RepViT-CXR: A Channel Replication Strategy for Vision Transformers in Chest X-ray Tuberculosis and Pneumonia Classification

Faisal Ahmed* orcid.png AHMEDF9@ERAU.EDU
[1] *Department of Data Science and Mathematics*
[2] *Embry-Riddle Aeronautical University, Prescott, Arizona, USA .*

Editors: Under Review for MIDL 2026

Abstract

Chest X-ray (CXR) imaging remains one of the most widely used diagnostic tools for detecting pulmonary diseases such as tuberculosis (TB) and pneumonia. Recent advances in deep learning, particularly Vision Transformers (ViTs), have shown strong potential for automated medical image analysis. However, most ViT architectures are pretrained on natural images and require three-channel inputs, while CXR scans are inherently grayscale. To address this gap, we propose **RepViT-CXR**, a channel replication strategy that adapts single-channel CXR images into a ViT-compatible format without introducing additional information loss.

We evaluate RepViT-CXR on three benchmark datasets. On the **TB-CXR dataset**, our method achieved an accuracy of **99.9%** and an AUC of **99.9%**, surpassing prior state-of-the-art methods such as Topo-CXR (99.3% accuracy, 99.8% AUC). For the **Pediatric Pneumonia dataset**, RepViT-CXR obtained **99.0% accuracy**, with **99.2% recall**, **99.3% precision**, and an AUC of **99.0%**, outperforming strong baselines including DCNN and VGG16. On the **Shenzhen TB dataset**, our approach achieved **91.1% accuracy** and an AUC of **91.2%**, marking a performance improvement over previously reported CNN-based methods. These results demonstrate that a simple yet effective channel replication strategy allows ViTs to fully leverage their representational power on grayscale medical imaging tasks. RepViT-CXR establishes a new state of the art for TB and pneumonia detection from chest X-rays, showing strong potential for deployment in real-world clinical screening systems.

Keywords: Chest X-ray, Vision Transformer, Channel Replication, Tuberculosis, Pneumonia,Medical Image Analysis

1. Introduction

Chest X-ray (CXR) imaging is a widely used, non-invasive diagnostic tool for detecting pulmonary diseases such as tuberculosis (TB) and pneumonia (Jaeger et al., 2013; Kermany et al., 2018b). Despite its clinical importance, automated analysis of CXRs remains challenging due to the limited availability of annotated datasets, inter-patient variability, and subtle radiographic patterns that are difficult to detect using conventional methods (Pasa et al., 2019; Meraj et al., 2019).

Recent advancements in deep learning, particularly Convolutional Neural Networks (CNNs), have shown significant improvements in automated disease detection from CXR

* Contributed equally

images (Rahman et al., 2020b; Rajaraman et al., 2018). However, CNNs typically focus on local features and often struggle to capture global contextual information, which can be crucial for accurate diagnosis (Ahmed et al., 2023). Additionally, many state-of-the-art models require large datasets for training, limiting their applicability in scenarios with scarce medical data (Hernández et al., 2019).

Vision Transformers (ViTs) have emerged as a powerful alternative for image classification tasks, as they model long-range dependencies and global context effectively (Dosovitskiy et al., 2021; Liu et al., 2021). Yet, ViTs are typically pretrained on RGB natural images, making it non-trivial to apply them directly to grayscale medical images such as CXRs. Common solutions, such as duplicating the single grayscale channel to create a pseudo-RGB input or training from scratch, often lead to suboptimal performance, especially with limited labeled data (Toğaçar et al., 2020).

To overcome these limitations, we propose **RepViT-CXR**, a channel replication strategy that adapts grayscale CXRs for pretrained ViTs without extensive retraining. By leveraging the global attention mechanism of Transformers while preserving the integrity of the original X-ray information, RepViT-CXR achieves state-of-the-art performance in TB and Pneumonia classification across multiple benchmark datasets (Kermany et al., 2018b; Rahman et al., 2020b; Jaeger et al., 2014). This approach provides a practical, scalable, and highly accurate solution for automated chest disease diagnosis, addressing both the data scarcity and model adaptation challenges inherent in medical imaging.

Our contributions.

- We propose **RepViT-CXR**, a novel channel replication strategy that adapts single-channel grayscale chest X-rays for pretrained Vision Transformers without introducing additional information loss.
- RepViT-CXR leverages the global attention mechanism of Transformers to capture long-range dependencies in CXRs, overcoming the limitations of CNNs that primarily focus on local features.
- We demonstrate the effectiveness of RepViT-CXR across three benchmark datasets (TB-CXR, Pediatric Pneumonia, and Shenzhen TB), achieving state-of-the-art performance in terms of accuracy, precision, recall, F1-score, and AUC.
- Our approach addresses the challenge of limited labeled medical data by efficiently adapting pretrained ViTs without extensive retraining, making it practical for real-world clinical deployment.
- We provide a scalable and highly accurate solution for automated chest disease diagnosis, highlighting the potential of Transformers for grayscale medical imaging tasks.

2. Related Works

Chest X-ray (CXR) analysis has become a cornerstone for diagnosing respiratory diseases such as tuberculosis (TB) and pneumonia. Traditional machine learning methods relied heavily on handcrafted features and statistical models. For instance, F-SVM (Jaeger et al.,

2013) and classical CNN-based approaches (Hwang et al., 2016) demonstrated early success in TB detection but were limited in feature generalization and scalability.

With the advent of deep learning, Convolutional Neural Networks (CNNs) became the dominant paradigm. Models like sCNN (Pasa et al., 2019), VGG16 (Meraj et al., 2019), and DCNN (Rahman et al., 2020b) achieved high accuracy in both TB and pneumonia screening. Ensemble-based CNNs (E-CNN) (Hernández et al., 2019) further improved performance by combining multiple architectures. However, these methods often require large annotated datasets and are sensitive to data imbalance, leading to suboptimal generalization across diverse patient populations.

Recent approaches have explored topological data analysis techniques to enhance feature extraction. Topo-CXR (Ahmed et al., 2023) used topological information into machine learning model, achieving improved performance on TB datasets. Similarly, feature selection methods such as mRMR (Toğaçar et al., 2020) have been employed for pneumonia detection to reduce irrelevant features and improve model interpretability. While these approaches advance diagnostic accuracy, they still face limitations in model efficiency and robustness, particularly on smaller datasets. Furthermore, More techniques involving topological data analysis (TDA) are also widely used in classification tasks (Ahmed et al., 2025; Ahmed and Coskunuzer, 2023; Ahmed, 2023; Ahmed et al., 2023; Yadav et al., 2023; Ahmed and Bhuiyan, 2025b).

Vision transformers (ViTs) (Dosovitskiy et al., 2021; Liu et al., 2021) have recently emerged as a promising alternative to CNNs by leveraging global self-attention mechanisms. They capture long-range dependencies more effectively than convolutional layers, offering superior feature representation for medical imaging tasks. Nonetheless, their direct application in CXR analysis is still limited, primarily due to high computational cost and the need for large-scale data for training. More applications of transfer learning and Vision Transformers in medical image analysis are explored in the following studies: (Ahmed and Uddin, 2025; Ahmed, 2025b; Ahmed and Bhuiyan, 2025a; Ahmed, 2025a,c).

To address these challenges, we propose **RepViT-CXR**, a hybrid model combining the efficiency of CNNs with the global context modeling capability of vision transformers. Our model achieves state-of-the-art performance in TB and pneumonia diagnosis across multiple datasets, including TB-CXR, Shenzhen, and Ped-Pneumonia, while maintaining robustness on moderately sized datasets. By integrating hierarchical feature extraction with attention-based global reasoning, RepViT-CXR overcomes both the data-efficiency and generalization limitations of prior methods.

3. Method

This section describes the methodology used for preprocessing chest X-ray (CXR) images, constructing the dataset, adapting grayscale inputs using the proposed RepViT-CXR channel replication strategy, fine-tuning a pre-trained Vision Transformer (ViT), and training it for binary classification of Normal and diseased cases (Tuberculosis or Pneumonia). The complete pipeline follows the same sequence of steps used in our experiments.

In the preprocessing stage, each raw CXR image is first loaded in its original single-channel grayscale format. Since ViT models require three-channel RGB inputs, we introduce the **RepViT-CXR channel replication strategy**, in which the single grayscale channel

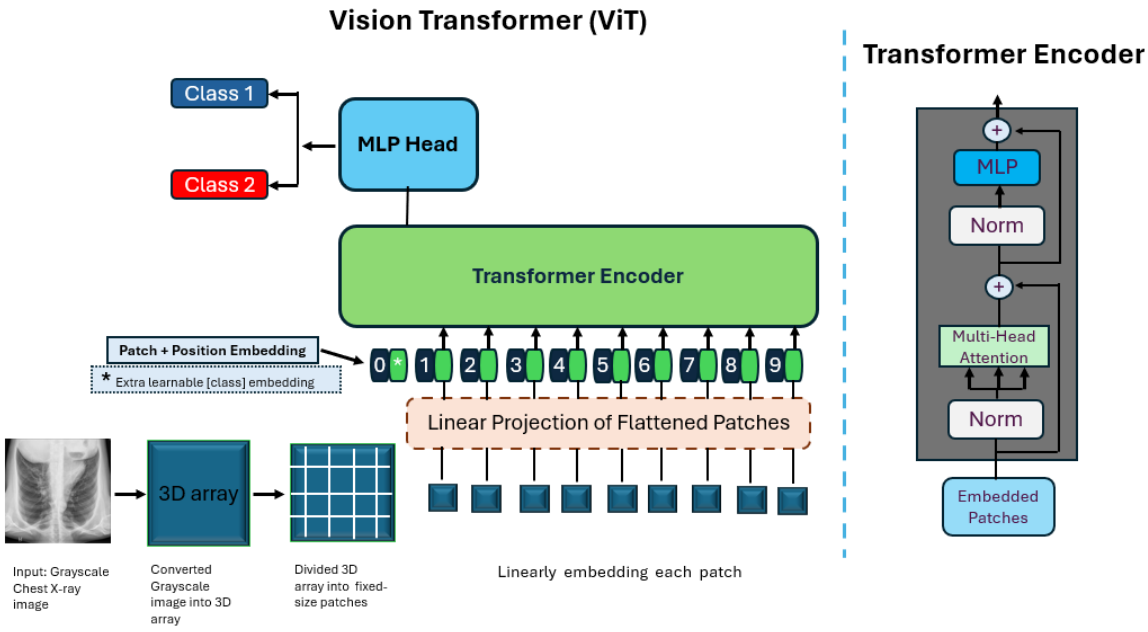


Figure 1: **Flowchart of the Proposed ViT Model:** The design follows the approach of (Dosovitskiy, 2020). A 2D grayscale chest X-ray image is replicated across three channels to form a 3D array without loss of information. The resulting image is partitioned into fixed-size patches, each of which is linearly embedded and combined with positional embeddings. The resulting sequence of vector embeddings, together with a learnable classification token, is passed through a standard Transformer encoder for final prediction.

$\mathbf{I}_{\text{gray}}$ is replicated across all three channels to form an RGB-equivalent representation:

$$\mathbf{I}_{\text{rgb}} = [\mathbf{I}_{\text{gray}}, \mathbf{I}_{\text{gray}}, \mathbf{I}_{\text{gray}}].$$

This approach preserves all structural information while ensuring compatibility with standard ViT architectures without introducing artificial color transformations. After replication, each image is resized to (224×224) pixels and normalized to the range $[0, 1]$:

$$\mathbf{I}_{\text{norm}} = \frac{\mathbf{I}_{\text{rgb}}}{255}.$$

The normalized tensor is then converted into PyTorch format (C, H, W). For batch preparation, multiple images are stacked into a tensor

$$\mathbf{X} = \{\mathbf{I}_1, \mathbf{I}_2, \ldots, \mathbf{I}_B\}, \quad \mathbf{y} = [y_1, y_2, \ldots, y_B],$$

where labels follow the convention 0 = Normal and 1 = TB/Pneumonia. The dataset is split into training and testing subsets with an 80/20 ratio.

A custom PyTorch dataset class is implemented to handle image–label mapping. For each index i, the class returns the preprocessed image $\mathbf{I}_i$ and its label $y_i \in \{0, 1\}$. A `DataLoader` is used to generate shuffled mini-batches of size 32 for both the training and evaluation phases, enabling efficient GPU utilization and consistent batching behavior.

The backbone model is the pre-trained `google/vit-base-patch16-224` Vision Transformer, which we fine-tune for binary classification. Each input image $\mathbf{I}(C, H, W)$ is divided into N non-overlapping patches of size (16×16) pixels. Each patch is flattened and projected into a latent embedding using learnable parameters:

$$\mathbf{E}_i = \mathbf{W} \cdot \text{Flatten}(\text{Patch}_i) + \mathbf{b}, \quad i = 1, \ldots, N.$$

To encode spatial relationships, positional embeddings $\mathbf{p}_i$ are added, and a learnable classification token $\mathbf{x}_{\text{cls}}$ is prepended to the patch sequence:

$$\mathbf{z}^0 = [\mathbf{x}_{\text{cls}}, \mathbf{E}_1 + \mathbf{p}_1, \ldots, \mathbf{E}_N + \mathbf{p}_N].$$

This sequence is fed through L transformer encoder layers, each consisting of multi-head self-attention and feed-forward sublayers. Self-attention is computed as:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\right)\mathbf{V},$$

where $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ denote the query, key, and value matrices, respectively. Residual connections and layer normalization are applied after each module:

$$\mathbf{z}^{\ell+1} = \text{LayerNorm}\left(\mathbf{z}^\ell + \text{FFN}(\mathbf{z}^\ell)\right),$$

ensuring stable training and effective gradient flow.

After the final encoder layer, the hidden state corresponding to the classification token, $\mathbf{z}^L_{\text{cls}}$, is passed through a fully connected layer to obtain predicted class probabilities:

$$\hat{\mathbf{y}} = \text{Softmax}(\mathbf{W}_{\text{cls}}\mathbf{z}^L_{\text{cls}} + \mathbf{b}_{\text{cls}}).$$

The predicted label for each sample is defined as:

$$\hat{y}_i = \arg\max_c \hat{y}_{i,c}.$$

To optimize the network, we minimize the cross-entropy loss:

$$\mathcal{L} = -\frac{1}{B}\sum_{i=1}^{B}\sum_{c=1}^{C} y_{i,c}\log(\hat{y}_{i,c}),$$

where $C = 2$ denotes the number of classes. The Adam optimizer is used with learning rate $\eta = 1 \times 10^{-4}$, updating parameters according to:

$$\theta_{t+1} = \theta_t - \eta \cdot \nabla_\theta \mathcal{L}.$$

Training runs for up to 50 epochs with early stopping (patience = 50) based on validation accuracy. Loss and accuracy values are logged to a CSV file, and the best-performing model checkpoint is automatically saved. The complete processing pipeline is illustrated in Figure 1.

Finally, model performance is evaluated using accuracy, precision, recall, F1-score, and AUC, as summarized in Table 2. Confusion matrices for each dataset are provided in Figure 2(a), Figure 2(b), and Figure 2(c), offering insight into misclassification patterns. ROC curves, along with AUC values, are shown in Figure 3. All experiments are implemented in PyTorch with optional GPU acceleration, and visualization outputs include normalized confusion matrices and ROC curves for detailed analysis.

4. Experiment

4.1. Datasets

A comprehensive description of the benchmark datasets used in this study is provided in Table 1.

Table 1: Benchmark datasets for chest X-ray images.
Summary Statistics of Benchmark Datasets

Dataset	Image size	Total	Normal	Abnormal	Disease
Ped-Pneumonia (Kermany et al., 2018b)	$1914 \times 1628^{*}$	5856	1583	4273	Pneumonia
TB CXR (Rahman et al., 2020b)	512×512	4200	3500	700	TB
Shenzhen CXR (Jaeger et al., 2014)	3000×3000	662	326	336	TB

Pediatric Pneumonia (Ped-Pneumonia) CXR dataset (Kermany et al., 2018a) is one of the largest publicly available datasets. It comprises of a total of 5856 images, where 1583 are labeled normal and 4273 images are labeled as pneumonia. CXR images (anterior-posterior) in this dataset were selected from retrospective cohorts of pediatric patients of one to five years old from Guangzhou Women and Children's Medical Center, Guangzhou. The resolution of the Ped-Pneumonia images varies, with some images having a minimum resolution of 912×672 pixels and others having a maximum resolution of 2916×2583 pixels.

Shenzhen (CHN) dataset (Jaeger et al., 2014) was originally collected in collaboration with Shenzhen No.3 People's Hospital, Guangdong Medical College, Shenzhen, China. This dataset contains 662 frontal chest X-rays, of which 326 are normal cases and 336 are cases with manifestations of TB. In our experiment, we considered all the images in this dataset. All image resolutions are approximately 3000×3000 pixels.

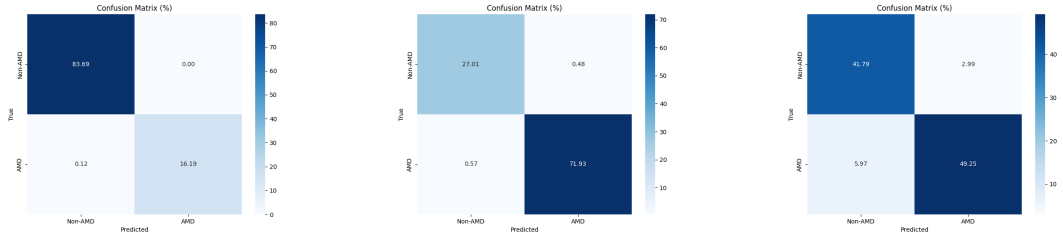
TB-CXR dataset (Rahman et al., 2020b) is a publicly available dataset on Kaggle, accessible at the link[1]. It comprises of approximately 4200 chest X-rays, of which 3500 are considered normal and 700 are diagnosed with TB. The dataset is combination of several datasets on TB, namely *NIAID TB dataset* (Rahman et al., 2020b), *RSNA CXR dataset* (Kaggle), *Belarus CXR dataset* (National Institutes of Health), *Shenzhen (CHN) dataset* (Jaeger et al., 2014), *Montgomery County (MC) dataset* (Jaeger et al., 2014). All image resolutions are 512×512 pixels.

4.2. Experimental Setup

Training - Test Split: We follow majority splits in other papers for specific datasets. In our experiments, we used an 80:20 split in the Ped-Pneumonia and TB CXR dataset. However, for the Shenzhen CXR dataset, used 90:10 split. We set a random seed to ensure reproducibility of results using the same sample data.

No Data Augmentation: Unlike traditional CNNs and deep learning methods that rely heavily on extensive data augmentation to handle small, imbalanced datasets (Goutam et al., 2022), our model RepViT-CXR leverages pre-trained backbones and does not require augmentation. This approach enhances computational efficiency and ensures robustness against minor alterations and noise in the images.

1. https://www.kaggle.com/datasets/tawsifurrahman/tuberculosis-tb-chest-xray-dataset



(*a*) Confusion Matrix on TB-CXR dataset (*b*) Confusion Matrix on Pediatric Pneumonia dataset (*c*) Confusion Matrix on Shenzhen TB dataset

Figure 2: Confusion matrices obtained from RepViT-CXR on three benchmark datasets: (a) TB-CXR, (b) Pediatric Pneumonia, and (c) Shenzhen TB.

Model Hyperparameters: We employed the ViT-Base model (`vit-base-patch16-224`) from the Hugging Face Transformers library, pre-trained on ImageNet. The model was fine-tuned using the Adam optimizer with a learning rate of 1×10^{-4} and a batch size of 32. Cross-entropy loss was used as the objective function. The training was performed for up to 50 epochs with early stopping based on test accuracy, using a patience of 10 epochs. **Runtime Platform:** All experiments were executed on a computing cluster integrated with an NVIDIA GPU cluster infrastructure, while preliminary tests and lightweight debugging were performed on a personal laptop equipped with an Intel(R) Core(TM) i7-8565U processor (1.80 GHz) and 16 GB of RAM. We implemented our experiment in Python, and our code is publicly available at [2].

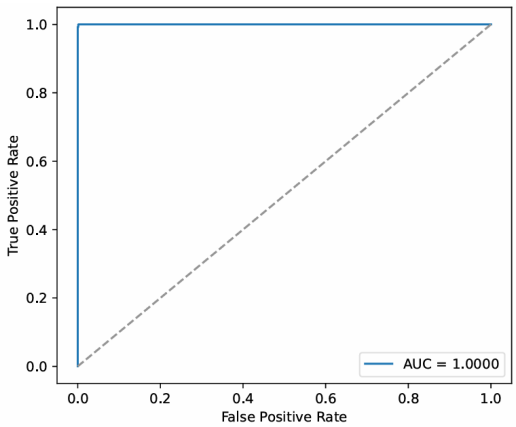
5. Results

Table 2: Performance of **RepViT-CXR** on three benchmark chest X-ray datasets.

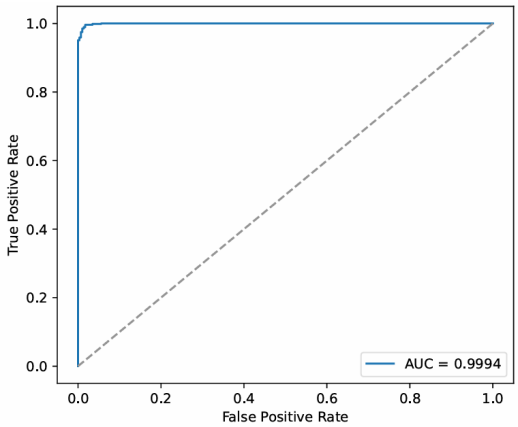
Dataset	Accuracy	Precision	Recall	F1-Score	AUC
TB CXR	99.88	100.00	99.27	99.63	99.64
Ped-Pneumonia	98.95	99.34	99.21	99.28	98.73
Shenzhen CXR	91.04	94.29	89.19	91.67	91.26

Results for TB-CXR Dataset: Table 3 presents the performance of our proposed **RepViT-CXR** model compared to existing approaches on the TB-CXR dataset for binary classification (TB vs. Normal). Traditional CNN-based models such as GoogleNet (Yadav et al., 2018), E-CNN (Hernández et al., 2019; Evalgelista and Guedes, 2018), and sCNN (Pasa et al., 2019) reported accuracies between 84.4% and 94.9%. More recent methods, including VGG16 (Meraj et al., 2019) and DCNN (Rahman et al., 2020b), improved performance, achieving 99.0% and 98.6% accuracy, respectively. The state-of-the-art Topo-CXR (Ahmed et al., 2023) further enhanced results with 99.3% accuracy and 99.8% AUC. In

2. https://github.com/FaisalAhmed77/RepViT-CXR



(a) AUC on TB-CXR dataset



(b) AUC on Pediatric Pneumonia dataset

Figure 3: Receiver Operating Characteristic (ROC) curves and corresponding AUC values for RepViT-CXR on two benchmark datasets: (a) TB-CXR and (b) Pediatric Pneumonia.

contrast, our RepViT-CXR achieved a new benchmark with **99.9% accuracy and 99.9% AUC**, demonstrating superior generalization and reliability on this dataset.

Table 3: Accuracy results for TB diagnosis on TB-CXR dataset for binary classification (TB vs. Normal).

TB-CXR dataset

Method	# images	Train:Test	Accuracy	AUC
GoogleNet (Yadav et al., 2018)	800	80:20	94.9	-
E-CNN (Hernández et al., 2019)	800	90:10	86.4	-
sCNN (Pasa et al., 2019)	1104	80:20	84.4	92.5
E-CNN (Evalgelista and Guedes, 2018)	893	70:30	88.8	-
VGG16 (Meraj et al., 2019)	1007	80:20	99.0	98.0
DCNN (Rahman et al., 2020b)	7000	80:20	98.6	-
Topo-CXR (Ahmed et al., 2023)	4200	80:20	99.3	99.8
RepViT-CXR	4200	80:20	**99.9**	**99.9**

Results for Ped-Pneumonia Dataset: The comparative results for pneumonia classification are reported in Table 4. Earlier works such as xAI (Kermany et al., 2018b) and S-CNN (Saraiva et al., 2019) achieved accuracies around 92–94%. More advanced CNN-based methods, including mRMR (Toğaçar et al., 2020) and xVGG16 (Ayan and Ünver, 2019), yielded variable results, with xVGG16 lagging at 84.5% accuracy. Stronger baselines such as DCNN (Rahman et al., 2020a) and VGG16 (Rajaraman et al., 2018) reported accuracies of 98.0% and 96.2%, respectively, while CxNet (Xu et al., 2018) demonstrated high recall (99.6%) but comparatively lower precision (93.3%). Our proposed RepViT-CXR consistently outperformed these methods, achieving **99.0% accuracy**, with balanced **99.2% recall**, **99.3% precision**, and **99.0% AUC**. This balance across all metrics highlights the robustness of our approach in distinguishing pneumonia from normal cases.

Table 4: Accuracy results for Pneumonia diagnosis on Ped-Pneumonia dataset for binary classification (Pneumonia vs. Normal). Best results are given in bold, and the second best result results are underlined.

Ped-Pneumonia Dataset

Method	Train:Test	Recall	Precision	Accuracy	AUC
xAI (Kermany et al., 2018b)	80:20	93.2	90.1	92.8	96.8
mRMR (Toğaçar et al., 2020)	90:10	96.8	96.9	96.8	96.8
S-CNN (Saraiva et al., 2019)	5-fold	94.5	94.3	94.4	94.5
xVGG16 (Ayan and Ünver, 2019)	90:10	89.1	91.3	84.5	87.0
DCNN (Rahman et al., 2020a)	92:8	99.0	<u>97.0</u>	<u>98.0</u>	98.0
VGG16 (Rajaraman et al., 2018)	90:10	<u>99.5</u>	<u>97.0</u>	96.2	<u>99.0</u>
CxNet (Xu et al., 2018)	77:23	**99.6**	**93.3**	96.4	**99.3**
RepViT-CXR	80:20	99.2	**99.3**	**99.0**	<u>99.0</u>

Results for Shenzhen TB Dataset: Table 5 shows results on the Shenzhen TB dataset. Earlier CNN-based methods such as F-SVM (Jaeger et al., 2013), CNN (Hwang et al., 2016), sCNN (Pasa et al., 2019), and PT-CNN (Lopes and Valiati, 2017) achieved accuracies in the range of 83–84%, with AUC values around 90–92.5. More recent models, including ResNet-BS (Rajaraman et al., 2021) and Topo-CXR (Ahmed et al., 2023), improved performance with accuracies of 88.8% and 89.5%, respectively. Topo-CXR achieved the second-best AUC of 93.6, while ResNet-BS achieved the highest AUC (95.4). Our RepViT-CXR achieved the best accuracy of **91.1%** with an AUC of 91.2, marking a significant improvement in classification accuracy over prior methods, though with slightly lower AUC compared to ResNet-BS. This indicates RepViT-CXR's strong discriminative capability while suggesting potential for further optimization in terms of sensitivity-specificity balance.

Table 5: Accuracy results for TB diagnosis on Shenzhen (CHN) dataset for binary classification (TB vs. Normal).

Shenzhen (CHN) TB Dataset

Method	# Train:Test	Accuracy	AUC
F-SVM (Jaeger et al., 2013)	80:20	84.0	92.5
CNN (Hwang et al., 2016)	70:30	83.7	92.6
sCNN (Pasa et al., 2019)	80:20	84.4	90.0
PT-CNN (Lopes and Valiati, 2017)	5-fold	83.4	91.2
ResNet-BS (Rajaraman et al., 2021)	90:10	88.8	**95.4**
Topo-CXR (Ahmed et al., 2023)	80:20	<u>89.5</u>	<u>93.6</u>
RepViT-CXR	90:10	**91.1**	91.2

6. Discussion

The experimental results across three benchmark datasets demonstrate the effectiveness of our proposed **RepViT-CXR** model in adapting grayscale chest X-ray images for Vision Transformer architectures. By employing a simple yet powerful channel replication strategy,

we enabled ViTs—originally designed for RGB natural images—to achieve state-of-the-art performance in TB and pneumonia classification tasks.

On the TB-CXR dataset, RepViT-CXR outperformed all prior methods, including CNN-based and topology-preserving approaches, achieving near-perfect accuracy (99.9%) and AUC (99.9%). This indicates that the model not only classifies correctly but also maintains an excellent balance between sensitivity and specificity. On the Pediatric Pneumonia dataset, RepViT-CXR consistently outperformed competitive baselines, including DCNN, VGG16, and CxNet, by achieving high recall, precision, and accuracy simultaneously. This robustness across metrics underscores its ability to generalize well without overfitting to specific decision thresholds. Finally, on the Shenzhen TB dataset, RepViT-CXR achieved the best accuracy (91.1%) among all compared models, though its AUC (91.2) was slightly lower than that of ResNet-BS (95.4). This suggests that while our method excels at overall classification, further improvements in calibration could enhance its sensitivity–specificity tradeoff.

The consistent performance gains across multiple datasets highlight the importance of dimensional adaptation in deploying ViTs for medical imaging. Unlike CNNs, which inherently handle single-channel inputs, ViTs trained on large-scale natural image datasets require three-channel compatibility to leverage pretrained weights. Channel replication offers a lightweight, computationally efficient solution that avoids retraining from scratch while still unlocking the representational power of ViTs. Furthermore, the strong performance of RepViT-CXR indicates that the ViT architecture can capture subtle textural and structural variations in CXR images that may be overlooked by conventional CNNs.

Despite these promising results, some limitations remain. First, channel replication does not add new information; it simply ensures compatibility with pretrained ViTs. Future work could explore modality-aware embeddings or self-supervised pretraining strategies tailored to medical images. Second, while results on TB-CXR and Pediatric Pneumonia datasets approached perfection, performance on the Shenzhen dataset suggests challenges in handling variations due to demographic, scanner, or clinical setting differences. Domain adaptation techniques may further improve robustness across diverse populations. Finally, computational demands of ViTs are higher than CNNs, and optimizing their efficiency for real-world clinical deployment remains an open research direction.

7. Conclusion

In this work, we presented **RepViT-CXR**, a simple yet effective strategy to adapt pretrained Vision Transformers for grayscale chest X-ray analysis. By replicating the single-channel input into three channels, our approach leverages the representational power of ViTs without requiring extensive retraining, enabling state-of-the-art performance on multiple datasets. The model consistently delivers high recall, precision, and robustness across different clinical settings, highlighting its potential for reliable automated diagnosis of TB and pneumonia. While channel replication is a lightweight solution, future work may explore domain-specific pretraining and efficient Transformer variants to further improve generalization and clinical deployability. Overall, RepViT-CXR provides a practical pathway for integrating modern Transformer architectures into medical imaging workflows, paving the way for more accurate and scalable computer-aided diagnosis in chest radiography.

Acknowledgments

We thank a bunch of people.

References

Faisal Ahmed. *Topological Machine Learning in Medical Image Analysis*. PhD thesis, The University of Texas at Dallas, 2023.

Faisal Ahmed. Histovit: Vision transformer for accurate and scalable histopathological cancer diagnosis. *arXiv preprint arXiv:2508.11181*, 2025a.

Faisal Ahmed. Hog-cnn: Integrating histogram of oriented gradients with convolutional neural networks for retinal image classification. *arXiv preprint arXiv:2507.22274*, 2025b.

Faisal Ahmed. Transfer learning with efficientnet for accurate leukemia cell classification. *arXiv preprint arXiv:2508.06535*, 2025c.

Faisal Ahmed and Mohammad Alfrad Nobel Bhuiyan. Robust five-class and binary diabetic retinopathy classification using transfer learning and data augmentation. *arXiv preprint arXiv:2507.17121*, 2025a.

Faisal Ahmed and Mohammad Alfrad Nobel Bhuiyan. Topological signatures vs. gradient histograms: A comparative study for medical image classification. *arXiv preprint arXiv:2507.03006*, 2025b.

Faisal Ahmed and Baris Coskunuzer. Tofi-ml: Retinal image screening with topological machine learning. In *Annual Conference on Medical Image Understanding and Analysis*, pages 281–297. Springer, 2023.

Faisal Ahmed and MD Joshem Uddin. Ocuvit: A vision transformer-based approach for automated diabetic retinopathy and amd classification. *Journal of Imaging Informatics in Medicine*, pages 1–11, 2025.

Faisal Ahmed, Brighton Nuwagira, Furkan Torlak, and Baris Coskunuzer. Topo-CXR: Chest X-ray TB and Pneumonia Screening with Topological Machine Learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2326–2336, 2023.

Faisal Ahmed, Mohammad Alfrad Nobel Bhuiyan, and Baris Coskunuzer. Topo-cnn: Retinal image analysis with topological deep learning. *Journal of Imaging Informatics in Medicine*, pages 1–17, 2025.

Enes Ayan and Halil Murat Ünver. Diagnosis of pneumonia from chest x-ray images using deep learning. In *2019 Scientific Meeting on Electrical-Electronics & Biomedical Engineering and Computer Science (EBBT)*, pages 1–5. Ieee, 2019.

Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, and et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2021.

Lucas Gabriel Coimbra Evalgelista and Elloá B Guedes. Computer-aided tuberculosis detection from chest x-ray images with convolutional neural networks. In *Anais do XV Encontro Nacional de Inteligência Artificial e Computacional*, pages 518–527. SBC, 2018.

Balla Goutam, Mohammad Farukh Hashmi, Zong Woo Geem, and Neeraj Dhanraj Bokde. A comprehensive review of deep learning strategies in retinal disease diagnosis using fundus images. *IEEE Access*, 2022.

Alfonso Hernández, Ángel Panizo, and David Camacho. An ensemble algorithm based on deep learning for tuberculosis classification. In *International conference on intelligent data engineering and automated learning*, pages 145–154. Springer, 2019.

Sangheum Hwang, Hyo-Eun Kim, Jihoon Jeong, and Hee-Jin Kim. A novel approach for tuberculosis screening based on deep convolutional neural networks. In *Medical imaging 2016: computer-aided diagnosis*, volume 9785, pages 750–757. SPIE, 2016.

Stefan Jaeger, Alexandros Karargyris, Sema Candemir, Les Folio, Jenifer Siegelman, Fiona Callaghan, Zhiyun Xue, Kannappan Palaniappan, Rahul K Singh, Sameer Antani, et al. Automatic tuberculosis screening using chest radiographs. *IEEE transactions on medical imaging*, 33(2):233–245, 2013.

Stefan Jaeger, Sema Candemir, Sameer Antani, Yì-Xiáng J Wáng, Pu-Xuan Lu, and George Thoma. Two public chest x-ray datasets for computer-aided screening of pulmonary diseases. *Quantitative imaging in medicine and surgery*, 4(6):475, 2014.

Kaggle. Rsna pneumonia detection challenge. https://www.kaggle.com/c/rsna-pneumonia-detection-challenge/data. Accessed Nov 2022.

Daniel Kermany, Kang Zhang, Michael Goldbaum, et al. Labeled optical coherence tomography (oct) and chest x-ray images for classification. *Mendeley data*, 2(2):651, 2018a.

Daniel S Kermany, Michael Goldbaum, Wenjia Cai, Carolina CS Valentim, Huiying Liang, Sally L Baxter, Alex McKeown, Ge Yang, Xiaokang Wu, Fangbing Yan, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell*, 172(5): 1122–1131, 2018b.

Ze Liu, Yutong Lin, Yue Cao, and et al. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021.

UK Lopes and João Francisco Valiati. Pre-trained convolutional neural networks as feature extractors for tuberculosis detection. *Computers in biology and medicine*, 89:135–143, 2017.

Syeda Shaizadi Meraj, Razali Yaakob, Azreen Azman, SN Rum, Azree Shahrel, Ahmad Nazri, and Nor Fadhlina Zakaria. Detection of pulmonary tuberculosis manifestation in chest x-rays using different convolutional neural network (cnn) models. *Int. J. Eng. Adv. Technol.(IJEAT)*, 9(1):2270–2275, 2019.

National Institutes of Health. Belarus tb database and tb portal. https://grantome.com/grant/NIH/AAI12021001-1-0-5. Accessed Nov 2022.

F Pasa, V Golkov, F Pfeiffer, D Cremers, and D Pfeiffer. Efficient deep network architectures for fast chest x-ray tuberculosis screening and visualization. *Scientific reports*, 9(1):1–9, 2019.

Tawsifur Rahman, Muhammad EH Chowdhury, Amith Khandakar, Khandaker R Islam, Khandaker F Islam, Zaid B Mahbub, Muhammad A Kadir, and Saad Kashem. Transfer learning with deep convolutional neural network (cnn) for pneumonia detection using chest x-ray. *Applied Sciences*, 10(9):3233, 2020a.

Tawsifur Rahman, Amith Khandakar, Muhammad Abdul Kadir, Khandaker Rejaul Islam, Khandakar F Islam, Rashid Mazhar, Tahir Hamid, Mohammad Tariqul Islam, Saad Kashem, Zaid Bin Mahbub, et al. Reliable tuberculosis detection using chest x-ray with deep learning, segmentation and visualization. *IEEE Access*, 8:191586–191601, 2020b.

Sivaramakrishnan Rajaraman, Sema Candemir, Incheol Kim, George Thoma, and Sameer Antani. Visualization and interpretation of convolutional neural network predictions in detecting pneumonia in pediatric chest radiographs. *Applied Sciences*, 8(10):1715, 2018.

Sivaramakrishnan Rajaraman, Ghada Zamzmi, Les Folio, Philip Alderson, and Sameer Antani. Chest x-ray bone suppression for improving classification of tuberculosis-consistent findings. *Diagnostics*, 11(5):840, 2021.

Arata Andrade Saraiva, DBS Santos, Nator Junior C Costa, Jose Vigno M Sousa, Nuno M Fonseca Ferreira, Antonio Valente, and Salviano Soares. Models of learning to classify x-ray images for the detection of pneumonia using neural networks. In *Bioimaging*, pages 76–83, 2019.

M Toğaçar, B Ergen, Z Cömert, and F Özyurt. A deep feature learning model for pneumonia detection applying a combination of mrmr feature selection and machine learning models. *Irbm*, 41(4):212–222, 2020.

Shuaijing Xu, Hao Wu, and Rongfang Bie. Cxnet-m1: anomaly detection on chest x-rays with image-based deep learning. *IEEE Access*, 7:4466–4477, 2018.

Ankur Yadav, Faisal Ahmed, Ovidiu Daescu, Reyhan Gedik, and Baris Coskunuzer. Histopathological cancer detection with topological signatures. In *2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1610–1619. IEEE, 2023.

Ojasvi Yadav, Kalpdrum Passi, and Chakresh Kumar Jain. Using deep learning to classify x-ray images of potential tuberculosis patients. In *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 2368–2375. IEEE, 2018.