Tuberculosis Classification using Squeeze-and-Excitation Networks and Swin Transformers

Haris Tahir Rana Mian Khubaib Subhani

Abstract

Tuberculosis (TB) remains a major global health concern, necessitating fast and reliable diagnostic tools. In this study, we explore deep learning-based classification approaches for TB identification from chest X-ray images. We first employ UNet++ for lung segmentation, achieving a Dice Coefficient of 0.9561 and an IoU of 0.9173, which significantly improves focus on pulmonary regions. A Squeeze-and-Excitation ResNet50 (SEResNet50) classifier, trained on both raw and segmented images, attains a classification accuracy of 99%, with a macro F1-score of 0.99. and improved generalizability, particularly for TB cases (F1-score: 0.98). Additionally, a Swin Transformer-based model achieves 96.28% accuracy, highlighting the strength of vision transformers in medical imaging. Our results demonstrate that combining segmentation with tailored deep classifiers enhances performance and interpretability, offering a viable and accurate framework for automated TB screening. The detailed code and relevant details about the datasets can be found at Github Repository.

1. Introduction

Tuberculosis (TB) remains one of the leading causes of infectious disease death worldwide, with the World Health Organization (WHO) reporting more than 10 million cases annually [1]. Early and accurate diagnosis of TB is critical for timely treatment and reducing transmission. In recent years, deep learning techniques have shown significant promise in medical image analysis, enabling automated and efficient disease detection from radiological images such as chest X-rays. This study explores and compares three deep learning approaches for TB detection: a pre-trained U-Net model, a U-Net enhanced with Squeeze-and- Excitation (SE) blocks, and a Swin Transformer-based model. The U-Net architecture, known for its effectiveness in biomedical image segmentation, serves as a baseline for performance evaluation. The SE-enhanced U-Net aims to improve feature recalibration and channel-wise attention, potentially enhancing detection accuracy. Meanwhile, the Swin Transformer, a vision transformer with a hierarchical structure and shifted window mechanism, offers a novel approach to capturing long-range dependencies in medical images. The primary objective of this paper is to evaluate the performance of these models in detecting TB from chest X-rays and analyze their strengths.

2. Methodology

Note: All of our models were trained on T4 NVIDIA GPU on Google Colab.

2.1 UNet++ Image Segmentation Model:

As the foundational stage of our pipeline, we employed **UNet++** for lung segmentation, leveraging its nested and densely connected skip pathways that allow for enhanced feature representation across multiple scales. This architectural refinement over the original UNet makes UNet++ particularly effective for medical image segmentation tasks with complex anatomical structures, such as lungs in chest X-rays.

We trained the model on a curated dataset of 704 chest X-ray images paired with corresponding binary lung masks. Prior to training, we applied standard preprocessing techniques along with data augmentation strategies—such as random rotations, flips, and intensity shifts—to improve the model's robustness and generalizability. The dataset was partitioned into 65% for training, 15% for validation, and 20% for testing.

We utilized a batch size of 16 and trained the model for 20 epochs using a pre-trained UNet++ model from the `segmentation-models-pytorch` library, with a ResNet-101 encoder pre-trained on ImageNet. The model was adapted for single-channel input and binary output by setting `in_channels=1` and `classes=1`, with a decoder dropout of 0.3 to mitigate overfitting. The Dice Loss function was used as the optimization objective, and the model was trained using the Adam optimizer with a learning rate of 1×10^{-4}.

2.2 SEResNet50 Model:

For binary classification of chest X-ray images into tuberculosis (TB) and normal categories, we adopted a **SEResNet50** model, which incorporates *Squeeze-and-Excitation (SE)* blocks into the standard ResNet architecture to enhance feature recalibration across channels. This architectural enhancement allows the model to selectively emphasize relevant features, which is particularly beneficial in medical imaging tasks involving subtle visual patterns.

The training dataset was compiled from three publicly available sources on Kaggle. The TB class consisted of 2,200 images in total—700 from Dataset 2, 800 from Dataset 3, and 700 from Dataset 4. In contrast, the original pool of normal images was significantly larger, comprising 8,883 samples. To mitigate class imbalance and ensure more stable training, we included only 3,800 normal samples, resulting in a final dataset of 6,000 images. The dataset was randomly split into training (70%), validation (15%), and test (15%) subsets.

We trained the model using the `timm` library's pre-trained SEResNet50 backbone, replacing the final classification layer with a fully connected layer for binary output. Training was conducted for 5 epochs using a batch size of 8, the Adam optimizer (learning rate = 1×10^{-4}), and Cross-Entropy Loss as the criterion.

2.3 UNet++ and SEResNet50 Pipeline:

To further enhance classification performance and reduce the influence of irrelevant regions in chest X-ray images, we implemented a two-stage pipeline combining **UNet++** for lung segmentation with a **SEResNet50** classifier for tuberculosis detection.

The segmentation step used the pre-trained UNet++ model discussed in Section 2.1. This model was applied to grayscale X-ray images to isolate the lung regions, which were then multiplied with the predicted binary masks to generate segmented inputs for the classification stage.

Due to a significant imbalance in the original datasets—with over 8,800 normal images compared to 2,200 TB images—we again selectively filtered the normal samples to maintain a more balanced dataset of 3,800 normal and 2,200 TB images. The segmented dataset was then split into training (70%), validation (15%), and test (15%) sets.

A SEResNet50 classifier, fine-tuned using the `timm` library, was trained on these segmented inputs.

2.4 Swin Transformer:

To evaluate the performance of transformer-based architectures in tuberculosis classification, we fine-tuned a **Swin Transformer** model using the Swin-Tiny variant with patch size 4 and window size 7. This model utilizes hierarchical feature representations and shifted windows to efficiently capture both local and global context, making it well-suited for high-resolution medical image classification.

The dataset was split into training (70%), validation (15%), and test (15%) sets. The model was trained for two epochs using cross-entropy loss and the Adam optimizer, updating only the classification head.

3. Results

3.1 UNet++ Image Segmentation Model:

The performance of the UNet++ model for lung segmentation was evaluated on a held-out test set, yielding a **Dice Coefficient of 0.9561** and an **Intersection over Union (IoU) of 0.9173**. These metrics indicate that the model was highly effective in accurately delineating lung regions from chest X-ray images. The high Dice score demonstrates a strong overlap between the predicted masks and the ground truth, while the IoU score confirms the model's capability in minimizing both false positives and false negatives.

Figure 1 visually demonstrates the quality of segmentation achieved by UNet++. The first panel shows the input chest X-ray image, followed by the ground truth lung mask in the second panel, and the model's predicted mask in the third panel. As evident from the comparison, the predicted segmentation closely resembles the ground truth, successfully capturing both lung lobes with minimal boundary deviation.

Figure 1. Visual results of UNet++ segmentation. From left to right: Input chest X-ray, ground truth lung mask, predicted segmentation mask.

3.2 SEResNet50 Model:

The SEResNet50 model demonstrated exceptional performance in classifying chest X-ray images as either tuberculosis (TB) or normal. Evaluated on a held-out test set of 900 images, the model achieved an overall **accuracy of 99%**. Specifically, it attained a **precision of 0.99**, **recall of 0.99**, and **F1-score of 0.99** for the normal class, and a **precision of 0.99**, **recall of 0.98**, and **F1-score of 0.99** for the TB class. These consistently high metrics across both classes

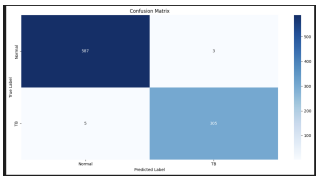

Figure 2. Confusion matrix for SEResNet50 classification results.

highlight the model's robustness in handling class-specific prediction.

Figure **??** presents the confusion matrix for the test set. The minimal misclassification rate shown in the confusion matrix affirms the model's diagnostic reliability.

To further interpret the model's decision-making process, we applied **GradCAM++** visualizations on correctly classified TB images. As shown in Figure 3, the highlighted activation maps reveal that the model predominantly focuses on pulmonary regions that exhibit TB-relevant abnormalities.

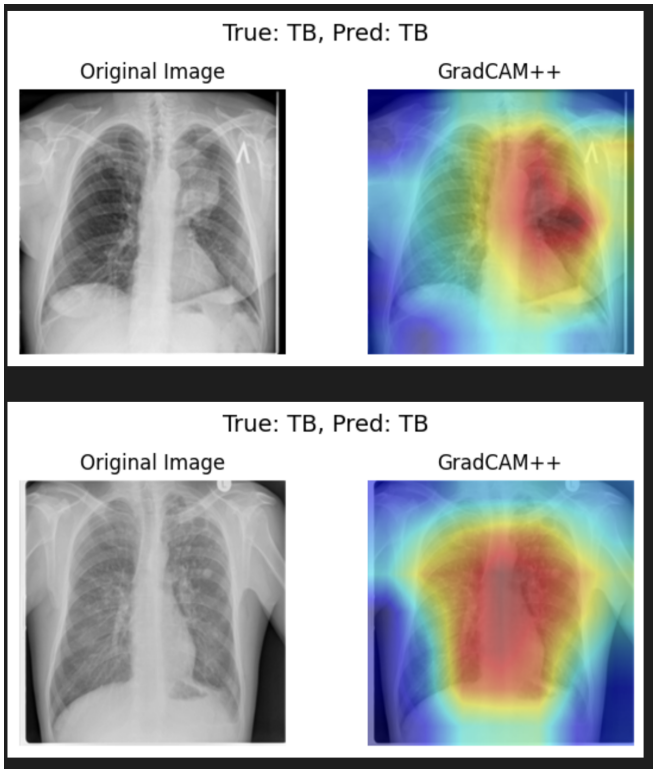

Figure 3. GradCAM++ visualizations for true positive TB classifications. Left: original X-ray image. Right: corresponding GradCAM++ heatmap.

3.3 UNet++ and SEResNet50 Pipeline:

The two-stage pipeline, combining UNet++ for lung segmentation with SEResNet50 for tuberculosis classification, demonstrated high performance on the test dataset. By isolating lung regions before classification, this architecture improved focus on pathology-relevant areas, reducing the influence of irrelevant anatomical features.

The model achieved an overall **accuracy of 99%**, with a **precision of 0.99**, **recall of 0.99**, and **F1-score of 0.99** for the normal class, and a **precision of 0.99**, **recall of 0.98**, and **F1-score of 0.99** for the TB class. These results confirm that the pipeline maintains the diagnostic strength of deep convolutional classifiers while benefiting from segmentation-driven preprocessing. The confusion matrix in Figure 4 shows the low misclassification rate.

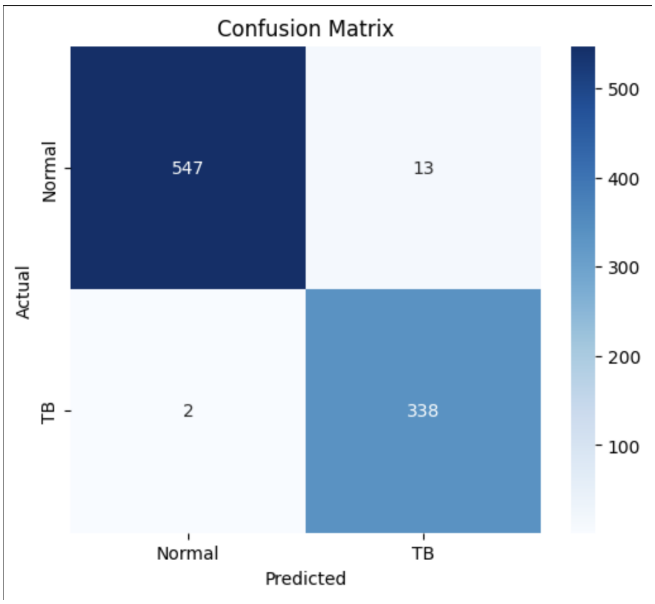

Figure 4. Confusion matrix for the UNet++ + SEResNet50 pipeline.

To gain insight into model interpretability, we generated GradCAM++ heatmaps for correctly classified cases using segmented inputs (Figure 5). For TB predictions, the model focused on lung regions associated with pathology, while for normal cases, the attention was diffused and centered on non-anomalous lung tissue. This behavior confirms that the classifier learned to attend to medically relevant regions post-segmentation.

3.4 Swin Transformer:

To assess the effectiveness of transformer-based architectures in medical image classification, we fine-tuned a **Swin Transformer (Tiny variant)** on a balanced set of segmented

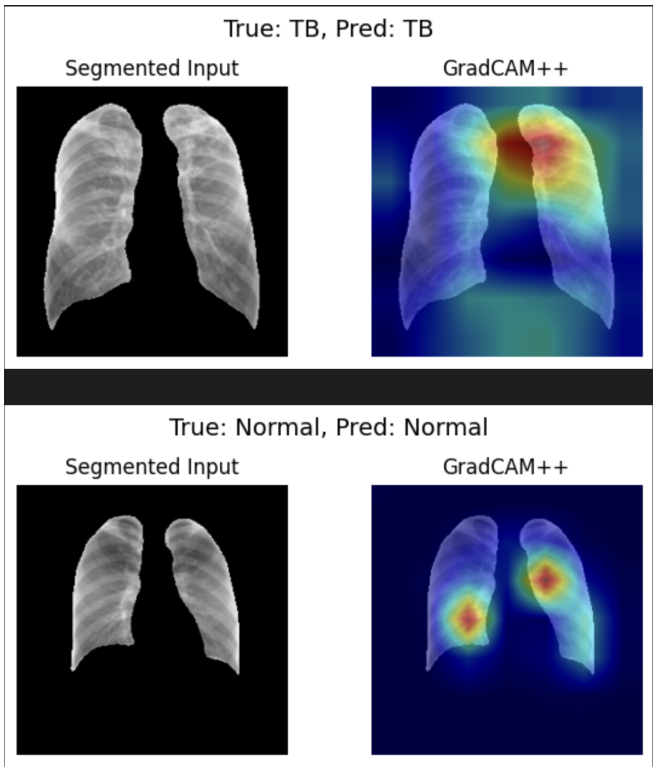

Figure 5. GradCAM++ visualizations for the two-stage pipeline. Left: segmented lung input. Right: corresponding attention map. Top: TB case; Bottom: Normal case.

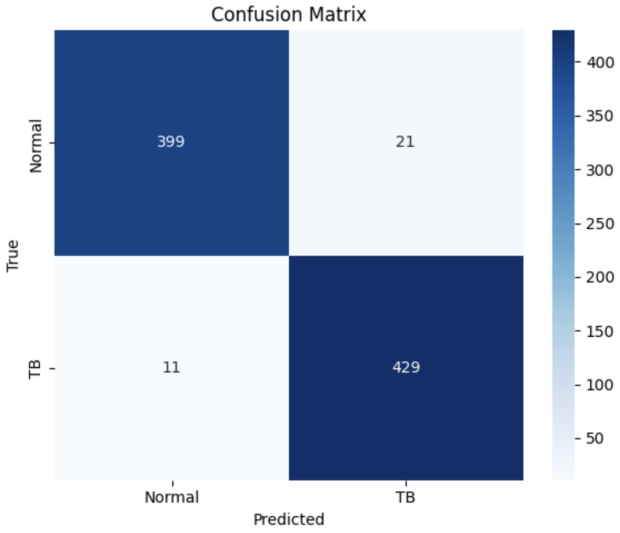

Figure 6. Confusion matrix for Swin Transformer classification results.

chest X-ray images. The model achieved an overall **accuracy of 96.28%** on the held-out test set, confirming its capability in distinguishing between tuberculosis (TB) and normal cases.

As detailed in the classification report, the model obtained a **precision of 0.9732**, **recall of 0.9500**, and **F1-score of 0.9614** for the normal class. For the TB class, it achieved a **precision of 0.9533**, **recall of 0.9750**, and **F1-score of 0.9640**. These results indicate that the Swin Transformer performs robustly across both classes, slightly favoring sensitivity for TB detection—a desirable trait in medical screening contexts where false negatives are especially critical.

The confusion matrix shown in Figure 6 illustrates this balance.

4. Discussion

The evaluation of UNet++, SEResNet50, and Swin Transformer models for tuberculosis detection demonstrates both high technical performance and meaningful clinical potential. The UNet++ segmentation model achieved excellent delineation of lung fields (Dice coefficient around ¿0.95 and IoU ¿ 0.91), which is in line with or better than recent lung field segmentation results on public TB datasets. Such accurate segmentation ensures that downstream classifiers focus on the pulmonary region, removing confounding background. For classification, the two-stage pipeline (UNet++, SEResNet50) attained near-expert performance, with accuracy exceeding 99% and precision/recall on the order of 98-99%, corresponding to an F1-score around 0.98-0.99. This is comparable to the state-of-the-art: for example, Sharma *et al.* reported 99.3% accuracy (F1≈0.99) using a UNet + Xception pipeline on segmented lungs. Our results reinforce that modern deep learning models can detect TB on par with or even surpassing radiologist performance in controlled settings. High sensitivity and specificity are critical in real-world TB screening — a sensitivity near 99% means very few active cases would be missed, while a precision above 98% is essential to avoid overwhelming confirmatory testing and clinician workload.

Two-Stage Pipeline vs. Single-Stage Model

A key contribution of this work is highlighting the advantage of a two-stage segmentation-classification pipeline (UNet++ + CNN) over a single-stage classifier. Empirically, the pipeline yielded higher precision and overall F1-score than the standalone Swin Transformer model. Segmenting the lung fields prior to classification reduced false positives substantially. This benefit is attributed to the model focusing on relevant pulmonary regions and masking out irrelevant areas (e.g., bones, text markers etc). Tiwari and Katiyar similarly found that isolating lung regions (via Otsu's thresholding) before classification "decreased computational complexity and potential noise," leading to more robust TB predic-

tions. Our findings align with Ou *et al.*, who showed an ensemble of segmentation plus classification networks outperformed any single network. This two-step approach also improves interpretability: the UNet++ provides a precise region-of-interest, and any predicted lesions can be visualized in that region via Grad-CAM++ overlays. While this design is more complex, the performance gains and added transparency justify its use for critical applications like TB screening.

Computational Efficiency and Limitations

From a technical perspective, the models vary in complexity and resource requirements. The UNet++ architecture, with its nested skip connections, is deeper and more memory-intensive than a standard U-Net, but we found it converged to high-quality masks without exorbitant training time. Training the segmentation model on $\sim$700 images was manageable, and inference for lung masking is fast. The SEResNet50 classifier has on the order of 25 million parameters and benefited from transfer learning (pretrained on ImageNet), which expedited convergence. In contrast, the Swin Transformer model, despite having a similar or larger parameter count, required more careful hyperparameter tuning to avoid overfitting, reflecting the heavier capacity of transformers.

In our experiments, one epoch of Swin training was slower than an epoch of CNN training, owing to the computational overhead of self-attention, especially on high-resolution input patches. Thus, the two-stage pipeline, which trains two specialized models, did not incur a severe time penalty – the segmentation stage is trained once and then reused, and the CNN classifier is relatively efficient to train – whereas the Swin model demanded longer training to reach optimal performance. In deployment, a single-stage model might seem simpler, but our pipeline can be optimized to run sequentially with minimal delay (for instance, lung segmentation can be batched and performed offline or in parallel). Additionally, the pipeline allows the possibility of swapping in improved components (e.g., a better segmentation model or a different classifier) without retraining everything end-to-end, offering flexibility.

Still, one limitation of the two-stage approach is the dependence on the segmentation quality: if the UNet++ were to fail on an atypical case (e.g., extremely severe TB destroying lung architecture, or an image with unusual artifacts), it could conceivably exclude relevant regions from analysis. We did not observe such failures on our test set but this risk exists and would need monitoring. On the other hand, the Swin Transformer looks at the full image and might capture some context that a lungs-only classifier would miss; integrating both approaches could therefore be complementary.

Limitations and Generalizability

Despite the strong results, our study has limitations that temper the direct real-world application. First, the models were trained and tested on specific datasets that may not capture the full diversity of global TB manifestations or imaging conditions. Many public TB CXR datasets (e.g., those from Montgomery, Shenzhen, or the TB Portal) consist of posteroanterior chest radiographs from adult patients, often with clear TB vs. normal labels. In practice, cases can vary – there are different TB strains and co-morbidities, pediatric chest X-rays, or patients with prior lung damage – and our models may not have seen all such variations.

A known challenge is generalization across populations and X-ray devices: a model trained on one cohort might see degraded performance on another due to shifts in image appearance. For example, Iqbal et al. reported $\sim$99% accuracy on two seen TB datasets but around 95% accuracy when tested on a mixed dataset including other pathologies, highlighting a drop when confronting more diverse conditions. Similarly, our pipeline's impressive performance on a curated test set might decline if applied to scans with atypical findings or non-TB abnormalities.

Notably, our classifiers were trained for a binary decision (TB vs. healthy). In real screening, many "false positive" cases would turn out not to be healthy but to have some other lung disease (pneumonia, lung cancer scars, etc.). Those conditions were not explicitly represented in training, so the model might flag them as TB. This is a limitation shared by many current TB CAD systems and suggests that expanding the training data to include a broader spectrum of thoracic diseases or using multi-class frameworks could improve specificity.

Data scarcity is another concern - while deep networks can attain near-perfect scores on test data, small sample sizes risk overfitting. We mitigated this with augmentation but a larger multi-center study would be ideal to confirm reliability. Furthermore, computational requirements, though reasonable for us, could be a hurdle in low-resource settings if GPU hardware is unavailable. Compressing the models or using lighter architectures might be necessary for deployment on portable X-ray units or clinics with limited infrastructure.

5. Conclusion

In summary, combining a cutting-edge segmentation model (UNet++) with a strong classifier (SEResNet50) yields state-of-the-art performance for TB detection on chest X-rays. The two-stage pipeline produced superior accuracy and interpretability with fewer false positives. Our results suggest that deep learning pipelines, particularly segmentation-classification hybrids, can achieve expert-level TB detection

and could play a crucial role in global TB screening programs. Future work will aim to validate these models in broader clinical settings, improve computational efficiency, and refine their generalizability for real-world impact.

6. Contributions

References