International Journal of Computer Techniques–IJCT Volume 13 Issue 4, July-August - 2026
Open Access and Peer Review Journal ISSN 2394-2231 https://ijctjournal.org/

RGCA-Swin-Tiny CXRNet: A Residual Gated Convolutional Attention Guided Swin Transformer for Chest X-ray Classification

Akshay Mool, Ashutosh Pandey, Rahul Kumar, Ankit Yadav
Assistant Professor, Delhi Technological University

Abstract—Automatic chest X-ray categorization is a crucial research topic for radiological screening and clinical decision-making. However, subtle disease-specific patterns, overlapping thoracic abnormalities, and considerable visual resemblance across chest illnesses make correct categorization difficult. RGCA-Swin-Tiny CXRNet, a hybrid deep learning model for chest X-ray classification, combines an ImageNet-pretrained Swin-Tiny Transformer with a lightweight convolutional attention branch to solve these issues. The convolutional attention branch catches local radiography signals such texture fluctuations, boundary changes, and opacity-like patterns, whereas the Swin-Tiny branch recovers hierarchical global contextual representations. The proposed model's **Residual Gated Convolutional Attention (RGCA)** method is novel, in which the convolutional descriptor creates a centered residual gate to adaptively suppress or boost the Swin-Tiny feature representation before classification. The suggested model is tested on a balanced five-class single-label ChestX-ray8 classification setting: Atelectasis, Hernia, No Finding, Pneumonia, and Pneumothorax. The dataset has 7500 images—5000 training, 1000 validation, and 1500 testing. Experimental findings demonstrate that RGCA-Swin-Tiny CXRNet has test accuracy of 64.67%, weighted precision of 64.48%, recall of 64.67%, F1-score of 64.44%, MCC of 0.5583, and AUC of 0.823. The confusion matrix and t-SNE visualization show that the proposed model learns meaningful discriminative representations, although visually overlapping classes like Atelectasis and No Finding remain difficult. Residual gated convolutional refinement is promising for chest X-ray classification using transformer-based global feature learning and local convolutional attention.

Keywords: Chest X-ray classification, Swin Transformer, convolutional attention, residual gated attention, medical image classification, deep learning.

---

1. Introduction

Due to its low cost, wide availability, and clinical significance, chest X-ray imaging is one of the most used thoracic illness diagnostic methods. Because some disorders have mild visual manifestations, overlapping radiographic patterns, and significant inter-class similarity, manual chest radiograph interpretation is difficult. Thus, deep learning-automated chest X-ray categorization is a key study area for radiological decision-making and preclinical illness screening.

Convolutional neural networks (CNNs) are commonly employed for medical picture categorization because they can capture local spatial patterns including edges, textures, and lesion-like features. Local characteristics in chest X-rays can reveal opacity, lung border alterations, and disease-specific texture abnormalities. CNNs may struggle to simulate long-term contextual interactions across anatomical areas. However, Vision Transformers and its derivatives can capture wider dependencies through self-attention methods, but they may underemphasize fine-grained local inductive biases that convolutional procedures describe.

This study offers RGCA-Swin-Tiny CXRNet, a hybrid model with an ImageNet-pretrained Swin-Tiny Transformer and a lightweight convolutional attention branch, to overcome this issue. Hierarchical transformer-based feature representation is provided by Swin-Tiny, while convolutional attention extracts supplementary local radiographic cues. The first patch-embedding layer of Swin-Tiny accepts one-channel input by averaging the pretrained RGB filters because chest X-ray pictures are grayscale. The other Swin Transformer levels keep their ImageNet initialization.

Residual Gated Convolutional Attention (RGCA) is the model's key innovation. Convolutional branch focused attention gate adaptively suppresses or increases final Swin feature dimensions. For five-class chest X-ray classification, a classifier receives the revised Swin representation and convolutional descriptor. This design preserves the transformer's global contextual power while adding local convolutional attention characteristics for radiographic illness classification.

This work's primary innovations and contributions are:

- **Hybrid Swin–convolutional architecture:** This study introduces RGCA-Swin-Tiny CXRNet, a lightweight hybrid model for chest X-ray classification that combines an ImageNet-pretrained Swin-Tiny Transformer with a convolutional attention branch.
- **Grayscale adaptation of pretrained Swin-Tiny:** The initial patch-embedding layer accommodates one-channel grayscale X-ray pictures by averaging the pretrained RGB filters while keeping the ImageNet transformer weights.
- **Residual Gated Convolutional Attention mechanism:** Local radiography signals drive a centered gate in a convolutional attention branch to adaptively suppress or augment Swin feature dimensions.
- **Complementary local-global feature learning:** Swin-Tiny's global contextual representation and convolutional local feature sensitivity for minor chest X-ray patterns are combined in the suggested architecture.
- **Five-class ChestX-ray8 evaluation:** The model is assessed on a five-class single-label ChestX-ray8 classification setup with Atelectasis, Hernia, No Finding, Pneumonia, and Pneumothorax utilizing accuracy,

precision, recall, F1-score, MCC, AUC, confusion matrix, and t-SNE visualization.

2. Literature Review

Clinical relevance of automated thoracic illness screening has led to much research on deep learning-based chest X-ray categorization. Baltruschat et al. [1] compared deep learning methods for multi-label chest X-ray classification, examining transfer learning, fine-tuning, network design, and non-image characteristics. Their work showed that ImageNet-pretrained CNNs can depict chest radiographs, although performance relies on picture resolution, training approach, and auxiliary information. Transfer learning was shown to be relevant for chest X-ray classification, although the model was mostly CNN-based and did not explicitly link convolutional local information with transformer-based global context.

Recently, attention-based CNN models have improved chest X-ray illness localization and discriminative feature learning. For multi-label chest X-ray classification, Guan and Huang [2] presented category-wise residual attention learning using attention maps to identify disease-specific discriminative areas. Chen et al. [3] proposed DualCheXNet, a dual asymmetric feature learning model for thoracic illness categorization using feature-level and decision-level learning. For multi-label thoracic illness detection, Chen et al. [4] introduced LLAGnet, a lesion location attention-guided network that incorporates region- and channel-level attention. These research indicate that attention processes can lead the model to disease-relevant areas, although they employ convolutional backbones and not a pretrained hierarchical transformer representation as the principal feature extractor.

More spatial, semantic, and disease-specific modeling has been added to attention-based chest X-ray categorization. Wang et al. [5] suggested a triple attention learning framework to diagnose 14 thoracic illnesses from chest radiographs using various attention methods to enhance feature discrimination. Pixel-wise classification and attention network PCAN was developed by Zhu et al. [8] for thoracic illness classification and poorly guided localization. In the absence of pixel-level illness annotations, fine-grained attention is useful for weakly supervised chest X-ray processing. These approaches mostly improve CNN-based feature mapping or localization, while utilizing convolutional attention to directly optimize transformer-derived feature embeddings is rarely investigated.

Graph-based models of image-disease label connections have been studied alongside attention-based CNN models. ImageGCN, a multi-relational image graph convolutional network for chest X-ray illness diagnosis, was proposed by Mao et al. [6]. Their method modeled visual associations to improve illness prediction. Lee et al. [7] presented CheXGAT, a disease correlation-aware graph attention network for multi-label chest X-ray diagnosis using implicit thoracic illness correlations. These graph-based algorithms emphasize relational thinking in thoracic illness categorization. However, they complicate graph modeling and do not directly address local convolutional cue fusing with transformer-based global feature representations.

Because they simulate long-range dependencies, Vision Transformer-based medical picture categorization models are gaining popularity. A robust Vision Transformer for generic medical image classification, MedViT, was suggested by Manzari et al. [9] using convolutional inductive bias and transformer-style global modeling. Wu et al. [10] suggested CTransCNN for multi-label medical picture categorization using Transformer and CNN. These results imply CNN–Transformer hybridization is promising since CNNs are good at local pattern extraction and Transformers for contextual representation learning. Many hybrid models use parallel feature extraction, direct concatenation, or generic CNN–Transformer integration instead of a convolutional attention branch to provide a residual gate for adaptively refining pretrained transformer features.

Reviewing the literature reveals significant research gaps. CNN-based chest X-ray models may capture local texture and lesion-like patterns, but they may struggle to simulate contextual connections. Second, transformer-based models can capture global dependencies but may understate radiography abnormality-related fine-grained convolutional inductive biases. Third, attention-guided approaches enhance CNN features or enable weak localization without modulating transformer embeddings with convolutional attention. Fourth, CNN–Transformer hybrid models often employ feature concatenation or generic hybrid blocks, but a lightweight convolutional branch to provide a centered residual gate for suppressing or augmenting pretrained Swin Transformer features is underexplored. These gaps inspired the RGCA-Swin-Tiny CXRNet, which uses residual gated feature refinement to merge a grayscale-adapted ImageNet-pretrained Swin-Tiny branch with a convolutional attention branch.

3. Proposed Methodology

The Residual Gated Convolutional Attention Swin-Tiny Network (RGCA-Swin-Tiny CXRNet) is presented here. The model classifies grayscale chest X-rays into five classes. ImageNet-pretrained Swin-Tiny Transformers and lightweight convolutional attention branches are used in the proposed architecture. Swin-Tiny recovers hierarchical transformer-based contextual features, whereas convolutional attention extracts supplementary local radiography information. The model's main innovation is employing the convolutional attention branch to build a residual feature-wise gate that adaptively refines the Swin-Tiny feature representation.

3.1 Overview of the Proposed RGCA-Swin-Tiny CXRNet

The model has two parallel feature extraction branches. Main contextual feature extractor Swin-Tiny Transformer is the first branch. The second branch captures local radiographic patterns such texture changes, boundary information, and opacity-like structures using lightweight convolutional attention. Before concatenating the two branches' outputs, the convolutional branch creates a feature-wise attention gate that modulates the final Swin feature vector.

Let the input grayscale chest X-ray image be denoted by X. For a batch of Nimages, the input tensor is represented as:

$$X \in \mathbb{R}^{N\times 1\times 256\times 256}$$

The Swin-Tiny branch extracts a final transformer feature vector z_s:

$$z_s = f_{\text{swin}}(X)$$

where:

$$z_s \in \mathbb{R}^{768}$$

In parallel, the convolutional attention branch extracts a local convolutional descriptor z_c:

$$z_c = f_{\text{conv}}(X)$$

where:

$$z_c \in \mathbb{R}^{128}$$

The final prediction is obtained by refining $z_s$using a residual gate generated from z_c, followed by concatenation of the refined Swin feature and the convolutional descriptor. The overall process can be summarized as:

$$z_f = \text{Fusion}(z_s, z_c)$$
$$\hat{y} = \text{Classifier}(z_f)$$

where $\hat{y}$ denotes the output logits corresponding to the five chest X-ray classes.

3.2 Grayscale-Adapted Swin-Tiny and Convolutional Attention Branch

Main branch of suggested model based on Swin-Tiny. Swin-Tiny's initial patch-embedding convolution requires three input channels because it was pretrained on ImageNet with three-channel RGB pictures. X-rays of the chest are inherently grayscale. Swin-Tiny's initial patch-embedding layer accepts one-channel input instead of three-channel repeated images for grayscale radiographs.

The initial pretrained RGB patch-embedding weights are:

$$W_{\text{RGB}} \in \mathbb{R}^{96\times3\times4\times4}$$

The pretrained RGB filters are averaged across the three input channels to adjust this layer for grayscale input:

$$W_{\text{gray}} = \frac{W_R + W_G + W_B}{3}$$

This yields grayscale patch-embedding weights:

$$W_{\text{gray}} \in \mathbb{R}^{96\times1\times4\times4}$$

This adaption lets the model process one-channel chest X-rays while retaining ImageNet pretraining. This modifies only the first patch-embedding convolution. All Swin Transformer layers keep their ImageNet-pretrained weights. Final transformer representation z_s with 768 dimensions comes from Swin-Tiny. In parallel, the convolutional attention branch creates a 128-dimensional descriptor z_c. Convolutional branch includes convolutional stem, conventional convolutional blocks, depthwise separable convolution block, channel attention, spatial attention, and global average pooling.

Convolutional branch provides local information that transformer branch may not capture. Channel attention prioritizes diagnostically helpful feature channels, while spatial attention focuses instructive visual areas. Our convolutional branch may focus on local radiographic patterns including lung border alterations, aberrant opacity, and texture irregularity.

3.3 Residual Gated Convolutional Attention Fusion and Classification

Residual Gated Convolutional Attention is the model's main innovation. CNN and Transformer characteristics are concatenated in several hybrid CNN–Transformer models. The suggested approach generates a feature-wise gate from the convolutional descriptor to modify the Swin-Tiny representation before classification.

First, the convolutional descriptor z_c is projected from 128 dimensions to 768 dimensions using a linear transformation:

$$a_c = W_g z_c + b_g$$

where:

$$W_g \in \mathbb{R}^{768\times128}, b_g \in \mathbb{R}^{768}$$

The projected vector a_c is passed through a sigmoid function and then centered to obtain a gate in the range $[-1,1]$:

$$g_c = 2\sigma(a_c) - 1$$

Therefore:

$$g_c \in [-1,1]^{768}$$

This centered gate suppresses and enhances Swin feature dimension If a component of g_c is positive, the corresponding Swin feature is enhanced.

The residual gated refinement is defined as:

$$z_r = z_s \odot (1 + \alpha g_c)$$

where z_r is the refined Swin feature, z_s is the original Swin feature, $g_c$is the centered convolutional attention gate, $\odot$denotes element-wise multiplication, and αcontrols the strength of gated modulation.

In this work, α is set to 0.5. Therefore, the effective scaling factor becomes:

$$1 + 0.5g_c$$

Since g_c lies in the range $[-1,1]$, the effective scaling range becomes:

$$1 + 0.5g_c \in [0.5,1.5]$$

Each Swin feature dimension can be muted or boosted by 50%.The pretrained Swin-Tiny representation is preserved while local convolutional attention cues modify it with this slight residual modulation.

After residual refinement, the refined Swin feature $z_r$is concatenated with the convolutional descriptor z_c:

$$z_f = \text{concat}(z_r, z_c)$$

where:

$$z_r \in \mathbb{R}^{768}, z_c \in \mathbb{R}^{128}$$

Therefore:

$$z_f \in \mathbb{R}^{896}$$

The fused feature vector $z_f$is passed through a lightweight multilayer classifier:

$$h = \text{ReLU}(W_1 z_f + b_1)$$
$$\hat{y} = W_2 h + b_2$$

where:

$$W_1 \in \mathbb{R}^{256\times896}, W_2 \in \mathbb{R}^{5\times256}$$

The final output is:

$$\hat{y} \in \mathbb{R}^5$$

corresponding to the five classes: Atelectasis, Hernia, No Finding, Pneumonia, and Pneumothorax.

Combining convolutional and transformer characteristics is regulated and interpretable with the RGCA technique. The convolutional attention branch refines the Swin representation by residual gating, not only extracting features. The model can classify chest X-rays using transformer-based global contextual representation and convolutional local radiographic sensitivity.

4. Experimental Setup

This section outlines the experimental setup for RGCA-Swin-Tiny CXRNet evaluation. The last experiment in this publication employs α = 0.5 residual gating strength. Next section results and comments are based only on this experimental configuration.

4.1 Dataset Description

A portion of the NIH ChestX-ray8 dataset was used for the investigation. The original dataset had several thoracic illness labels, however this study uses a five-class single-label classification issue. Selected classes are:

1. Atelectasis

2. Hernia

3. Nothing Found

4. Pneumonia

5. Pneumothorax

This experiment used just single-label images. To make the job a multi-class classification problem, images with multiple illness classifications were removed. A balanced dataset of 7500 chest X-ray pictures was created from 1500 images from each class.

The experiment's class-label mapping was:

Class	Label
Atelectasis	0
Hernia	1
No Finding	2
Pneumonia	3
Pneumothorax	4

A balanced five-class dataset reduces class imbalance during model training and assessment. Note that this experimental configuration differs from the entire multi-label ChestX-ray14 classification job.

4.2 Data Preprocessing and Augmentation

Grayscale chest X-rays were all processed. The suggested model alters Swin-Tiny's initial patch-embedding layer to accept one-channel input. Therefore grayscale pictures were not reproduced into three RGB channels. Image inputs were transformed to a single-channel format and scaled to 256 × 256 pixels.

The input tensor for a batch of images was therefore represented as:

$$X \in \mathbb{R}^{N\times1\times256\times256}$$

where Ndenotes the batch size.

The following preprocessing and augmentation operations were applied:

Operation	Description
Grayscale conversion	Images were converted to one-channel grayscale format.
Resizing	Each image was resized to 256 × 256 pixels.
Random horizontal flip	Applied with probability 0.5 during training.
Random vertical flip	Applied with probability 0.5 during training.
Tensor conversion	Images were converted into PyTorch tensors.

The model was exposed to spatial differences in the training data during augmentation operations to increase generalization.

4.3 Train-Validation-Test Split

The balanced dataset has 7500 images. The dataset has training, validation, and testing subsets:

Subset	Number of Images
Training set	5000
Validation set	1000
Test set	1500
Total	7500

The training set learned model parameters, the validation set monitored training, and the test set evaluated performance.

4.4 Model Configuration

Grayscale-adapted Swin-Tiny Transformer and convolutional attention branches make up the proposed RGCA-Swin-Tiny CXRNet.

The Swin-Tiny branch started using ImageNet-pretrained weights. Since Swin-Tiny originally expected three-channel RGB input, its initial patch-embedding convolution was changed to allow one-channel grayscale. The pretrained RGB filters were averaged over the channel dimension. Other Swin-Tiny layers kept their ImageNet-pretrained initialization.

The convolutional attention branch extracted complementing local radiographic information. Convolutional layers, depthwise separable convolution blocks, channel, spatial, and global average pooling were included. This branch yielded a 128-dimensional convolutional descriptor.

The convolutional attention branch generated a 128-dimensional descriptor, whereas Swin-Tiny produced a 768-dimensional feature vector. A residual gate was created using the convolutional descriptor to refine Swin feature representation. The final experiment's residual gating strength was:

$$\alpha = 0.5$$

The residual gated refinement was performed as:

$$z_r = z_s \odot (1 + 0.5 g_c)$$

where $z_s$is the Swin feature vector, g_c is the convolutional attention gate, and $z_r$is the refined Swin feature vector.

Concatenating the 768-dimensional refined Swin feature with the 128-dimensional convolutional descriptor produced a fused feature vector of size: $768 + 128 = 896$.

A lightweight classifier with a fully connected layer, ReLU activation, dropout, and output layer processed this fused 896-dimensional feature vector. Finally, the output layer produced logits for the five chest X-ray classes.

4.5 Training Configuration

Supervised learning with cross-entropy loss trained the model. Each image has one class label since the objective was a five-class single-label classification issue.

The training configuration used in the finalized experiment is summarized below:

Setting	Value
Model	RGCA-Swin-Tiny CXRNet
Input size	1 × 256 × 256
Number of classes	5
Batch size	128
Number of epochs	25
Loss function	Cross-Entropy Loss
Optimizer	Adam
Learning rate	0.0001
Scheduler	StepLR
Step size	1
Gamma	0.8
Gating strength	α = 0.5
Dropout in classifier	0.3

0.0001 was the initial learning rate for the Adam optimizer. A StepLR scheduler with a decay factor of 0.8 reduced learning rate after each epoch. This model was trained for 25 epochs.

The loss function used for optimization was:

$$L = \text{CrossEntropy}(\hat{y}, y)$$

where $\hat{y}$ represents the predicted logits and y represents the ground-truth class label.

4.6 Evaluation Metrics

The 1500-image independent test set assessed the trained model. A complete categorization performance assessment employed many performance measures.

Following metrics were used:

Metric	Purpose
Accuracy	Measures the overall proportion of correctly classified test samples.
Precision	Measures the correctness of positive predictions.
Recall	Measures the ability to identify samples of each class.
F1-score	Harmonic mean of precision and recall.
Matthews Correlation Coefficient	Measures balanced classification quality across classes.
AUC	Measures class separability using probability scores.
Confusion matrix	Shows class-wise correct and incorrect predictions.
t-SNE visualization	Visualizes the separability of learned feature representations.

For each class's sample count, weighted precision, recall, and F1-score were calculated. We also utilized the Matthews Correlation Coefficient since it evaluates multi-class classification performance more fairly than accuracy alone.

The residual-refined Swin feature and convolutional descriptor were merged to create the t-SNE visualization using the final fused feature representation z_f. The suggested model was qualitatively tested for separable feature clusters for distinct chest X-ray classes using this visualization.

5. Experimental Results and Discussion

Experimental results of the proposed RGCA-Swin-Tiny CXRNet on the five-class single-label ChestX-ray8 classification challenge are presented here. The data presented here are from the finished experiment with residual gating strength α = 0.5. Overall performance measures, class-wise classification, confusion matrix analysis, and t-SNE-based feature visualization are used.

5.1 Training and Validation Performance

Figure 1 shows the proposed model's training and validation curves. The graphic shows 25-epoch accuracy and loss patterns. Validation accuracy improved and matched training performance as learning progressed. This shows that the model learned meaningful discriminative representations from chest X-ray pictures without overfitting.

These findings are supported by loss curves. Training loss dropped consistently throughout epochs, indicating steady model optimization. Although modest changes occurred, validation loss decreased. Because multiple thoracic illness groups have overlapping visual patterns and modest radiographic distinctions, chest X-ray categorization fluctuates. Training and validation results indicate that the residual gated convolutional attention technique supported robust feature learning.

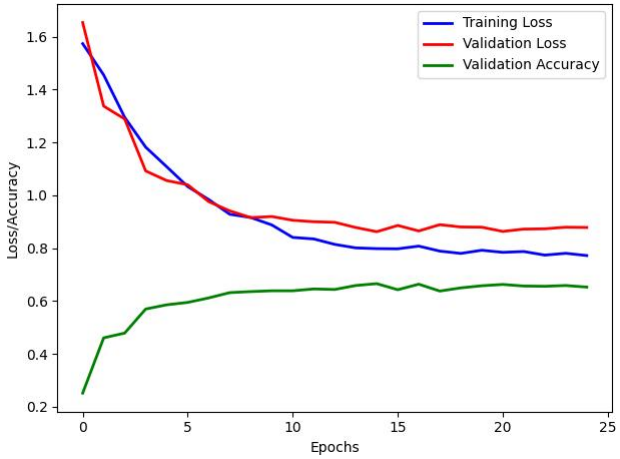

Figure 1. Training and validation accuracy and loss curves of the proposed RGCA-Swin-Tiny CXRNet over 25 epochs.

5.2 Overall and Class-wise Classification Performance

Overall test performance of the suggested model is shown in Table 1. RGCA-Swin-Tiny CXRNet scored 64.67% on 1500 independent test images. The weighted accuracy, recall, and F1-score were close to 64%, showing that the model balanced categorization across the five classes. The model's MCC score of 0.5583 indicates a moderate positive correlation between predicted and actual class labels.

Table 1. Overall performance of RGCA-Swin-Tiny CXRNet

Metric	Value
Test Accuracy	64.67%
Weighted Precision	64.48%
Weighted Recall	64.67%
Weighted F1-score	64.44%
MCC	0.5583
AUC	0.8236

Performance by class is shown in Table 2. Hernia had the best classification performance among the five classes, with 92.62% accuracy, 94.95% recall, and 93.77% F1-score. This suggests that the learnt fused representation discriminated this class well in the experimental context. With a recall of 70.21% and F1-score of 65.29%, pneumonia performed well. The F1-score of 62.85% for pneumothorax indicates modest classification ability.

Atelectasis and No Finding were harder to classify. The lowest F1-score was 44.58% for No Finding and 51.85% for Atelectasis. This decreased performance may be due to visual overlap between normal and abnormal radiographs and modest thoracic anomalies. The No Finding class is difficult since it reflects no disease patterns rather than an aberrant structure.

Table 2. Class-wise performance analysis of RGCA-Swin-Tiny CXRNet

Class	Precision	Recall	F1-score	Support
Atelectasis	51.27%	52.44%	51.85%	307
Hernia	92.62%	94.95%	93.77%	317
No Finding	46.64%	42.69%	44.58%	260
Pneumonia	61.01%	70.21%	65.29%	292
Pneumothorax	66.90%	59.26%	62.85%	324

The hybrid model appears to acquire relevant local-global feature representations based on its performance trend. Swin-Tiny gives global contextual modeling, while convolutional attention provides local radiography sensitivity. More adaptable than direct feature concatenation, the residual gated technique suppresses or enhances transformer characteristics before classification using local convolutional cues.

5.3 Confusion Matrix and Feature Space Visualization

Figure 2 shows the suggested model's confusion matrix. The matrix shows class-wise prediction behavior more clearly. Most hernia test samples were successfully categorized, with 301 rightly classified out of 317. This confirms the strong class-wise performance observed in Table 2.

Atelectasis, No Finding, Pneumonia, and Pneumothorax are also misclassified in the confusion matrix. Atelectasis was occasionally misdiagnosed as pneumonia and no finding. Atelectasis, pneumonia, and pneumothorax were also mistaken with No Finding. These misclassifications are clinically comprehensible since certain radiographic abnormalities are subtle and distinct thoracic illnesses may have comparable visual features such opacity, lung texture variation, or aberrant air-space patterns.

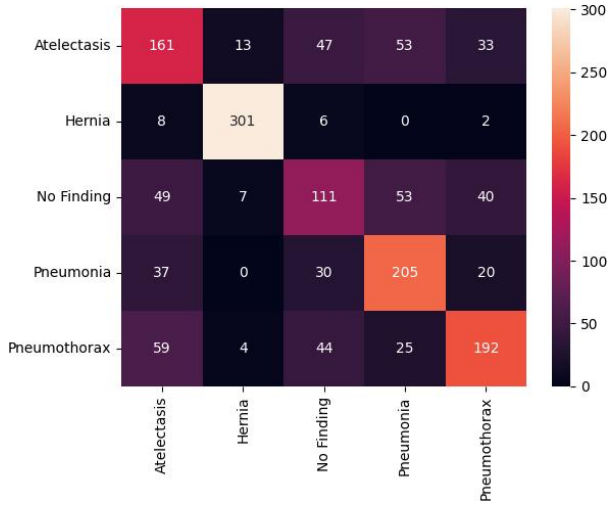

Figure 2. Confusion matrix of the proposed RGCA-Swin-Tiny CXRNet on the test set.

Figure 3 shows t-SNE visualization of learnt fused feature representations. Residual gated refinement and concatenation produced the fused feature vector for the t-SNE visualization. Visualization demonstrates the model learnt partly separable feature clusters for five classes. The Hernia class has good accuracy, recall, and F1-score and seems more distinct than the others.

The t-SNE figure also shows overlap between Atelectasis, No Finding, Pneumonia, and Pneumothorax. This overlap supports confusion matrix analysis and suggests these classes have fewer separable feature distributions in learnt representation space. Because many thoracic disorders have similar visual features, chest X-ray categorization should overlap.

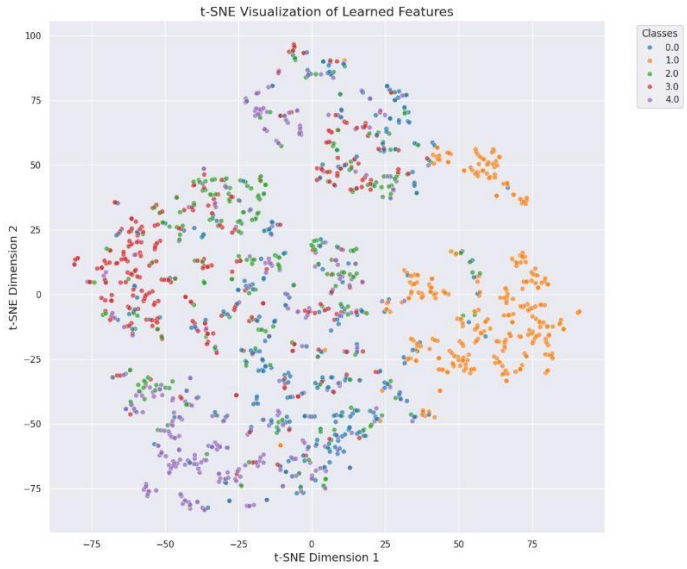

Figure 3. t-SNE visualization of the fused feature representations learned by the proposed RGCA-Swin-Tiny CXRNet.

The experiments show that RGCA-Swin-Tiny CXRNet can learn meaningful discriminative representations for five-class chest X-ray classification. This innovative residual gating strategy for pretrained Swin-Tiny features utilizing a convolutional attention branch is also supported by the results. The proposed model adaptively modulates transformer representation using local convolutional attention instead of concatenating CNN and Transformer features. The remaining class overlap and misclassification patterns imply patient-wise splitting, larger-scale multi-label training, better data augmentation methodologies, and comparative ablation studies may enhance performance.

6. Conclusion and Future Research Directions

This research proposes RGCA-Swin-Tiny CXRNet, a hybrid deep learning model for five-class chest X-ray classification. ImageNet-pretrained Swin-Tiny Transformers and lightweight convolutional attention branches are used. The Swin-Tiny branch collects hierarchical global contextual representations, while the convolutional attention branch extracts supplementary local radiography characteristics. The Residual Gated Convolutional Attention (RGCA) technique, where the convolutional descriptor creates a centered residual gate to adaptively suppress or boost the Swin-Tiny feature representation before classification, is the fundamental innovation.

The trial used a balanced five-class single-label ChestX-ray8 configuration with Atelectasis, Hernia, No Finding, Pneumonia, and Pneumothorax. The model has test accuracy of 64.67%, weighted precision of 64.48%, recall of 64.67%, F1-score of 64.44%, MCC of 0.5583, and AUC of 0.8236. The class-wise study indicated that the model performed best for Hernia, although visual overlap and slight radiographic changes made Atelectasis and No Finding harder. The confusion matrix and t-SNE visualization demonstrated that the model learnt valid feature representations, yet several illness categories had overlapping feature distributions.

Transformer-based contextual modeling and convolutional attention-guided local feature refinement show promise for chest X-ray categorization. The proposed RGCA technique lets local convolutional cues dynamically control the pretrained Swin-Tiny representation, unlike CNN–Transformer feature concatenation. This allows radiographic image processing to combine local and global information with controlled and adaptive fusion.

Despite these promising outcomes, this study has drawbacks. The experiment started using a five-class single-label subset of ChestX-ray8 instead of the full multi-label setup. Second, the study employed a balanced selection of pictures, which may not reflect real-world radiology dataset class imbalance. Third, while the proposed model shows beneficial feature learning, baseline comparisons and ablation experiments would increase the empirical validity of the RGCA process.

The suggested approach may be expanded to multi-label ChestX-ray14 classification in the future. Use patient-wise data separation to improve clinical validity and prevent data leakage. Ablation of the convolutional attention branch and residual gate, comparison with CNN, Transformer, and CNN–Transformer baselines, and assessment on CheXpert or MIMIC-CXR may increase performance. To improve clinical interpretability, Grad-CAM, attention map visualization, and class activation analysis can be used. The model may also learn the residual gating intensity to automatically calculate the appropriate convolutional attention-based feature modulation during training.

References

[1] I. M. Baltruschat, H. Nickisch, M. Grass, T. Knopp, and A. Saalbach, "Comparison of deep learning approaches for multi-label chest X-ray classification," *Scientific Reports*, vol. 9, no. 1, Art. no. 6381, 2019.

[2] Q. Guan and Y. Huang, "Multi-label chest X-ray image classification via category-wise residual attention learning," *Pattern Recognition Letters*, vol. 130, pp. 259–266, 2020.

[3] B. Chen, J. Li, X. Guo, and G. Lu, "DualCheXNet: Dual asymmetric feature learning for thoracic disease classification in chest X-rays," *Biomedical Signal Processing and Control*, vol. 53, Art. no. 101554, 2019.

[4] B. Chen, J. Li, G. Lu, and D. Zhang, "Lesion location attention guided network for multi-label thoracic disease classification in chest X-rays," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 7, pp. 2016–2027, 2020.

[5] H. Wang, S. Wang, Z. Qin, Y. Zhang, R. Li, and Y. Xia, "Triple attention learning for classification of 14 thoracic diseases using chest radiography," *Medical Image Analysis*, vol. 67, Art. no. 101846, 2021.

[6] C. Mao, L. Yao, and Y. Luo, "ImageGCN: Multi-relational image graph convolutional networks for disease identification with chest X-rays," *IEEE Transactions on Medical Imaging*, vol. 41, no. 8, pp. 1990–2003, 2022.

[7] Y. -W. Lee, S.-K. Huang, and R.-F. Chang, "CheXGAT: A disease correlation-aware network for thorax disease diagnosis from chest X-ray images," *Artificial Intelligence in Medicine*, vol. 132, Art. no. 102382, 2022.

[8] X. Zhu, S. Pang, X. Zhang, J. Huang, Z. Lu, and Y. Feng, "PCAN: Pixel-wise classification and attention network for thoracic disease classification and weakly supervised localization," *Computerized Medical Imaging and Graphics*, vol. 102, Art. no. 102137, 2022.

[9] O. N. Manzari, H. Ahmadabadi, H. Kashiani, S. B. Shokouhi, and A. Ayatollahi, "MedViT: A robust vision transformer for generalized medical image classification," *Computers in Biology and Medicine*, vol. 157, Art. no. 106791, 2023.

[10] X. Wu, H. Chen, J. Wang, H. Troiano, V. Loia, and H. Fujita, "CTransCNN: Combining Transformer and CNN in multilabel medical image classification," *Knowledge-Based Systems*, vol. 281, Art. no. 111030, 2023.