A Transfer learning-Based ViT Model For Tuberculosis Detection And Analysis Using Chest X-ray Images.

Mr. Athukuri Prasad
Research Scholar, JJTU
mailsofprasad@gmail.com
Dept. of ECE

Dr. Alok Agarwal
Internal Guide
alokagarwal26aaa@gmail.com
Dept. of ECE

Dr.Ch. Rami Reddy
Co - Guide
crreddy229@gmail.com
Dept. of EEE

Abstract:

This study used a dataset consisting of chest X-ray images to classify the images into three distinct categories Tuberculosis (TB), Pneumonia and Normal images. This study aims to investigate the topic of classifying TB images by using Deep Learning (DL) methodologies. Ongoing research efforts are now being conducted in this field to offer assistance to medical practitioners in the treatment of patients. Vision transformer (ViT) pre-trained models have been used. The core idea behind ViT is to apply the self-attention mechanism, which is central to transformer architectures, to images. Transfer learning (TL) is a very effective computer vision methodology that enables the generation of precise models in a resource-efficient manner. In this study, two pre-trained models, including ViTB-16 and ViTB-32 based on TL were used. The ViTB-16 model demonstrated superior performance in Tuberculosis identification as compared to the other model. The use of all two models in conjunction with our suggested strategy has resulted in a notable improvement in the classification accuracy score, reaching about 95.44%. There is an observed improvement in comparison to the previous studies in the field of Artificial Intelligence.

Key words: Deep Learning, Vision Transformer, Transfer learning

1 Introduction

Tuberculosis (TB) is a severe infectious disease caused by the bacterium Mycobacterium tuberculosis. It remains a global health concern, with a significant impact on mortality and morbidity. Early and accurate diagnosis of TB is crucial for effective disease management. TB diagnosis traditionally relies on methods such as sputum microscopy, culture, molecular tests and Imaging techniques. Chest X-ray (CXR) images play a pivotal role in diagnosing TB as they can reveal specific patterns and abnormalities associated with the disease[1].With the advancements in AI, computer-aided diagnosis (CAD) systems have emerged, enhancing the accuracy and efficiency of TB diagnosis through automated analysis of CXR images[2].

Vision Transformers, Dosovitskiy et al. (2020),[3]often referred to as "ViTs," are a class of neural network architecture that have gained prominence in the field of computer vision. They were introduced as an alternative to convolutional neural networks (CNNs) for image classification and other vision tasks. Unlike traditional CNNs, ViTs leverage self-attention mechanisms and sequential processing, making them particularly well-suited for capturing long-range dependencies in images. Their application in medical image analysis, including TB image classification, is gaining traction.ViTs were inspired by the success of transformer models in natural language processing, particularly models like BERT and GPT. Several studies have explored ViT's potential in medical image analysis, including segmentation, detection, and classification tasks. ViT's ability to learn global image features makes it promising for diseases like TB, where understanding overall lung conditions is crucial. Several studies have explored ViT's potential in medical image analysis, including

segmentation, detection, and classification tasks. ViT's ability to learn global image features makes it promising for diseases like TB, where understanding overall lung conditions is crucial[4].

The core idea behind Vision Transformers is to apply the self-attention mechanism, which is central to transformer architectures, to images. This enables the model to capture long-range dependencies and relationships between different image regions, which can be important for understanding the content of an image.An image is divided into fixed-size non-overlapping patches. Each patch is treated as a separate token and is processed by the model. This patch-based approach allows ViTs to handle images of different sizes without resizing. To account for the spatial information of the patches, positional encodings are added to the patch embeddings. These encodings help the model understand the relative positions of the patches in the image.This mechanism enables the model to weigh the importance of different patches when making predictions. The multi-head attention allows the model to focus on different aspects of the image simultaneously.Vision Transformers typically consist of multiple transformer layers, which process the patches and their features iteratively.After the self-attention mechanism, there are feedforward neural networks applied to the output of each patch to capture more complex features.At the end of the network, a classification head is added to make predictions for various tasks, such as image classification, object detection, and semantic segmentation. Fine-tuning a vision model like the ViTs for image classification involves adjusting and training specific parameters to adapt the pre-trained model to a particular classification task. ViT models are a type of deep learning model that has shown significant success in various computer vision tasks. Two common variants of ViT are ViT-B16 and ViT-B32, which differ in terms of their architecture size and complexity[5]. Figure 2 demonstrates the flow diagram of CXR image classification

2. **TB Image Classification: Related Work**

Li et al. [6]employed ViT for TB classification from CXR images, achieving competitive accuracy compared to traditional CNNs. They highlighted ViT's capability to discern subtle patterns indicative of TB. *Zhang et al.* [7]proposed a ViT-based model for TB severity assessment from CXR images. Their approach leveraged ViT's attention mechanism to focus on regions of interest, leading to precise severity classification.

The base models are pre-trained on a large dataset (e.g., ImageNet) and contain various layers, including a set of self-attention layers and feedforward neural networks. ViT-B16 refers to a specific variant of the Vision Transformer (ViT) model architecture. The "B16" in ViT-B16 typically refers to the model's configuration. The "B" often denotes the size or capacity of the model, and the "16" might refer to the input patch size. This means that the ViT-B16 model is likely a variant of the ViT architecture with a certain configuration and input patch size.The "B32" in "ViT-B32" likely refers to the model's architecture and size. In the context of ViT models, "B" usually denotes the base model, and "32" suggests the patch size used in the input images. In ViT models, the input image is divided into fixed-size non-overlapping patches, and each patch is treated as a token. The number after "B" typically represents the number of transformer layers in the model.ViT models, such as ViT-B16 and

ViT-B32, are variants with different model sizes and capacity. B16 is smaller and faster, while B32 is larger and potentially more accurate[8].

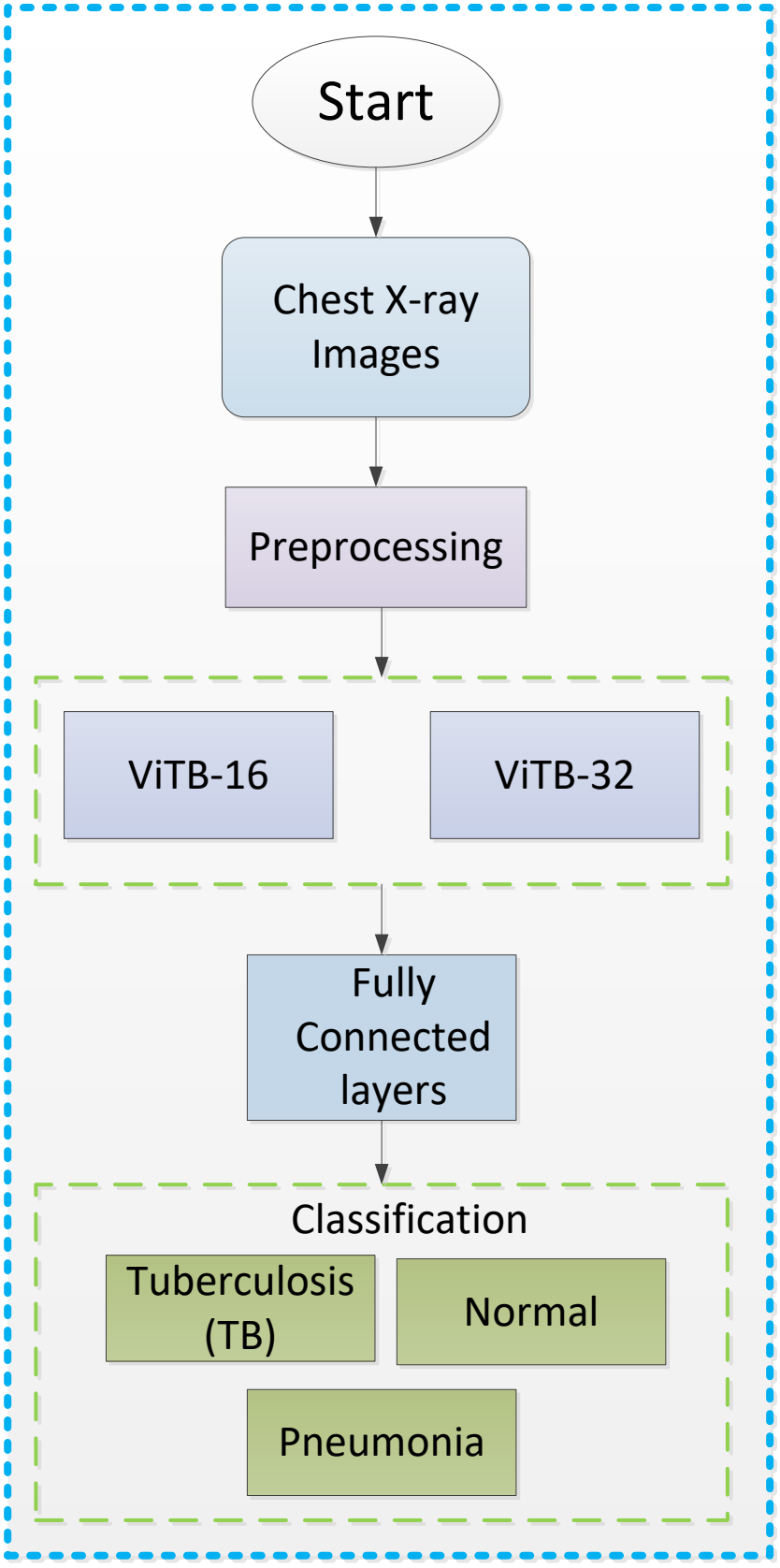

Figure 1: Flow diagram of CXR Image classification using ViT models

3. Transfer learning

Transfer learning using Vision Transformer (ViT) models like ViT-B16 and ViT-B32 for image classification involves leveraging pre-trained ViT models to solve new image classification tasks[9][10]. ViT models are typically pre-trained on large-scale image datasets, such as ImageNet or JFT-300M. During pre-training, these models learn to recognize a wide range of features and patterns within images.Transfer learning begins by taking a pre-trained ViT model, like ViT-B16 or ViT-B32, and adapting it for a new, specific image classification task[4].

5. Data set Preparation

The dataset is divided into training, validation, and test sets.Figure 2 demonstrates the x-ray images of TB, pneumonia and Normal lung Images.Data preprocessing techniques such as data augmentation (e.g., random cropping, flipping) are applied to increase model robustness. To improve the model's generalization and robustness, data augmentation techniques applied during training. Common augmentations include random cropping, flipping, rotation, and color adjustments. The choice and strength of augmentations can be important hyperparameters[11]. The classification head is typically composed of fully connected layers that map the extracted features to the number of classes in the specific classification task. The top (classification) layers of the pre-trained ViT model are replaced with new layers designed for the specific task. Typically, a few fully connected layers are added to transform the model's feature representation into the desired number of classes for your classification task[12].

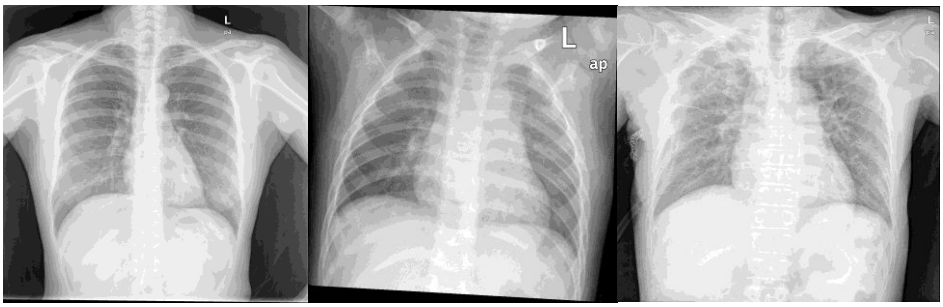

Figure 2: X-ray images of TB, Pneumonia and Normal.

6. Fine-tuningViT Models

The modified ViT model is then fine-tuned on the training dataset. During this phase, the model adjusts its weights to better fit the new data.The original weights of the pre-trained model are usually kept fixed, and only the newly added layers are updated.The model's performance is evaluated on the validation set, and hyper-parameters (e.g., learning rate, batch size) may be tuned to optimize performance.The learning rate determines how much the model's parameters are updated during training. The choice of the optimizer such as Adam, SGD, or others can influence the convergence speed and the final model performance. Common choices include cross-entropy loss for single-label classification and focal loss for imbalanced datasets or multi-label classification. The batch size determines how many data samples are processed in each training step[13]. A larger batch size can increase training

speed, but it may also require more memory. The batch size can be a crucial parameter for fine-tuning. The optimal number of epochs depends on the dataset and model complexity. regularization techniques like dropout or weight decay to prevent overfitting. Appropriate evaluation metrics used to assess the model's performance during and after training. Common metrics for classification tasks include accuracy, precision, recall, and F1-score After fine-tuning, the model's performance is assessed on the test dataset to ensure it generalizes well to new, unseen data.Once the model is fine-tuned and evaluated, it can be used for inference on new images for classification.Vision Transformer models, particularly ViT-B16 and ViT-B32, have shown considerable potential for TB CXR image classification. Their ability to capture global context and spatial relationships, coupled with their adaptability to medical imaging, makes them promising candidates for automated TB diagnosis [14].

A confusion matrix is a fundamental tool used in classification tasks to assess the performance of a model in TB detection[15]. It is particularly useful for binary classification problems, such as detecting whether a patient has TB or not. The confusion matrix consists of four key components

TP: Patients who have TB and are correctly identified as having TB.
TN: Patients who do not have TB and are correctly identified as not having TB.
FP: Patients who do not have TB but are incorrectly identified as having TB.
FN: Patients who have TB but are incorrectly identified as not having TB.

The confusion matrix allows to compute various performance metrics for a TB detection model, such as Sensitivity measures the proportion of actual TB cases that are correctly identified by the model. It is calculated as TP / (TP + FN).Specificity measures the proportion of non-TB cases that are correctly identified by the model. It is calculated as TN / (TN + FP).Precision measures the proportion of correctly identified TB cases among all the predicted TB cases. It is calculated as TP / (TP + FP).Accuracy measures the overall correct predictions, both for TB and non-TB cases. It is calculated as (TP + TN) / (TP + TN + FP + FN).The F1 score is the harmonic mean of precision and recall. It is a useful metric when there is an imbalance between TB and non-TB cases. It is calculated as 2 * (Precision * Recall) / (Precision + Recall). Figure 3 shows the confusion matrix of ViTB16 and ViTB32. Table 1 shows the different performance metrics values.

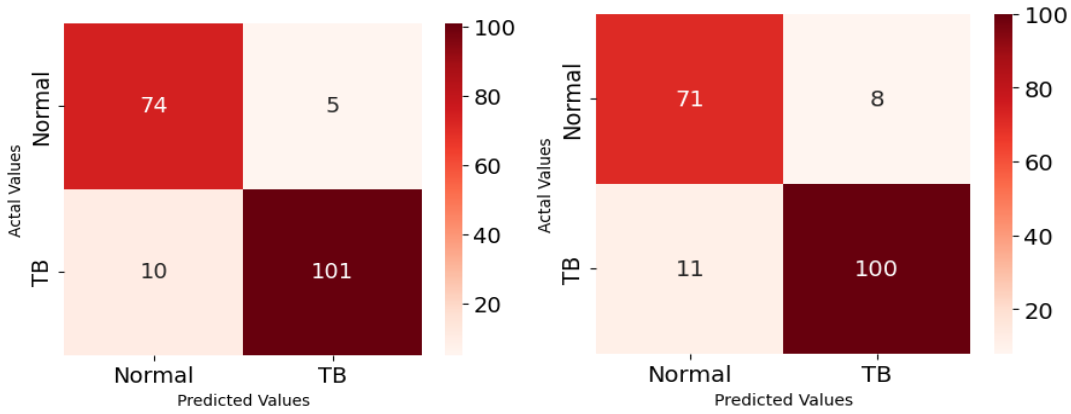

Figure 3: Confusion matrix of two class classification.

Table 1: performance metrics of differentViT models

Model	Accuracy	Recall	Precesion	F1-score	AUROC	AUPRC	Kappa
ViT-B16	0.924	0.946	0.966	0.956	0.914	0.972	0.623
ViT-B32	0.913	0.98	0.84	0.976	0.9441	0.991	0.639
Ensemble Of ViT Models	0.9644	0.98	0.97	0.965	0.967	0.993	0.658

7. Conclusion and Future Directions

The integration of ViT models in TB image classification using CXR images represents a significant advancement in the field of medical imaging. The ability of ViT to capture complex patterns and long-range dependencies makes it a promising tool for automated TB diagnosis. Continued research efforts, focusing on addressing existing challenges and exploring innovative approaches, are essential to fully harness the potential of ViT in improving TB diagnosis and patient outcomes. The use of Vision Transformers for TB image classification using CXR images represents a novel and promising approach. ViTs offer the potential to improve the accuracy of TB diagnosis and contribute to more effective disease management. Continued research in this field, including the development of specialized ViT models, data augmentation techniques, and interpretability tools, will be essential to fully leverage the capabilities of these models for TB diagnosis. With ongoing advancements in deep learning and medical imaging, ViTs may play a pivotal role in the fight against tuberculosis.

Continued research efforts should focus on refining ViT models, addressing dataset limitations, and exploring interpretability techniques. Additionally, collaborative initiatives between clinicians and machine learning experts can further enhance the practical application of ViT in TB diagnosis and monitoring. In conclusion, the use of Vision Transformers for TB image classification using CXR images represents a novel and promising approach. ViTs offer the potential to improve the accuracy of TB diagnosis and contribute to more effective disease management. Continued research in this field, including the development of specialized ViT models, data augmentation techniques, and interpretability tools, will be essential to fully leverage the capabilities of these models for TB diagnosis. With ongoing advancements in deep learning and medical imaging, ViTs may play a pivotal role in the fight against tuberculosis. Despite promising results, challenges such as dataset scarcity, class imbalance, and interpretability persist. Integrating domain knowledge, transfer learning, and attention mechanism refinement are potential strategies to address these challenges[16]. Interpreting deep learning models for medical image analysis is vital for gaining clinicians' trust and ensuring the transparency of decision-making processes. Techniques for model interpretability, such as Grad-CAM and feature visualization, have been explored to provide insights into the predictions made by ViTs.The scarcity of large, annotated TB image datasets remains a significant hurdle. Additionally, addressing class imbalance and handling data from diverse sources, including different imaging devices and quality levels, requires further investigation.

References

[1] E. Çallı, E. Sogancioglu, B. van Ginneken, K. G. van Leeuwen, and K. Murphy, “Deep

learning for chest X-ray analysis: A survey," *Med. Image Anal.*, vol. 72, p. 102125, 2021, doi: 10.1016/j.media.2021.102125.

[2] L. Brunese, F. Mercaldo, A. Reginelli, and A. Santone, "Explainable Deep Learning for Pulmonary Disease and Coronavirus COVID-19 Detection from X-rays," *Comput. Methods Programs Biomed.*, vol. 196, p. 105608, 2020, doi: 10.1016/j.cmpb.2020.105608.

[3] A. Dosovitskiy *et al.*, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," 2021, [Online]. Available: http://arxiv.org/abs/2010.11929.

[4] L. T. Duong, N. H. Le, T. B. Tran, V. M. Ngo, and P. T. Nguyen, "Detection of tuberculosis from chest X-ray images: Boosting the performance with vision transformer and transfer learning," *Expert Syst. Appl.*, vol. 184, no. July, p. 115519, 2021, doi: 10.1016/j.eswa.2021.115519.

[5] K. He *et al.*, "Transformers in Medical Image Analysis: A Review," pp. 1–19, 2022, [Online]. Available: http://arxiv.org/abs/2202.12165.

[6] X. LI, C. F. LIU, L. GUAN, S. WEI, X. YANG, and S. Q. LI, "Deep Learning in Chest Radiography: Detection of Pneumoconiosis," *Biomed. Environ. Sci.*, vol. 34, no. 10, pp. 842–845, 2021, doi: 10.3967/bes2021.116.

[7] J. Zhang *et al.*, "Viral Pneumonia Screening on Chest X-Rays Using Confidence-Aware Anomaly Detection," vol. 40, no. 3, pp. 879–890, 2021, doi: 10.1109/TMI.2020.3040950.

[8] G. Liang and L. Zheng, "A transfer learning method with deep residual network for pediatric pneumonia diagnosis," *Comput. Methods Programs Biomed.*, vol. 187, p. 104964, 2020, doi: 10.1016/j.cmpb.2019.06.023.

[9] D. M. Ibrahim, N. M. Elshennawy, and A. M. Sarhan, "Deep-chest: Multi-classification deep learning model for diagnosing COVID-19, pneumonia, and lung cancer chest diseases," *Comput. Biol. Med.*, vol. 132, p. 104348, 2021, doi: 10.1016/j.compbiomed.2021.104348.

[10] S. Rajaraman, G. Zamzmi, L. R. Folio, and S. Antani, "Detecting Tuberculosis-Consistent Findings in Lateral Chest X-Rays Using an Ensemble of CNNs and Vision Transformers," *Front. Genet.*, vol. 13, no. February, pp. 1–13, 2022, doi: 10.3389/fgene.2022.864724.

[11] L. Vogado, F. Araújo, P. S. Neto, J. Almeida, J. M. R. S. Tavares, and R. Veras, "A ensemble methodology for automatic classification of chest X-rays using deep learning," *Comput. Biol. Med.*, vol. 145, no. December 2021, p. 105442, 2022, doi: 10.1016/j.compbiomed.2022.105442.

[12] M. K. Mahbub, M. Biswas, L. Gaur, F. Alenezi, and K. C. Santosh, "Deep features to detect pulmonary abnormalities in chest X-rays due to infectious diseaseX: Covid-19, pneumonia, and tuberculosis," *Inf. Sci. (Ny).*, vol. 592, pp. 389–401, 2022, doi: 10.1016/j.ins.2022.01.062.

[13] S. I. Nafisah and G. Muhammad, "Tuberculosis detection in chest radiograph using convolutional neural network architecture and explainable artificial intelligence," *Neural Comput. Appl.*, vol. 6, 2022, doi: 10.1007/s00521-022-07258-6.

[14] S. Rajaraman, P. Ganesan, and S. Antani, "Deep learning model calibration for improving performance in class-imbalanced medical image classification tasks," *PLoS One*, vol. 17, no. 1 January, pp. 1–34, 2022, doi: 10.1371/journal.pone.0262838.

[15] S. Rajaraman, G. Cohen, L. Spear, L. Folio, and S. Antani, "DeBoNet: A deep bone suppression model ensemble to improve disease detection in chest radiographs," *PLoS One*, vol. 17, no. 3 March, pp. 1–22, 2022, doi: 10.1371/journal.pone.0265691.

[16] S. Bharati, P. Podder, and M. R. H. Mondal, "Hybrid deep learning for detecting lung diseases from X-ray images," *Informatics Med. Unlocked*, vol. 20, p. 100391, 2020, doi: 10.1016/j.imu.2020.100391.