation function is employed for

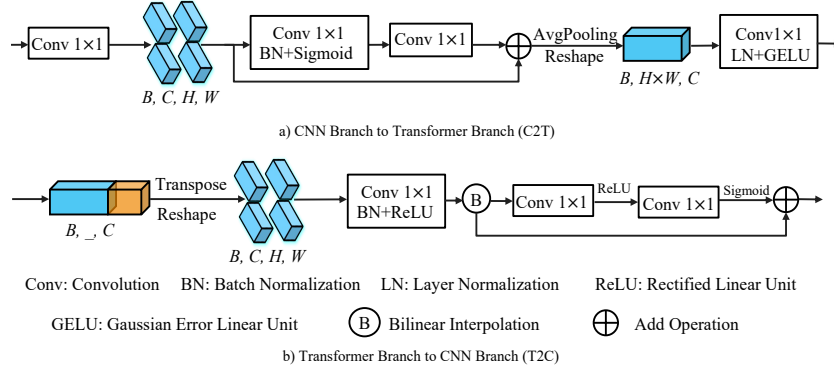


Fig. 3. Structure of Information Interaction Module (IIM), which includes the C2T and T2C.

fast convergence and reduced training time, thereby improving the efficiency of the model training. Layer normalization is used for feature regularization.

The T2C process involves aligning the spatial scale by employing appropriate up-sampling techniques. Batch normalization is used to regularize the features. We use the commonly used ReLU activation function in the convolutional operations. Bilinear interpolation is applied to upsample the feature maps, enhancing the spatial resolution and capturing finer details. Similar to C2T, multiple 1×1 convolutions are utilized for optimizing feature information exchange. After the 1×1 convolutions, we incorporate both ReLU and Sigmoid activations in a cross-interaction manner to improve the nonlinear fitting capability. Finally, a residual connection is introduced between the output of the bilinear interpolation and the output after a series of operations to preserve important information and enhance the model’s representational capacity.

4 Experiments

4.1 Dataset

The Catheter and Line Position Challenge on Kaggle¹ is a competition that involves classifying 40000 images to detect misplaced catheters. In this study, 30083 CXR image training data were used as multi-label sample classification, which was named CXR11.

The NIH ChestX-ray14 dataset², which was named CXR14, includes 112120 frontal X-ray images from 30805 unique patients with annotations for 14 common diseases. Limited by computer equipment, we only use part of the data. A detailed description of CXR11 and CXR14 is provided in the supplementary material.

¹ CXR11: kaggle.com/competitions/ranzcr-clip-catheter-line-classification/data

² CXR14: nihcc.app.box.com/v/ChestXray-NIHCC

4.2 Comparison to the State-of-the-Arts

To ensure fair comparisons, we utilized the aforementioned parameter settings and classical methods to calculate the AUC scores for each category and the average AUC score for all diseases, as presented in Tables 1 and 2 for the CXR11 and CXR14 datasets, respectively. To further validate the feasibility of our proposed approach, we compared it with eight other state-of-the-art medical image classification networks, achieving the best performance in multi-label classification. Furthermore, we conducted paired t-tests to assess the statistical significance of performance differences between our proposed model and those proposed by other authors. Based on the p-values, we can conclude that there are statistically significant differences in the performance of each model for this specific task.

Table 1. Comparison of the classification performance of different models on CXR11 datasets, where Swan Ganz denotes Swan Ganz Catheter Present. The best results are shown in bold.

AUC score (%)		Res-Net34 [5]	Res-Net 50 [5]	Res-NeXt 50 [18]	SE-ReNet 50[6]	ViT [3]	Swin Transformer [8]	Conv-NeXt [9]	DeiT [15]	CheX-Net (ours)
ETT	Abnormal	78.26	82.97	74.16	85.95	77.68	67.23	74.96	79.36	92.49
	Borderline	88.51	89.33	85.96	89.96	83.02	78.49	81.99	83.56	88.04
	Normal	96.96	97.37	96.95	97.29	88.18	83.84	94.96	87.95	97.92
NGT	Abnormal	78.15	80.69	79.63	77.26	77.18	74.09	72.54	76.48	81.38
	Borderline	78.10	79.35	81.14	78.58	73.41	69.74	69.40	74.51	81.72
	Incompletely Imaged	92.94	93.09	93.20	92.37	87.88	82.97	88.66	87.25	94.19
	Normal	91.19	92.72	91.57	91.55	85.09	81.86	88.74	85.93	92.72
CVC	Abnormal	59.64	61.26	62.11	61.66	60.60	59.29	56.05	61.44	62.88
	Borderline	58.80	58.89	59.26	58.99	56.36	56.18	58.57	57.44	59.36
	Normal	57.20	59.14	61.32	57.89	55.80	53.38	59.19	56.23	60.70
Swan Ganz		92.95	95.72	97.11	93.44	87.48	84.99	90.51	89.31	96.72
Mean		79.34	80.96	80.22	80.45	75.70	72.01	75.97	76.32	82.56
p-value		.0108	.0435	.0896	.0054	.0001	.0001	.0006	.0001	-

Table 1 presents the results of different models for CXR11 classification. From Table 1, it can be observed that our proposed method achieved the highest average AUC score (82.56%). Among the compared models, ResNet50 attained the highest average AUC score of 80.96%, while Swin Transformer achieved the lowest average AUC score of 72.01%, which is 1.60% and 10.56% lower than our proposed method, respectively. Regarding ETT-Abnormal, our method outperformed the other top-performing model, ResNet50, by 9.52%.

Table 2 provides the results of different models for CXR14 classification. From Table 2, it can be seen that our proposed method obtained the highest average AUC score (76.80%). Among the compared models, ConvNeXt achieved

the highest average AUC score of 76.73%, while ViT obtained the lowest average AUC score of 54.00%, which are only 0.07% and 22.80% lower than our proposed method, respectively. Although ConvNeXt performed similarly to CheXNet, our method exhibited superior performance in multiple categories. For example, for Fibrosis, CheXNet outperformed ConvNeXt by 4.33%.

Table 2. Comparison of the classification performance of our different models on CXR14 datasets. The best results are shown in bold.

AUC score (%)	Res-Net34 [5]	Res-Net 50 [5]	Res-NeXt 50 [18]	SE-ReNet 50[6]	ViT [3]	Swin Transformer [8]	Conv-NeXt [9]	DeiT [15]	CheX-Net (ours)
Atelectasis	73.35	73.87	73.51	73.81	53.53	62.48	73.31	70.19	74.24
Cardiomegaly	91.00	89.18	89.03	89.72	50.18	56.67	91.42	80.81	91.76
Consolidation	70.19	71.72	72.98	71.73	52.43	64.33	71.19	68.06	71.69
Edema	85.40	85.08	84.32	86.02	68.78	76.52	84.89	83.85	83.76
Effusion	81.02	81.79	82.09	81.40	44.74	64.95	82.77	77.17	81.96
Emphysema	76.39	83.29	80.63	80.88	53.19	55.94	80.79	75.45	81.09
Fibrosis	76.66	76.37	78.26	77.18	60.02	69.27	73.94	73.84	78.27
Hernia	69.00	79.62	80.96	78.88	67.29	69.35	88.84	69.99	89.04
Infiltration	69.12	68.85	69.24	69.78	56.07	61.89	68.87	67.16	68.41
Mass	74.96	77.04	76.07	74.92	47.56	57.07	73.97	66.77	76.29
Nodule	69.65	68.94	71.36	69.34	50.83	61.15	72.73	65.79	69.71
Pleural Thickening	67.83	66.99	69.23	67.43	46.46	57.09	70.51	63.99	67.39
Pneumonia	59.02	59.21	58.32	60.39	48.97	54.07	55.68	50.50	59.61
Pneumothorax	82.39	83.32	81.81	85.75	55.90	61.65	85.34	74.86	81.96
Mean	74.71	76.09	76.27	76.23	54.00	62.32	76.73	70.60	76.80
p-value	.0706	.1638	.2041	.2400	.0000	.0000	.4559	.0001	-

For the CXR11 and CXR14 datasets, CheXNet is compared with other algorithms and the overall performance of the network is demonstrated. In the supplementary material, a qualitative analysis of the classification performance of each compared method and the AUC for each disease is presented.

4.3 Ablation Study

To assess the effectiveness of the proposed CheXNet network and the contribution of each module within the overall network, we conducted a series of step-wise ablation experiments on the CXR11 dataset. The following models were compared to evaluate their performance:

Baseline: The Transformer branch is a standard Transformer encoder, and the CNN branch consists of residual blocks from the ResNet network, without any gated mechanisms for information exchange and communication.

Model 1: Based on the baseline, the standard Transformer encoder is replaced with the MEMSP module, which includes label embedding but not the MSP block.

Model 2: Based on the baseline, the residual blocks of the ResNet network are replaced with IB, introducing internal nesting.

Model 3: Built upon Model 1, the residual blocks are replaced with IB.

Model 4: Built upon Model 2, the IIM modules are added.

Model 5: Built upon Model 3, the IIM modules are added. In this case, the MEMSP module does not include the MSP block.

CheXNet (ours): Built upon Model 5, the MSP module is added.

Table 3. Classification performance of different models in our system on the CXR11 dataset, where Swan Ganz denotes Swan Ganz Catheter Present. The best results are in bold.

AUC socre (%)		Baseline	Model 1	Model 2	Model 3	Model 4	Model 5	CheXNet (ours)
ETT	Abnormal	80.63	93.70	83.74	90.87	88.75	90.08	92.49
	Borderline	88.18	85.00	87.08	87.93	84.93	85.46	88.04
	Normal	97.57	97.05	97.85	97.88	97.84	97.84	97.92
NGT	Abnormal	78.39	74.76	81.02	79.39	81.22	80.92	81.38
	Borderline	78.64	78.97	80.40	81.53	81.86	82.05	81.72
	Incompletely Imaged	92.86	92.17	93.39	93.60	94.26	94.33	94.19
CVC	Normal	91.96	89.93	91.73	92.57	92.18	92.18	92.72
	Abnormal	60.14	63.90	64.05	64.56	63.11	61.83	62.88
	Borderline	57.66	58.63	58.74	56.39	59.99	60.00	59.36
	Normal	58.04	58.32	60.02	57.32	58.33	58.70	60.70
Swan Ganz		98.11	94.68	97.73	97.50	95.65	96.36	96.72
Mean		80.20	80.65	81.50	81.78	81.65	81.79	82.56

Table 3 provides a quantitative analysis of the experimental results on the CXR11 dataset for different modules. Compared to the Baseline and Models 1 to 5, CheXNet demonstrates improvements in AUC of 2.36%, 1.91%, 1.06%, 0.78%, 0.91%, and 0.77%, respectively. The proposed CheXNet achieves the best classification results through a combination of several modules. From the AUC values for each category in Table 3, it can be observed that each module plays a role, confirming the effectiveness of these modules. Fig. 4 provides a visual representation of our classification results on the CXR11 dataset.

5 Conclusion

In this paper, we propose a hybrid deep learning network named CheXNet. The label embedding automatically captures label dependencies, effectively alleviating label dependency issues. We incorporate multi-scale pooling to fuse features

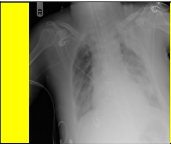
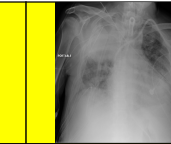
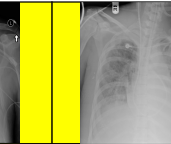
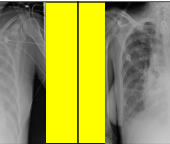
Image				
				
Baseline	ETT-Abnormal: 0.0559 ETT-Borderline: 0.2854 ETT-Normal: 0.2424 NGT-Abnormal: 0.1427 NGT-Borderline: 0.1501 NGT-Incompletely Imaged: 0.3459 NGT-Normal: 0.4547 CVC-Abnormal: 0.3872 CVC-Borderline: 0.6118 CVC-Normal: 0.9623 Swan Ganz Catheter Present: 0.0581	ETT-Abnormal: 0.0090 ETT-Borderline: 0.0099 ETT-Normal: 0.9971 NGT-Abnormal: 0.0914 NGT-Borderline: 0.1635 NGT-Incompletely Imaged: 0.0876 NGT-Normal: 0.6131 CVC-Abnormal: 0.3603 CVC-Borderline: 0.6336 CVC-Normal: 0.9855 Swan Ganz Catheter Present: 0.0691	ETT-Abnormal: 0.0593 ETT-Borderline: 0.3508 ETT-Normal: 0.1184 NGT-Abnormal: 0.1030 NGT-Borderline: 0.1299 NGT-Incompletely Imaged: 0.1789 NGT-Normal: 0.6926 CVC-Abnormal: 0.3446 CVC-Borderline: 0.4913 CVC-Normal: 0.9968 Swan Ganz Catheter Present: 0.9537	ETT-Abnormal: 0.0040 ETT-Borderline: 0.0021 ETT-Normal: 0.9985 NGT-Abnormal: 0.0761 NGT-Borderline: 0.1471 NGT-Incompletely Imaged: 0.4195 NGT-Normal: 0.0794 CVC-Abnormal: 0.2797 CVC-Borderline: 0.7185 CVC-Normal: 0.9789 Swan Ganz Catheter Present: 0.0312
CheXTransCNN (ours)	ETT-Abnormal: 0.057 ETT-Borderline: 0.2563 ETT-Normal: 1.0000 NGT-Abnormal: 0.2371 NGT-Borderline: 0.2770 NGT-Incompletely Imaged: 0.8600 NGT-Normal: 0.8551 CVC-Abnormal: 0.4750 CVC-Borderline: 0.7234 CVC-Normal: 0.9724 Swan Ganz Catheter Present: 0.7545	ETT-Abnormal: 0.0293 ETT-Borderline: 0.1147 ETT-Normal: 1.0000 NGT-Abnormal: 0.1370 NGT-Borderline: 0.2177 NGT-Incompletely Imaged: 0.6251 NGT-Normal: 0.9257 CVC-Abnormal: 0.4001 CVC-Borderline: 0.7223 CVC-Normal: 0.9853 Swan Ganz Catheter Present: 0.0973	ETT-Abnormal: 0.2175 ETT-Borderline: 0.9594 ETT-Normal: 0.2697 NGT-Abnormal: 0.2635 NGT-Borderline: 0.2741 NGT-Incompletely Imaged: 0.8079 NGT-Normal: 0.7961 CVC-Abnormal: 0.4184 CVC-Borderline: 0.6121 CVC-Normal: 0.9939 Swan Ganz Catheter Present: 0.9819	ETT-Abnormal: 0.0112 ETT-Borderline: 0.0060 ETT-Normal: 1.0000 NGT-Abnormal: 0.1360 NGT-Borderline: 0.2325 NGT-Incompletely Imaged: 0.9370 NGT-Normal: 0.2469 CVC-Abnormal: 0.3816 CVC-Borderline: 0.7446 CVC-Normal: 0.9769 Swan Ganz Catheter Present: 0.0593

Fig. 4. Comparison of the number of cases for each disease in the CXR11 dataset.

from different scales and an inner branch to capture more locally detailed features. Moreover, we employ the IIM module to facilitate information interaction between the CNN and Transformer, enabling the network to effectively utilize both local and global lesion features. We conducted one ablation experiment and two comparative experiments to analyze our method. The extensive experimental results on the two CXR datasets demonstrate the effectiveness and generalization ability of our approach in the field of multi-label medical classification.

References

1. Allaouzi, I., Ben Ahmed, M.: A novel approach for multi-label chest x-ray classification of common thorax diseases. *IEEE Access* **7**, 64279–64288 (2019)
2. Baltruschat, I.M., Nickisch, H., Grass, M., Knopp, T., Saalbach, A.: Comparison of deep learning approaches for multi-label chest x-ray classification. *Scientific reports* **9**(1), 1–10 (2019)
3. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020)
4. Guan, Q., Huang, Y.: Multi-label chest x-ray image classification via category-wise residual attention learning. *PATTERN RECOGNITION LETTERS* **130**(SI), 259–266 (2020)
5. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2016), seattle, WA, JUN 27-30, (2016)
6. Hu, J., Shen, L., Albanie, S., Sun, G., Wu, E.: Squeeze-and-excitation networks. vol. 42, pp. 2011–2023. *IEEE COMPUTER SOC* (2020)

7. Lee, Y.W., Huang, S.K., Chang, R.F.: Chexgat: A disease correlation-aware network for thorax disease diagnosis from chest x-ray images. *Artificial Intelligence in Medicine* **132**, 102382 (2022)
8. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: 2021 IEEE/CVF INTERNATIONAL CONFERENCE ON COMPUTER VISION (ICCV 2021). pp. 9992–10002 (2021), eLECTR NETWORK, OCT 11–17, (2021)
9. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 11976–11986 (June 2022), new Orleans, LA, JUN 18–24, (2022)
10. Ma, C., Wang, H., Hoi, S.C.H.: Multi-label thoracic disease image classification with cross-attention networks. In: MEDICAL IMAGE COMPUTING AND COMPUTER ASSISTED INTERVENTION - MICCAI 2019, PT VI. Lecture Notes in Computer Science, shenzhen, China (2019)
11. Majkowska, A., Mittal, S., Steiner, D.F., Reicher, J.J., McKinney, S.M., Duggan, G.E., Eswaran, K., Cameron Chen, P.H., Liu, Y., Kalidindi, S.R., Ding, A., Corrado, G.S., Tse, D., Shetty, S.: Chest radiograph interpretation with deep learning models: Assessment with radiologist-adjudicated reference standards and population-adjusted evaluation. *Radiology* **294**(2), 421–431
12. Pesce, E., Joseph Withey, S., Ypsilantis, P.P., Bakewell, R., Goh, V., Montana, G.: Learning to detect chest radiographs containing pulmonary lesions using visual attention networks. *Medical Image Analysis* **53**, 26–38 (2019)
13. Sahlol, A.T., Abd Elaziz, M., Tariq Jamal, A., Damaševičius, R., Farouk Hassan, O.: A novel method for detection of tuberculosis in chest radiographs using artificial ecosystem-based optimisation of deep neural network features. *Symmetry* **12**(7), 1146 (2020)
14. Taslimi, S., Taslimi, S., Fathi, N., Salehi, M., Rohban, M.H.: Swinchex: Multi-label classification on chest x-ray images with transformers. *arXiv preprint arXiv:2206.04246* (2022)
15. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jegou, H.: Training data-efficient image transformers & distillation through attention. In: INTERNATIONAL CONFERENCE ON MACHINE LEARNING, VOL 139. vol. 139, pp. 7358–7367. ELECTR NETWORK, JUL 18–24, (2021)
16. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.u., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems. vol. 30. Curran Associates, Inc. (2017)
17. Xiao, J., Bai, Y., Yuille, A., Zhou, Z.: Delving into masked autoencoders for multi-label thorax disease classification. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 3588–3600 (January), los Angeles, CA (2023)
18. Xie, S., Girshick, R., Dollar, P., Tu, Z., He, K.: Aggregated residual transformations for deep neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (July), honolulu, HI (2017)
19. Yang, M., Tanaka, H., Ishida, T.: Performance improvement in multi-label thoracic abnormality classification of chest x-rays with noisy labels. *INTERNATIONAL JOURNAL OF COMPUTER ASSISTED RADIOLOGY AND SURGERY* **18**(1, SI), 181–189 (2023)