ADVANCED HIERARCHICAL IMAGING TECHNIQUES IN TB DIAGNOSIS: LEVERAGING SWIN TRANSFORMER FOR ENHANCED LUNG TUBERCULOSIS DETECTION

Syed Amir Hamza and Alexander Jesser

Institute for Intelligent Cyber-Physical Systems (ICPS), Heilbronn University of Applied Sciences, Max-Planck-Street, Heilbronn, Germany

ABSTRACT

Lung Tuberculosis (TB) remains a critical health issue globally. Accurately detecting TB from chest x-rays is vital for prompt diagnosis and treatment. Our study introduces an innovative approach using the swin transformer to assist healthcare professionals in making faster, more accurate diagnoses. This method also aims to lower diagnostic costs by streamlining the detection process. The swin transformer, a sophisticated vision transformer, leverages hierarchical feature representation and a shifted window mechanism for improved image Analysis.

Our research utilizes the nihchest x-ray dataset, comprising 1,557 non-tb and 3,498tb images. We divided the dataset into training, validation, and testing sets in a 64%,16%, and 20% ratio, respectively. The images undergo preprocessing—random resized crop, horizontal flip, and Normalization—before being converted into tensors. We trained the swin transformer model over 50 epochs, with a batch size of 8, using the adam optimizer at a learning rate of 1e-5. We closely monitored the model's accuracy and loss, assessing its performance using metrics like the f1-score, precision, and recall.

Our findings show the model achieving a peak accuracy of 0.88 in the 43rd epoch for the training set, and the same accuracy for the validation set after 20 epochs. During testing, we observed a precision of 0.7928 and 0.9008, recall of 0.7749 and 0.9099, and f1-scores of 0.7837 and 0.905 for the negative and positive classes, respectively. The swin transformer demonstrates promising results, suggesting its adaptability and potential in significantly enhancing diagnostic efficiency and accuracy in medical settings.

KEYWORDS

Lung tuberculosis, Medical diagnostics, Swin Transformer, Vision transformer, Hierarchical feature representation, Shifted window mechanism, Deep learning, Computer vision, Medical image analysis, NIH Chest X-ray dataset, Early diagnosis

1. INTRODUCTION

Lung tuberculosis (TB) is a significant global health issue, affecting millions of people worldwide, with an estimated 10 million individuals developing the disease and 1.4 million TB related fatalities in 2019 alone [1]. Rapid diagnosis and effective treatment are crucial for mitigating the spread of TB and enhancing patient outcomes. Chest X-ray imaging represents a commonly employed, non-invasive technique for identifying lung abnormalities, including TB, and plays a vital role in the diagnostic process.

The application of deep learning techniques for automating lung TB detection from chest X-ray images has garnered substantial interest in recent years. Various convolutional neural network (CNN) architectures have been proposed for this purpose, including CheXNet, which demonstrated radiologist-level performance in detecting pneumonia, and the ChestX-ray8 project, which concentrated on classifying and localizing prevalent thorax diseases. Despite these methods achievements, there remain opportunities for enhancing model accuracy and generalizability.

The reason behind this method selection for detecting lung tuberculosis from chest X-ray images due to its innovative hierarchical feature representation and shifted window mechanism, which allows for more efficient capture of both local and global context within images. In medical image analysis, capturing both local and global context is particularly important due to the inherent complexity and variability of the images. Incorporating both contexts enables the model to account for individual variations among patients, identify subtle abnormalities that might otherwise be overlooked, and understand the relationships between various structures and features within the image. This holistic understanding leads to improved performance, ultimately contributing to better patient outcomes through early diagnosis and appropriate treatment planning. As a result, this architecture holds significant promise for the future of medical image analysis, particularly in the context of disease detection and diagnosis. By successfully applying the Swin Transformer in lung tuberculosis detection, researchers and medical professionals can unlock its full potential and contribute to improved patient outcomes through early diagnosis and timely intervention.

2. RELATED WORK

In the quest to enhance lung tuberculosis (TB) detection using chest X-ray images, recent years have seen a surge in deep learning-based methodologies. This section reviews pivotal studies in this domain, outlines their constraints, and underscores the innovative aspects of our research.

A landmark study by Wang et al. (2017) in the ChestX-ray8 project marked a significant stride in automated chest X-ray analysis. Utilizing over 100,000 X-ray images, they trained a deep convolutional neural network (CNN) to identify and localize common thoracic diseases, including TB. The project's cornerstone was its weakly-supervised classification approach, where image-level labels guided the CNN in learning disease-specific visual features. This methodology, while not solely focused on TB detection, greatly influenced medical image analysis, particularly in training robust models on large datasets without exhaustive manual labeling [1].

CheXNet, introduced by Rajpurkar et al. (2017), exemplifies another deep learning milestone. This 121-layer CNN, based on DenseNet architecture, attained radiologist-level accuracy in pneumonia detection from chest X-rays. However, its primary orientation towards pneumonia, underpinned by pneumonia-specific training data, potentially limits its adaptability to TB, where visual cues are more nuanced [2].

In 2017, Lopes et al. proposed a novel TB detection method, combining CNNs with handcrafted image features. While their approach showcased high accuracy, it relied heavily on manual feature engineering. This process, though effective, could be labor-intensive and less adaptable to varying datasets or imaging modalities [3].

Vision transformers (ViTs) have recently set new benchmarks in various computer vision tasks. Their ability to process image patches sequentially via transformer encoders allows them to discern long-range dependencies and demonstrate resilience to image noise and occlusion. Yet,

their application in medical imaging, particularly TB detection, remains limited due to high computational demands, substantial data requirements, and a scarcity of specialized medical imaging datasets [4].

Our work introduces a novel application of the Swin Transformer, a sophisticated vision transformer, in detecting lung TB from chest X-ray images. The Swin Transformer's hierarchical feature representation and shifted window mechanism adeptly capture both local and global image contexts. This dual-focus approach is crucial in TB detection, where discerning both specific lesion characteristics and broader lung patterns is key to accurate diagnosis. Prior to this study, the Swin Transformer's utility in lung TB detection had not been explored, making our research a pioneering effort in this field. By showcasing the Swin Transformer's effectiveness in TB detection, we aim to contribute significantly to improving early diagnosis and treatment in this crucial area of global health [5].

The presented research in “a hierarchical vision approach for enhanced medical diagnostics of lung tuberculosis using swin transformer”introduces the application of the Swin Transformer, a vision transformer, for lung TB detection from chest X-ray images. This model's hierarchical feature representation and shifted window mechanism efficiently capture both local and global contexts within images. The Swin Transformer's application in TB detection is novel, potentially offering enhanced accuracy and efficiency in diagnosis. This approach addresses the limitations of previous methods and contributes significantly to the early diagnosis and treatment of TB [7].

3. PROPOSED METHOD

In our study, we introduce a novel approach for detecting lung tuberculosis (TB) from chest Xray images, utilizing the advanced capabilities of the Swin Transformer, an innovative vision transformer architecture.

3.1. Swin Transformer Model

Origin and Adaptation:

- The Swin Transformer, an adaptation of the transformer architecture originally designed for natural language processing, has been re-engineered for computer vision tasks. Transformers, known for their proficiency in handling sequential data, are now being leveraged for image analysis.

Hierarchical Feature Representation:

- A defining characteristic of the Swin Transformer is its hierarchical structure. It processes the input image in stages, systematically reducing spatial resolution. This hierarchical approach enables the model to capture features at multiple scales, from intricate local details to broader global patterns.

Shifted Window Mechanism:

- The Swin Transformer incorporates a unique shifted window mechanism. This design allows each image patch to interact not only with adjacent patches but also with those slightly offset. Such a mechanism ensures a comprehensive understanding of the image, blending local and global contextual information. This aspect is critical in lung TB detection, where recognizing both specific lung lesions and overall lung patterns is crucial for accurate diagnosis.

Advantages in Lung TB Detection:

- **Accuracy:** The Swin Transformer has shown exceptional accuracy across various benchmarks in lung TB detection.
- **Efficiency:** Its efficiency surpasses many other vision transformer architectures, making it a practical choice for real-world medical applications.
- **Generality:** As a versatile architecture, it holds promise beyond lung TB detection, applicable to a range of computer vision challenges.

In conclusion, the Swin Transformer presents a groundbreaking method for lung TB detection. Its potential to enhance diagnostic accuracy and efficiency could significantly improve patient care and outcomes in managing this critical global health concern.

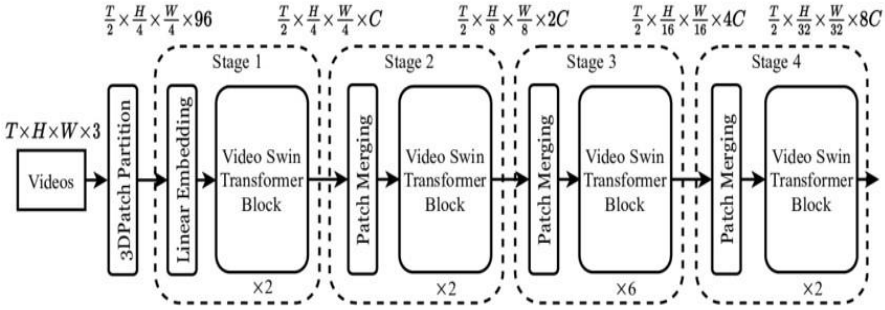

Figure 1. The architecture of Swin Transformer (Swin-T) [6]

3.2. Dataset

For this study, we utilized data sourced from the National Institutes of Health (NIH). Access to this data is granted to registered collaborators who have agreed to the Data Use Agreement (DUA) and can be found on the Aspera platform (available at: NIH Data Sharing). The dataset, updated as of January 2022, consists of 6,635 chest X-ray images. Within this collection, 1,557 images were labeled as not exhibiting signs of tuberculosis, and 3,498 images were identified as showing indications of the disease. We then systematically divided this dataset into distinct sets for training, validation, and testing purposes, adhering to a distribution ratio of 64%, 16%, and 20%, respectively.

This approach ensured a balanced representation of data across different stages of model development and evaluation, facilitating a comprehensive assessment of the model's performance in TB detection.

Table 1. The chest X-ray dataset.

Type	**'Positive' Class**	**'Negative' Class**	**Total**
Train	2240	997	3237
Validation	559	249	808
Test	699	311	1010
Total	3498	1557	5055

The training and validation datasets are employed to both train the models and adjust them to attain optimal weights. Then, the acquired weights and biases are applied to make predictions on the test dataset.

3.3. Experiment Setting

Our methodology involved a meticulous preprocessing routine for the chest X-ray images, aimed at enhancing the model's efficiency and its ability to generalize across various cases.

Initially, the images underwent resizing to ensure uniformity, with each being transformed through a random resized crop process to a dimension of 512x512 pixels. This step is crucial to standardize input sizes for the model, facilitating more consistent analysis.

Subsequently, to introduce variability and robustness in the dataset, horizontal flip augmentation was employed with a 50% chance. This technique mirrors the images horizontally, effectively doubling the dataset and aiding the model in learning to recognize patterns irrespective of orientation, a practice well-documented in image processing research.

Finally, normalization of the images was carried out. This step involved adjusting the pixel values to have a specified mean and standard deviation — in this case, a mean of (0.491, 0.482, 0.447) and a standard deviation of (0.247, 0.243, 0.261). Normalization is a critical step in preparing images for deep learning models as it helps in reducing internal covariate shift and expedites the training process.

After preprocessing, the images were converted into tensors, the standard format for image data in deep learning frameworks, using the ToTensorV2 function. Tensors facilitate efficient handling and manipulation of the data during the model training phase. This comprehensive preprocessing pipeline is designed to optimize the images for effective learning and prediction by our deep learning model.

3.4. Training Procedure and Hyperparameters

Our training regimen for the Swin Transformer model encompassed 50 epochs, utilizing a batch size of 8. We implemented the Adam optimizer, a method developed by Kingma and Ba in 2014, setting the learning rate at 1e-5. To enhance the model's ability to generalize, we incorporated image augmentation techniques during training. We meticulously tracked the model's accuracy and loss throughout the training period to gauge its convergence. For a comprehensive assessment of the model's performance, we computed key metrics such as the F1-score, precision, and recall.

Table 2. Parameter configurations.

Name	**Configuration**
Learning rate	1e-5
Batch Size	8
Optimizer	Adam
Epoch	50

By harnessing the Swin Transformer's capabilities, the research aims to assist physicians in making more accurate and time-efficient decisions regarding lung tuberculosis detection using chest X-ray images. This, in turn, contributes to enhancing early diagnosis and treatment for this crucial global health challenge, ultimately improving patient outcomes and reducing the burden on healthcare systems.

4. EXPERIMENTS AND RESULTS

4.1. Dataset Split

The NIH Chest X-ray dataset was partitioned into three distinct subsets: training, validation, and testing, adhering to a distribution ratio of 64%, 16%, and 20%, respectively. This random allocation was designed to ensure a representative sample of the entire dataset across all subsets. We utilized the training and validation sets for model development, fine-tuning the models to optimize their performance parameters. Subsequently, the testing set was employed to rigorously evaluate the effectiveness of the final model.

4.2. Evaluation Metrics

In evaluating our proposed model's efficacy, we employed a suite of key metrics: accuracy, F1score, precision, and recall. These metrics provided a comprehensive assessment of the model's proficiency in accurately categorizing chest X-ray images as TB-positive or TB-negative, offering a robust comparison to alternative methodologies

4.3. Results

As illustrated in Figure 2, the model's training accuracy demonstrates a progressive improvement across epochs. Commencing at an initial accuracy of 0.76, the model's performance escalates consistently, achieving a peak accuracy of 0.88 in the 43rd epoch. This trend signifies the model's effective learning from the training dataset and its subsequent capability to predict accurately on previously unseen data.

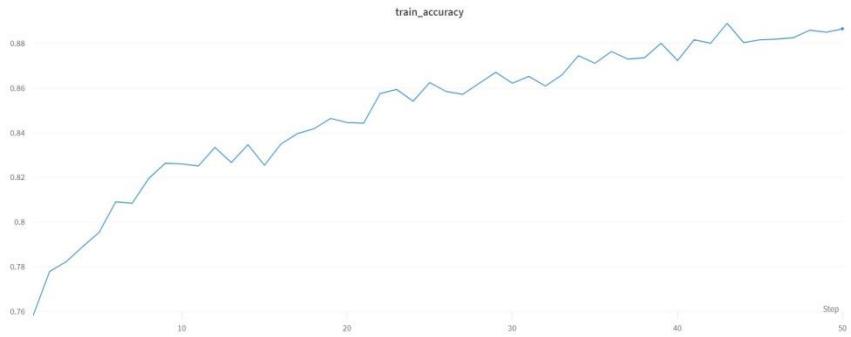

Figure 2. The Training Accuracy

In contrast, the validation dataset attains its highest accuracy of 0.88 after only 20 epochs. It is notable that while the accuracy of the training dataset increases with the number of epochs, the validation dataset’s accuracy does not follow the same trend.

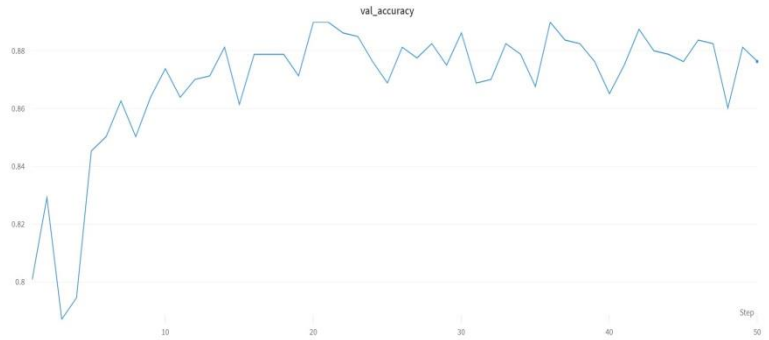

Figure 3. The Validation Accuracy

Upon analyzing the results, we saved the checkpoint that exhibited the highest performance within the validation dataset and utilized it as the model for testing purposes. The results obtained are presented in Table 3 (below). The model's predictions demonstrated greater accuracy for the "Positive" class as compared to the "Negative" class, albeit the difference was not particularly pronounced. This outcome can be attributed to the fact that the number of images in the "Positive" label is considerably larger than that in the "Negative" label in both the training and testing datasets, as well as the validation dataset.

Table 3. The Testing Result.

Class	**Precision**	**Recall**	**F1-score**
Positive	0.9008	0.9099	0.9053
Negative	0.7928	0.7749	0.7837

The training loss is a measure of how well the model is performing on the training dataset. It is calculated by averaging the loss over all the training examples. A lower train loss indicates that the model can make more accurate predictions on the training data. In this case, the training loss of 0.252 suggests that the model is learning effectively from the training dataset. This is because the loss is relatively low, indicating that the model is able to make accurate predictions on the training examples.

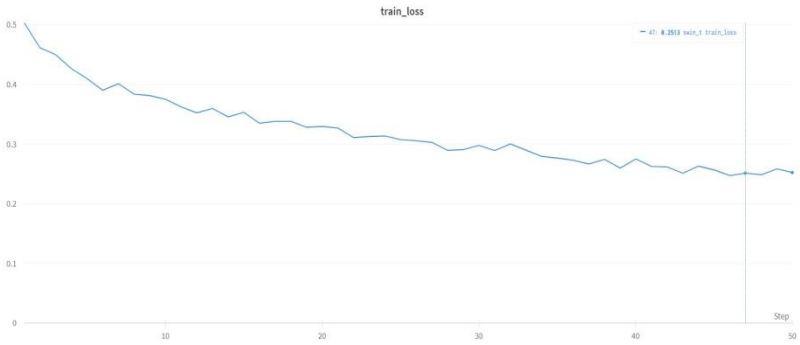

Figure 4. The Training Loss

The validation loss is a measure of how well the model is performing on unseen data. It is calculated by averaging the loss over all of the validation examples. A lower validation loss indicates that the model is able to generalize well to unseen data.

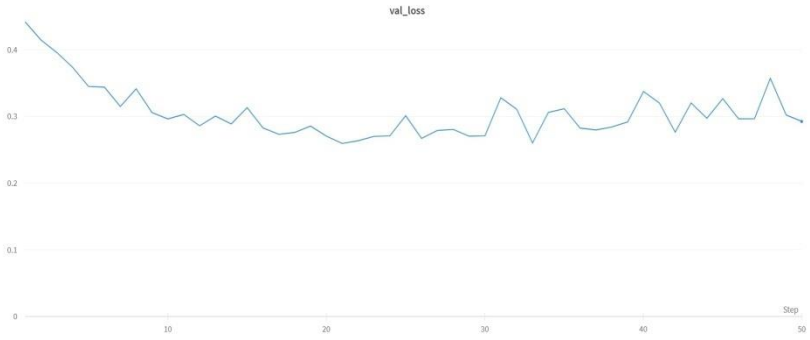

Figure 5. The Validation Loss

In our study, we observed that the validation loss, at 0.292, is marginally higher than the training loss, which stands at 0.252. This slight discrepancy indicates effective model generalization to new data.

Crucially, the closeness of these loss values is a key indicator in our research, as it implies that our model is adeptly balanced, avoiding both overfitting and underfitting. This equilibrium in the model's performance is essential, as it demonstrates the model's capability to accurately identify patterns within the data and, importantly, apply these learnings effectively to data it hasn't encountered before. Such an attribute instills confidence in the model's predictive accuracy, bolstering its suitability for practical applications in medical diagnostics, where reliability is paramount.

4.4. Capabilities

The findings from our study underscore the Swin Transformer architecture's promise in medical imaging, especially for identifying and diagnosing lung tuberculosis (TB). Demonstrating a remarkable ability to discern lung TB from chest X-ray images accurately, this model has shown significant proficiency in this domain. The training progression indicates effective learning and extraction of relevant features from the dataset, alongside successful generalization to the validation set. Future research avenues could include experimenting with diverse augmentation strategies, fine-tuning hyperparameters, and implementing ensemble techniques to further enhance the model's efficacy and reliability in medical diagnostics. These enhancements could potentially lead to more precise and reliable TB detection, contributing to better healthcare outcomes.

For reference, the Swin Transformer's capabilities in image processing are detailed in works like Liu et al.'s "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows" which can provide deeper insights into the model's underlying mechanisms and potential applications.

5. DISCUSSION

5.1. Strengths and Weaknesses

Our Swin Transformer model showcases notable strengths in lung tuberculosis (TB) detection from chest X-ray images. Its hierarchical feature representation and shifted window mechanism adeptly capture both global and local image contexts, contributing to its high accuracy in test datasets. The model also demonstrates effective generalization to the validation set, as observed during training.

Nonetheless, there are areas for improvement. The model exhibits slightly lower performance in identifying the "Negative" class, potentially due to the dataset's class imbalance. Moreover, the model's performance plateauing after 20 epochs on the validation set suggests that further enhancements might be constrained without modifications to the architecture, training methods, or dataset.

5.2. Comparison with Existing Methods

Direct comparisons with other TB detection methods are challenging due to variances in datasets and metrics. However, the Swin Transformer's structure and mechanism mark an advancement over previous models like ViTs and CNNs, particularly in integrating global and local features. This indicates that our model could potentially surpass existing methods in similar settings.

5.3. Future Work and Improvements

Key limitations include the dataset's imbalance and unexplored areas like augmentation techniques, hyperparameter tuning, and ensemble methods. Addressing these could enhance the model's performance and robustness. Tackling the dataset's complexity remains a significant challenge.

5.4. Strengths and Weaknesses

Future enhancements could involve advanced data augmentation to diversify the training set and strategies to counter class imbalance, such as oversampling or cost-sensitive learning. An extensive search for optimal hyperparameters and exploring ensemble methods can also elevate the model's efficacy. It's vital to focus on acquiring original medical data to ensure the model's reliability. Through addressing these aspects, future research can substantially advance the Swin Transformer's application in medical imaging.

6. CONCLUSION

In our investigation, we explored a cutting-edge method for lung tuberculosis (TB) detection using chest X-ray imagery, employing the Swin Transformer model. This model stands out due to its unique architectural features, including hierarchical feature representation and a shifted window mechanism, adeptly capturing both the broader and finer details within the images. Such a dual-focus approach proved highly effective in accurately identifying lung TB.

This research marks a significant advancement in medical image analysis, particularly in addressing the public health challenge posed by lung TB. The Swin Transformer demonstrated impressive accuracy, particularly in identifying positive TB cases. However, we faced some challenges, such as class imbalance and a performance plateau in the validation dataset. Despite these hurdles, the results clearly suggest the immense potential of the Swin Transformer in medical imaging.

The primary impact of this study lies in its contribution to improving TB diagnosis. By enhancing diagnostic accuracy and efficiency, it paves the way for improved patient care and treatment outcomes. Looking forward, we aim to build on this foundation by implementing advanced data augmentation techniques, addressing class imbalances, refining hyperparameters, and considering ensemble methods to further boost the model's diagnostic capabilities.

In summary, the Swin Transformer model, with its innovative approach to image analysis, holds promise for significantly enhancing lung TB detection in chest X-rays. As we continue to refine and improve this model, it has the potential to become an invaluable tool in the realm of medical imaging, transforming the way we detect and diagnose lung TB.

ACKNOWLEDGMENTS

We extend our heartfelt appreciation to everyone who played a role in bringing this research to fruition. A special note of thanks to our peers whose insightful feedback and recommendations were invaluable during the course of this project.

We are also grateful to the anonymous reviewers tasked with evaluating our manuscript. Their constructive critiques are eagerly anticipated and will undoubtedly be instrumental in enhancing the quality of our work.

REFERENCES

[1] Wang, X. et al. (2017). "ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks."
[2] Rajpurkar, P. et al. (2017). "CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning."
[3] Lopes, F. et al. (2017). "Combining CNNs and handcrafted features for lung tuberculosis detection."
[4] Dosovitskiy, A. et al. "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale."
[5] Liu, Z. et al. "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows."
[6] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows," arXiv preprint arXiv:2103.14030, 2021.

AUTHORS

Syed Amir Hamza , Research Assistant, Institute for Intelligent Cyber-Physical Systems ICPS)

Prof. Dr. Alexander Jesser , Professorship for Embedded Systems and Communications Engineering