

Intelligent Diagnosis Methods for Multi-label Chest X-ray Images Based on Self-supervised Learning

Kaiwen Luo

*Faculty of Health Science, University of Macau, Macau, China
cc21824@um.edu.mo*

Abstract. With the rapid increase in the volume of medical imaging data, self-supervised learning has shown significant potential in the intelligent diagnosis of chest X-ray images. This paper proposes a self-supervised pre-training method based on the SimCLR framework. It uses the large-scale ChestX-ray14 dataset for unsupervised pre-training and then performs transfer fine-tuning on the multi-label classification task. Specifically, the paper adopt EfficientNet-B0 as the encoder structure and optimize the network's feature extraction ability through contrastive learning. For the downstream multi-label classification task, the paper uses a fully connected layer for output, Sigmoid activation, and combine the Focal Loss function to handle the problem of class imbalance. The experimental results show that compared with the model without pre-training, the pre-trained model has an approximately 2.51% increase in the average AUC value and an approximately 3.04% increase in the F1 value on the validation set. In addition, the paper used the Grad-CAM visualization technique to analyze the regions of interest of the model and conducted in-depth analysis on the categories with a high error rate. The experiments also included dynamic threshold optimization, training curve plotting, and performance comparison across other categories to confirm the method's efficacy.

Keywords: Self-supervised learning, SimCLR, Multi-label classification, EfficientNet, Grad-CAM.

1. Introduction

Chest X-ray is an important imaging tool for diagnosing various diseases such as those in the lungs. Common chest diseases often have a multi-label nature (e.g., pneumonia and cardiac enlargement coexist). To improve the level of diagnostic automation, researchers have constructed large-scale annotated datasets, such as the ChestX-ray14 dataset released by NIH, which contains 112,120 frontal chest radiographs and is annotated with 14 common diseases [1]. Traditional supervised learning methods have achieved certain results on this dataset, but limited by the high cost of annotation, many studies have begun to focus on unsupervised or self-supervised methods. Self-Supervised Learning can learn useful features from a large amount of unannotated data and is particularly applicable to the field of medical images [2,3]. Self-supervised learning approaches can be broadly categorized as predictive, generative, or contrastive methods [4]. For example, the SimCLR framework proposed by Chen et al. achieved a representation quality comparable to that of

supervised learning on natural images through the contrastive learning method [2]. Meanwhile, the EfficientNet series of networks significantly improved the accuracy while consistently scaling the depth, width, and resolution to preserve model efficiency [5]. In addition, multi-label classification often faces the difficulties of severe class imbalance. Introducing the Focal Loss can effectively reduce the weights of easily classifiable samples and focus more attention on difficult samples [6]. To improve the model's interpretability, using gradient-weighted activation maps, the Grad-CAM approach displays the areas of the image that the model is interested in [7], which helps to analyze the basis for the model's decision-making. Based on the above technical background, this paper proposes a pre-training and fine-tuning process based on SimCLR, and conducts optimization research for the multi-label classification task of chest X-rays. The main work includes Conducting SimCLR self-supervised pre-training using the dataset to improve the feature representation ability. Using EfficientNet-B0 as the encoder and obtaining better features through contrastive learning after pre-training. In the downstream multi-label classification, a fully connected layer + Sigmoid output is adopted, and the Focal Loss is combined for optimization. The target metrics are AUC and F1. Comparing the performance of the pre-trained model and the randomly initialized model to verify the performance gain brought by pre-training. Visualizing and analyzing the regions that the model focuses on through Grad-CAM, and analyzing the performance of the category with the highest error rate. The experimental process includes a dynamic threshold optimization strategy, the plotting of training curves, and the comparison of the performance of various categories to comprehensively evaluate the model's effectiveness.

2. Method

The method in this paper mainly consists of three parts: self-supervised pre-training, fine-tuning for downstream multi-label classification, and visual analysis of the model.

Self-supervised pre-training (SimCLR): The SimCLR contrastive learning framework is adopted to pre-train the ChestX-ray14 data. Specifically, two randomly augmented views (including random cropping, horizontal flipping, color perturbation, etc.) are sampled from each input image. Two sets of representations are extracted through the shared EfficientNet-B0 encoder network and then mapped to the contrastive loss space via a small non-linear projection head. The normalized temperature cross-entropy loss (NT-Xent) is used as the contrastive loss to make the features of the two augmented views of the same image closer in the high-dimensional space and the features of different images farther apart [2]. This unsupervised contrastive learning enables the model to learn features useful for lesion detection without labels.

Encoder structure: EfficientNet - B0 is selected as the base network to balance the number of model parameters and performance [5]. EfficientNet uniformly adjusts the depth, width, and resolution of the network through a compound scaling strategy and can achieve high accuracy while maintaining a small scale [5]. During the SimCLR pre-training phase, EfficientNet - B0 serves as the backbone network to output feature maps, which participate in contrastive loss training through the projection head. Figure1 shows pre-training process framework.

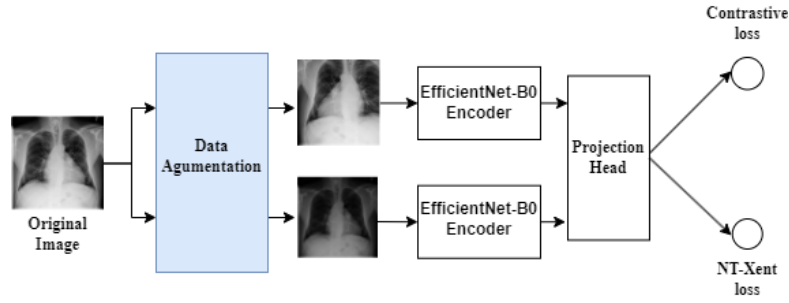


Figure 1. Pre-training process framework (picture credit: original)

Downstream multi-label classification: After completing self-supervised pre-training, remove the projection head of SimCLR, only keep the EfficientNet-B0 encoder, and then connect a fully connected layer to form a classification head. Global average pooling is the first thing the classification head does on the encoder output and then connects multiple parallel fully connected units. Each unit corresponds to a disease label, and the probability of this label is output after Sigmoid activation. Since chest X-ray classification is a multi-label task, each image can have multiple labels simultaneously. Therefore, independent Sigmoid outputs are used, and the loss is calculated for each label. To address the serious imbalance between positive and negative samples, Focal Loss is introduced to replace the traditional binary cross-entropy loss. Based on the calculation of the original cross-entropy [6], Focal Loss reduces the loss weight of easily classifiable samples (samples with high prediction confidence), enabling the model to pay more attention to the minority class samples that are difficult to classify [6]. Specifically, the loss of each label is defined as:

$$FL(p_t) = -\alpha(1-p_t)^\gamma \log(p_t) \quad (1)$$

Among them, p_t is the predicted probability of this label, γ is the focal adjustment parameter, and α is the category balance factor.

Dynamic Threshold Optimization: The model prediction output is the probability of each label. This paper uses the Dynamic Threshold Strategy to select the optimal threshold for each label based on the validation set data to maximize the F1 Score, thereby converting the probability into the final Binary Prediction. This strategy takes into account that the probability distributions and imbalance levels of different labels may vary, and improves the overall classification performance by tuning the thresholds.

Model Visualization Analysis: To understand the decision-making mechanism of the model, the Gradient-weighted Class Activation Mapping algorithm is used to visualize the classification results [7]. Grad-CAM uses the gradient information of the target category in the last convolutional layer to sum the Feature Map weighted by the gradients, generating a Class Activation Heatmap to highlight the areas of the image that the model is interested in [7]. In this paper, Grad-CAM is used to conduct a visual analysis on a part of the samples to verify whether the model has learned features related to Lesion Location and to analyse the regions of interest in the case of misclassification.

3. Experimental design

3.1. Dataset and preprocessing

Dataset and Preprocessing: This study uses the ChestX-ray14 Dataset published by National Institutes of Health (NIH) [1]. This dataset contains 112,120 Frontal Chest X-ray images, which are labeled with 14 common chest diseases (such as pneumonia), belonging to the Multi-label Classification Task. The images in the dataset have different sizes. In this experiment, the Image Resizing is uniformly set to 224×224 Resolution, and Pixel Normalization is performed. All images (without labels) are used in the Pre-training stage, and the corresponding label information is used for Supervised Learning in the Fine-tuning stage. In this paper, the dataset is divided into a Training Set and a Validation Set. The Validation Set accounts for approximately 20%, which is used for Model Selection and Metric Evaluation.

3.2. Pre-training and fine-tuning settings

Pre-training and Fine-tuning settings: In the SimCLR self-supervised pre-training stage, this paper uses EfficientNet-B0 as the backbone network, sets the training period to 100 epochs, the Adam optimizer is used, the initial learning rate is set to 0.001, and sets the temperature parameter T to 0.5. The augmentation strategies include random cropping, horizontal flipping, color jittering, and grayscale transformation. After pre-training is completed, transfer the weights of the encoder to the Multi-label Classification model. In the Fine-tuning stage, connect a fully connected classification head to the output of the encoder and use Focal Loss to train the model. The Adam optimizer is also used in the Fine-tuning training, initializes the learning rate to 0.0001, evaluates the performance of the Validation Set once per epoch, and saves the best model.

3.3. Evaluation metrics

Evaluation metrics: Use the average Area Under the Curve (Area Under ROC Curve) and macro-average F1 Score as the main evaluation metrics. The AUC can comprehensively evaluate the classification performance of the model under different thresholds, while the F1 Score takes both precision and recall into account, and is especially suitable for imbalanced multi-label scenarios. This paper records and plots the Validation Set Area Under the Curve and F1 curves during the training process to intuitively observe the convergence behavior of the model. In addition, the individual AUC and F1 values are calculated for each label for performance comparison between categories.

4. Experimental results and analysis

4.1. Overall comparison framework

This study focuses on the diagnosis task of 14 types of diseases in chest X-ray images and constructs two types of models: one with pre-training (a ResNet18 model based on SimCLR self-supervised pre-training, denoted as v3withamp2) and one without pre-training (a ResNet18 model trained from scratch, denoted as v3-1). Combining the model training loss curve and the F1 score and AUC graphs in Figure 2 and Figure 3, Table 1-3 and the bar chart in Figure 4 comparing the AUC and F1 score scores of each disease for the two types of models, through the F1 score (which combines precision and recall to measure the robustness of classification) and AUC (which reflects the ability

to distinguish between positive and negative samples, and the higher the value, the clearer the classification boundary) indicators, the diagnostic performance of the models in each category is deeply compared to explore the actual gains and limitations of the pre-training strategy.

4.2. Training curve and performance improvement

Figures 2 and 3 show the AUC and F1 curves of the pre-trained model and the non-pre-trained model on the Validation Set during the training process. It can be seen that the model with SimCLR pre-training exhibited higher AUC and F1 Score at the early stage of training and converged faster. Finally, the Average AUC Value of the pre-trained model was approximately 0.0251 (i.e., 2.51 percentage points) higher than that of the randomly initialized model, and the macro-average F1 Score was approximately 0.0304 (3.04 percentage points) higher. This improvement is consistent with the results of previous self-supervised studies. For example, Azizi et al. reported that in the chest X-ray classification task, the self-supervised pre-training model achieved an average AUC improvement of 1.1% compared to the ImageNet pre-trained baseline [8]. The experiment's findings also show how well Self-Supervised Learning works to enhance classification performance in the medical imaging domain (especially in the case of insufficient labels). In Figures 2 and 3, the Dynamic Threshold Optimization strategy continuously adjusts the thresholds of each label during the training process, causing the F1 score curve to fluctuate slightly but showing a smooth overall improvement, effectively enhancing the comprehensive indicators of Multi-label Classification.

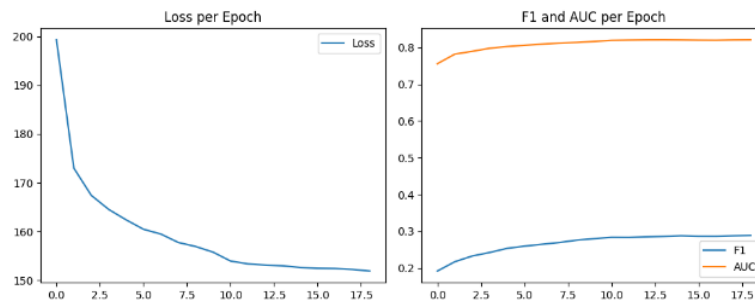


Figure 2. AUC and F1 curves of the pre-trained model on the validation set (picture credit: original)

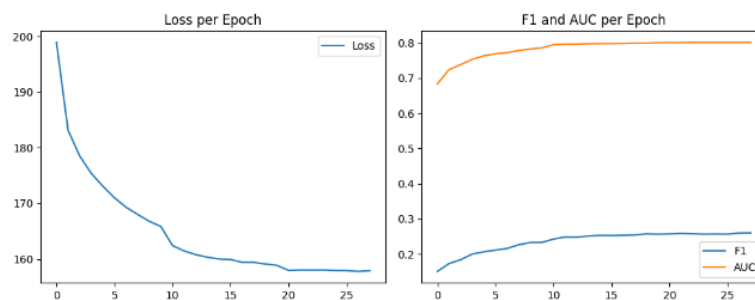


Figure 3. AUC and F1 curves on the validation set of the unpretrained model (picture credit: original)

4.3. Performance comparison by category

To understand the model's ability to recognize different diseases, the AUC and F1 scores of each disease label were calculated, and a bar graph shown in Figure 4 was plotted.

Table 1. Performance scores of each category for the model with pre-training

Class	F1-score	AUC	Class	F1-score	AUC
Atelectasis	0.37	0.7977	Pneumothorax	0.41	0.8826
Cardiomegaly	0.35	0.8994	Consolidation	0.22	0.8027
Effusion	0.53	0.8812	Edema	0.25	0.8904
Infiltration	0.41	0.7064	Emphysema	0.36	0.8924
Mass	0.37	0.8419	Fibrosis	0.13	0.7935
Nodule	0.23	0.7218	Pleural_Thickening	0.20	0.7882
Pneumonia	0.08	0.7402	Hernia	0.13	0.8520

Table 2. Performance scores of each category for the model without pre-training

Class	F1-score	AUC	Class	F1-score	AUC
Atelectasis	0.36	0.7884	Pneumothorax	0.38	0.8581
Cardiomegaly	0.35	0.9036	Consolidation	0.20	0.7929
Effusion	0.52	0.8721	Edema	0.26	0.8846
Infiltration	0.39	0.6974	Emphysema	0.26	0.8434
Mass	0.33	0.8264	Fibrosis	0.10	0.7717
Nodule	0.19	0.6826	Pleural_Thickening	0.20	0.7694
Pneumonia	0.07	0.7316	Hernia	0.03	0.7965

Table 3. Sample size distribution

Class	Sample size	Class	Sample size
Atelectasis	11559	Pneumothorax	5302
Cardiomegaly	2276	Consolidation	4667
Effusion	13317	Edema	2303
Infiltration	19894	Emphysema	2516
Mass	5782	Fibrosis	1686
Nodule	6331	Pleural_Thickening	3385
Pneumonia	1431	Hernia	227

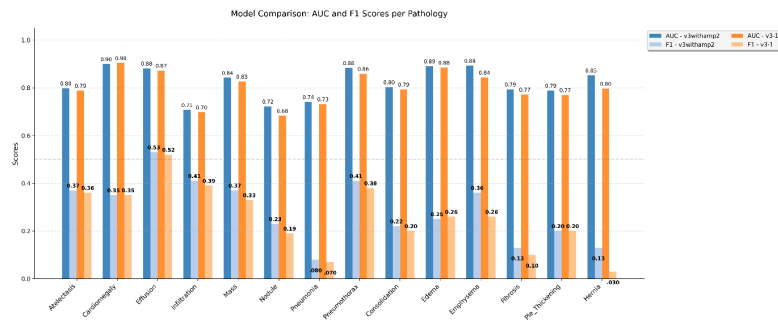


Figure 4. Comparison of AUC and F1 scores of the two types of models for each disease (picture credit: original)

4.3.1. High-sample supported categories (taking Effusion as an example)

The sample size of Effusion (fluid accumulation) reaches 13,317 cases (sample size distribution table), which is representative of high-sample categories in Dataset. Table 1 illustrates that the pre-training model's F1 score is 0.53, and the Area Under the Curve is 0.8812; Table 2 shows that the F1 score of the model without Pre-training is 0.52, and the Area Under the Curve is 0.8721. The bar chart also intuitively presents the differences in the indicators of the two types of models for this category.

The high sample size provides sufficient feature learning materials for the model. The model without Pre-training can also learn the typical imaging features of fluid accumulation (such as the gray - scale distribution and morphological boundaries of the fluid area) through repeated iterations to achieve relatively stable classification. The Pre-training model further strengthens the feature distinction between fluid accumulation and other similar diseases (such as Infiltration) by using the general features learned through the SimCLR framework, achieving a slight improvement in both F1 score and AUC. This verifies the “adaptability gain” of the high - sample categories to the Pre-training strategy. However, due to the high distinguishability of the features themselves, the performance gap between the two types of models is relatively small.

4.3.2. Low-sample challenge category (taking Hernia as an example)

The sample size of Hernia is only 227 cases (sample size distribution table), belonging to the category of low sample and high challenge. In Table 1, the Pre-training model has a score of 0.13 of F1 and an AUC of 0.8520; in Table 2, the non-Pre-training model has a F1 score of 0.03 and an AUC of 0.7965. The bar chart clearly shows the weakness of the performance of the non-Pre-training model in this category.

With a low sample size, when the non-Pre-training model learns from scratch, it is difficult to cover the diverse features of hernias (such as the morphological differences of hernias in different positions and the overlapping interference with surrounding tissues). The feature learning is seriously insufficient, resulting in a blurred classification boundary, and the F1 score, and AUC are at extremely low levels. Although the Pre-training model has not completely solved the problem of sample scarcity, the general features learned by SimCLR provide basic prior knowledge for the classification head, helping the model capture the subtle differences between hernias and normal tissues. The performance improvement is significant, highlighting the key value of Pre-training in "feature supplementation" for low-sample categories. At the same time, it also reflects the objective

bottleneck that the performance improvement of low-sample categories is limited by the sample size.

4.3.3. Complex and vague feature categories (taking Pneumonia as an example)

The sample size of Pneumonia is 1,431 cases (Sample Size Distribution Table). Moreover, the imaging features are interfered with by the infection morphology (such as ground-glass opacities and consolidation opacities) and the disease course stage, showing complex manifestations and highly overlapping with the features of other diseases (such as Infiltration). Table 1 shows that the Pre-training model has a score of 0.08 of F1 and an AUC of 0.7402; Table 2 presents that the non - Pre-training model has a score of 0.07 of F1 and the AUC is 0.7316. The bar chart also reflects the low performance of both types of models for this category.

The superposition of complex features and the relatively limited sample size greatly increases the difficulty of model learning. The non - Pre-training model lacks the foundation of general features, making it difficult to extract effective discriminative information from high - dimensional and vague imaging data, resulting in extremely poor classification results. Although the Pre-training model introduces general features, the pneumonia features have a low degree of fit with the general features learned through Pre-training, and there are obstacles in feature transfer and adaptation, resulting in only a slight improvement in performance. The poor performance of the two types of models in the pneumonia category verifies the "double negative impact" of feature complexity and insufficient sample size, which is a key area that needs to be addressed in model optimization.

4.3.4. Quantification and differences of the gain from the pre-training strategy

By comparing the changes in the indicators of the two types of models for each disease category in Tables 1 and 2, and with the assistance of the bar chart in Figure 4 for observation, the gain pattern of the Pre-training strategy can be quantified:

High-gain category: Taking Infiltration as an example, in Table 2, the F1 score-score of the model without Pre-training is 0.39 and the Area Under the Curve is 0.6974. In Table 1, the F1 score-score of the model with Pre-training increases to 0.41 and the Area Under the Curve increases to 0.7064. The bar chart also shows the improvement of the indicators for this category after Pre-training. The subtle features of infiltrative lesions (such as the infiltrative area with blurred edges) can enhance the capture ability through the general features of Pre-training. Pre-training effectively supplements the feature learning dimension and improves the classification accuracy. The learning difficulty of this category of features is moderate, and it has a good adaptability to the Pre-training features.

Low gain/negative gain categories: Cardiomegaly. In Tables 1 and 2, the score of F1 is 0.35 for both, and the AUC values are 0.8994 and 0.9036 respectively. The bar chart shows that the AUC of the model without Pre-training is slightly higher. The imaging features of cardiomegaly (such as the cardiac contour and size ratio) have high distinguishability. The model without Pre-training can already achieve stable recognition through direct learning. The general features of Pre-training do not optimize it. Instead, redundant information is introduced due to feature migration, resulting in a slight decrease in the AUC, which reflects that the gain of Pre-training is limited by the distinguishability of the inherent features of the category.

4.3.5. Key findings under multi-dimensional comparison

Combining the sample size distribution table, Table 1, Table 2, and the bar chart, the core findings can be summarized as follows:

Sample size dominates the basic performance: For high sample categories (such as Effusion), the performance of both types of models is relatively good. For low sample categories (such as Hernia), the performance is generally low, which verifies that the sample size is the basic support for model learning and directly affects the feature coverage and classification stability.

Adaptability differences of the Pre-training strategy: Pre-training has a significant gain for categories with low samples and medium difficulty in feature learning (such as Infiltration), but has limited or even negative gain for categories with complex and fuzzy features (such as Pneumonia) and categories with high inherent distinguishability (such as Cardiomegaly), which reflects that the Pre-training strategy needs to be deeply adapted to the category features. Direction guidance for model optimization: For low-performance categories (e.g., Pneumonia, Hernia), collaborative optimization is required from the data side (customized enhancement, sample synthesis), the model side (optimizing pre-trained feature transfer, lightweight classification head), and the training side (weighted loss, dynamic threshold adjustment) to break through the dual limitations of samples and features.

In summary, pre-trained models show performance advantages in most categories, verifying the value of SimCLR self-supervised pre-training. However, restricted by factors such as sample distribution and feature complexity, there are significant performance differences among different categories. In the future, attention should be focused on low-performance categories, and focused optimization techniques should be created to enhance the multi-label diagnostic model for chest X-ray pictures' generality and robustness.

4.4. Model interpretability analysis

4.4.1. Grad-cam visualization method and medical value

Grad-CAM calculates the gradient of the Feature Map through backpropagation to generate a heatmap of the model's decision, intuitively presenting the model's attention to image regions [9,10]. In the field of medical imaging, this technology can verify whether the model focuses on real lesions, avoiding situations of "attention to noise" or "misjudgment of artifacts", and it is a key bridge connecting algorithm performance and clinical credibility. This study visualizes the Grad-CAM of typical cases of the Pre-training model and the model without Pre-training and conducts an analysis from three dimensions: accuracy of lesion localization, consistency of feature attention, and attribution of category differences.

4.4.2. Visualization results and analysis of typical cases

High sample category: Effusion Visualization phenomenon: The heatmap of the Pre-training model is concentrated in the liquid density area along the lower edge of the thoracic cavity or the edge of the lung field, which highly coincides with the typical location of effusion marked by radiologists (Figure5a). Although the heatmap of the model without Pre-training covers the effusion area, the edge is slightly blurred, and there are occasional mis - focuses on rib shadows (Figure5b).

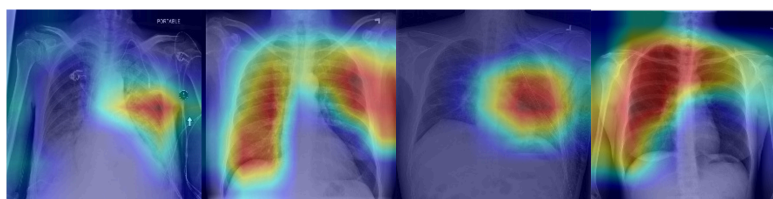


Figure 5. Grad-CAM visualization (a: Effusion with pre-training, b: Effusion without pre-training, c: Pneumonia with pre-training, d: Pneumonia without pre-training) (picture credit: original)

Quantitative association: The F1 score (0.53) and AUC (0.8812) of the Pre-training model on Effusion are better than those of the model without Pre-training (F1 score = 0.52, AUC = 0.8721). The precise positioning of the heatmap explains the performance advantage — the Pre-training features help the model capture the density gradient and morphological boundary of effusion more stably.

Complex feature category: Pneumonia Visualization phenomenon: The heatmaps of both types of models cover part of the ground-glass opacity area. However, the heatmap of the Pre-training model is more concentrated in the center of the consolidation (Figure5c), while the heatmap of the model without Pre-training widely covers the lung fields (Figure5d), and even includes normal lung markings in the area of concern.

Performance attribution: The imaging features of pneumonia (such as ground-glass opacity and air bronchogram) highly overlap with Infiltration. Although the Pre-training model enhances the sensitivity to density changes through general features, it is still difficult to fully distinguish subtle pathological differences. The model without Pre-training lacks the ability to screen features, resulting in "over-generalization" of the heatmap, which explains why its F1 score-score is only 0.07.

4.4.3. Clinical implications

Potential for reliable diagnosis: In well-defined categories such as Effusion and Pneumothorax, the model's heatmap is consistent with the doctor's diagnostic thinking, suggesting that the self-supervised pre-training model has the potential to assist in clinical decision-making.

Optimization direction: In the future, the attention mechanism (such as CBAM) can be combined to strengthen the weight of the lesion area, or the weakly supervised localization technology (WSL) can be introduced to refine the heatmap through label supervision, enhancing the model's capacity to identify intricate disease characteristics.

4.4.4. Visualization conclusion

Gradient-weighted Class Activation Mapping analysis shows that self-supervised pre-training significantly improves the consistency between the model's focus and medical semantics by optimizing the feature extraction mode, especially in categories with low samples and well-defined structural features. However, for diseases with vague features or extremely scarce samples (such as Pneumonia, Hernia), the model still has positioning deviations and needs to be further optimized by combining clinical prior knowledge. This analysis provides a clear direction for model iteration: by enhancing the learning ability of pathological textures and introducing anatomical structure constraints, the multi-label diagnosis model can be promoted to develop in a more clinically credible direction.

5. Conclusions

This paper studies a multi-label classification method for Chest X-ray based on SimCLR Self-Supervised Learning. The EfficientNet-B0 encoder is Pre-training on the large-scale Dataset of ChestX-ray14 and then transferred to the Downstream multi-label classification task. The experimental results show that compared with the randomly initialized model, the classification performance of the Pre-training model is significantly improved (the AUC is increased by about 2.51%, and the F1 score is increased by about 3.04%), which verifies the Effectiveness of Self-Supervised Learning in medical images. After incorporating Focal Loss, the model's ability to recognize minority sample categories is enhanced, alleviating the imbalance problem in multi-label task. The results of the Gradient-weighted Class Activation Mapping visual analysis show that the model can focus on disease-related regions, providing a certain degree of interpretability for the results. In future work, the authors plan to explore more self-supervised methods to further improve the quality of feature representation and consider multi-view or multi-modal inputs (such as combining CT images) to enrich the training information. Meanwhile, they will try more dynamic threshold adjustment strategies and label relationship modeling methods to further improve the classification performance of each category. In summary, this study provides an effective technical route for multi-label intelligent diagnosis of Chest X-ray, providing the opportunity to enhance the automated diagnosis system's functionality and interpretability.

References

- [1] Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., & 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 (CVPR 2017)* (pp. 3462–3471).
- [2] Chen, T., Kornblith, S., Norouzi, M., Hinton, G. (2020). A Simple Framework for Contrastive Learning of Visual Representations. *Proc. ICML*.
- [3] Zhou, Z., Sodha, V., Rahman Siddiquee, M. M., Feng, R., Tajbakhsh, N., Gotway, M. B., & Liang, J. (2020). Models genesis: Generic autodidactic models for 3D medical image analysis. *Medical Image Analysis*, 67, 101840.
- [4] Shurrab, S., & Duwairi, R. (2021). Self-supervised learning methods and applications in medical imaging analysis: A survey. *PeerJ Computer Science*, 8, e1045.
- [5] Tan, M., & Le, Q. V. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. In K. Chaudhuri & R. Salakhutdinov (Eds.), *Proceedings of the 36th International Conference on Machine Learning* (pp. 6105–6114). PMLR.
- [6] Lin, T.-Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (pp. 2999–3007).
- [7] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (pp. 618–626).
- [8] Azizi, S., Mustafa, B., Ryan, F., Fleet, D., & Norouzi, M. (2021). Big self-supervised models advance medical image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (pp. 3458–3468).
- [9] Chattopadhyay, A., Sarkar, A., Howlader, P., & Balasubramanian, V. N. (2018). Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)* (pp. 839–847). IEEE.
- [10] Baltrusaitis, T., Ahuja, C., & Morency, L. P. (2019). Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2), 423–443.