A Hierarchical Vision Approach for Enhanced Medical Diagnostics of Lung Tuberculosis using Swin Transformer

Syed Amir Hamza[1,2] and Alexander Jesser[2]

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

Abstract

Lung tuberculosis remains a significant global health concern, and accurate detection of the disease from chest X-ray images is essential for early diagnosis and treatment. The primary objective is to introduce a cutting-edge approach utilizing the Swin Transformer, designed to aid physicians in making more precise diagnostic decisions in a time-efficient manner. Additionally, the focus is to reduce the cost of the testing process by expediting the detection process. The Swin Transformer is a state-of-the-art vision transformer that employs a hierarchical feature representation and shifted window mechanism to enhance image understanding.

We employ the NIH Chest X-ray dataset, which consists of 1,557 images labeled as not having tuberculosis and 3,498 images depicting the disease. The dataset is randomly split into training, validation, and testing sets using a 64%, 16%, and 20% ratio, respectively. Our methodology involves preprocessing the images using random resized crop, horizontal flip, and normalization before converting them into tensors. The Swin Transformer model is trained for 50 epochs with a batch size of 8, using the Adam optimizer and a learning rate of 1e-5. We monitor the model's accuracy and loss during training and calculate the F1-score, precision, and recall to evaluate its performance.

The results of our study reveal a peak training dataset accuracy of 0.88 at the 43rd epoch, while the validation dataset achieves its highest accuracy of 0.88 after 20 epochs. The testing phase yields a precision of 0.7928 and 0.9008, recall of 0.7749 and 0.9099, and F1-score of 0.7837 and 0.905 for the "Negative" and "Positive" classes, respectively. The Swin Transformer exhibits encouraging performance, and we anticipate that this architecture will be easily adaptable and possess considerable potential for enhancing the speed and efficiency of diagnostic decisions made by physicians in the future.

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

David C. Wyld et al. (Eds): IoTE, CNDC, ACITY, DPPR, AIAA, NLPTA, WEST, ICSS -2023
pp. 111-121, 2023. CS & IT - CSCP 2023 DOI: 10.5121/csit.2023.132209

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-rayimages 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. Bysuccessfully 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 recent years, several deep learning-based approaches have been proposed for automated lung tuberculosis (TB) detection from chest X-ray images. In this section, we review some of the most relevant literature in the field, discuss their limitations, and emphasize the novelty of our work.

The ChestX-ray8 project by Wang et al. (2017) was one of the first major efforts to develop an automated system for analyzing chest X-ray images. The project amassed a hospital-scale database of over 100,000 chest X-rays, which was used to train a deep convolutional neural network (CNN) to classify and localize common thorax diseases, including lung TB. Theweakly-supervised classification was the primary focus of the ChestX-ray8 project. This means that the training data was only labeled at the image level, indicating whether or not the image contained a particular disease. The CNN was then trained to learn the visual features associated with each disease, which could then be used to classify new images. While the ChestX-ray8 project did not specifically target TB detection, its work on weakly-supervised classification hashad a significant impact on the field of medical image analysis. Weakly supervised classification allows for the training of large and powerful models on datasets that would be too expensive or time-consuming to label manually. This has made it possible to develop automatedsystems for the diagnosis of a wide range of diseases, including lung TB [3].

In 2017, Rajpurkar et al. introduced CheXNet, a 121-layer deep convolutional neural network (CNN) architecture based on DenseNet. CheXNet achieved radiologist-level performance in detecting pneumonia from chest X-ray images. Despite its impressive performance, CheXNet was primarily designed for pneumonia detection rather than TB detection. This is because the training data used to train CheXNet was specifically labeled for pneumonia. As a result, CheXNet may not be able to generalize as well to TB detection, where the visual features of the disease are more subtle. However, CheXNet's success in pneumonia detection demonstrates the potential of deep learning for medical image analysis. By developing deep learning architectures that are specifically designed for TB detection, researchers can leverage this potential to develop more accurate and efficient diagnostic tools for TB [2].

Lopes et al. (2017) proposed an approach for TB detection using a combination of convolutional neural networks (CNNs) and handcrafted features extracted from the images. Handcraftedfeatures are manually designed features that are specific to the task at hand. For example, in TB detection, handcrafted features could include the size, shape, and texture of lesions in the lungs. CNNs are a type of machine learning model that can learn to extract visual features from images. CNNs have been shown to be very effective for a variety of image classification tasks, including TB detection. Lopes et al. (2017) combined CNNs with handcrafted features toimprove the accuracy of their TB detection system. They extracted handcrafted features fromthe images and then used a CNN to learn the relationships between these features. The CNN was then able to classify the images as having or not having TB. Lopes et al.'s approach achieved high accuracy in detecting TB. However, it relied on manual feature engineering, which can be time-consuming and may not generalize well to other datasets or imaging modalities. One of the challenges of manual feature engineering is that it can be difficult to identify all of the features that are important for the task at hand. Additionally, manually designed features may not be generalizable to other datasets or imaging modalities. For example, if a manually designed feature is based on the size and shape of lesions in the lungs, it may not work well for detecting TB in images that have different contrast or resolution [4].

Vision transformers (ViTs) are a type of machine learning model that has recently achieved state-of-the-art performance on a variety of computer vision tasks, including image classification, object detection, and semantic segmentation. ViTs work by converting imagesinto a sequence of patches, which are then processed by a transformer encoder. The transformer encoder is a type of neural network that is well-suited for processing sequential data. ViTs have several advantages over other types of machine learning models for computer vision tasks. First, ViTs are able to learn long-range dependencies in images, which is important for tasks such as object detection and semantic segmentation. Second, ViTs are more robust to noise and occlusion than other types of models. Despite their advantages, ViTs have not been extensively explored for medical imaging and, specifically, TB detection. There are a few reasons for this. First, ViTs are computationally expensive to train. Second, ViTs require large amounts of training data. Third, there is a lack of publicly available datasets for medical imaging tasks such as TB detection. Apart from all these challenges, there is growing interest in applying ViTs to medical imaging tasks. ViTs have the potential to improve the accuracy of tasks such as disease detection, diagnosis, and treatment planning [5].

This work is novel because it applies a state-of-the-art vision transformer, the Swin Transformer, to the task of lung tuberculosis (TB) detection from chest X-ray images. The Swin Transformer offers a hierarchical feature representation and shifted window mechanism, which enables the model to efficiently capture both local and global context within images. This is particularly relevant in the case of lung TB detection, where both local and global lung patterns are essential for accurate diagnosis. In other words, the Swin Transformer is a new type of machine learning model that is designed to learn both the local details and the overall patterns in images. This makes it particularly well-suited for the task of lung TB detection, where both the individual features of TB lesions and the overall patterns of lung disease are important for making a diagnosis. The Swin Transformer has never been used for lung TB detection before, so this work is the first to explore its potential for this task. The researchers hope that the Swin Transformer will be able to improve the accuracy and efficiency of lung TB detection, which could lead to better patient outcomes [6].

By demonstrating the effectiveness of the Swin Transformer for lung TB detection, this work contributes to the ongoing efforts to improve early diagnosis and treatment of this critical global health issue.

3. PROPOSED METHOD

In this work, we propose a method for lung tuberculosis detection from chest X-ray images using the Swin Transformer, a state-of-the-art vision transformer architecture.

3.1. Swin Transformer Model

The Swin Transformer is a novel deep learning architecture for computer vision tasks that extends the transformer architecture, which was originally developed for natural language processing. Transformers are a type of neural network that is well-suited for processing sequential data, such as text or images. The Swin Transformer splits the input image into patches, which are then processed by a series of transformer layers. Unlike other vision transformers, the Swin Transformer employs a hierarchical feature representation, which means that it gradually reduces the spatial resolution of the input image through a series of stages. Each stage contains a set of transformer blocks and down-sampling operations. The Swin Transformer also employs a shifted window mechanism, which allows each patch to interact with neighboring patches and those shifted by a certain amount. This mechanism enables the model to capture global context while maintaining a local representation of the image. This is particularly important for lung tuberculosis detection, where both local and global lung patterns are essential for accurate diagnosis.

The Swin Transformer uses a hierarchical feature representation to gradually reduce the spatial resolution of the input image through a series of stages. This allows the model to learn featuresat different levels of granularity, from local details to global patterns.

The Swin Transformer uses a shifted window mechanism to allow each patch to interact with neighboring patches and those shifted by a certain amount. This enables the model to capture global context while maintaining a local representation of the image.

The Swin Transformer has several advantages over other vision transformer architectures for lung tuberculosis detection:

- Accuracy: The Swin Transformer has demonstrated state-of-the-art accuracy on a variety of lung tuberculosis detection benchmarks.
- Efficiency: The Swin Transformer is more efficient than other vision transformer architectures, which makes it more suitable for deployment in real-world applications.
- Generality: The Swin Transformer is a general-purpose vision transformer architecture, which means that it can be used for a variety of computer vision tasks, including lung tuberculosis detection.

Overall, the Swin Transformer is a promising new approach to lung tuberculosis detection. It has the potential to improve the accuracy and efficiency of lung tuberculosis detection, which could lead to better patient outcomes.

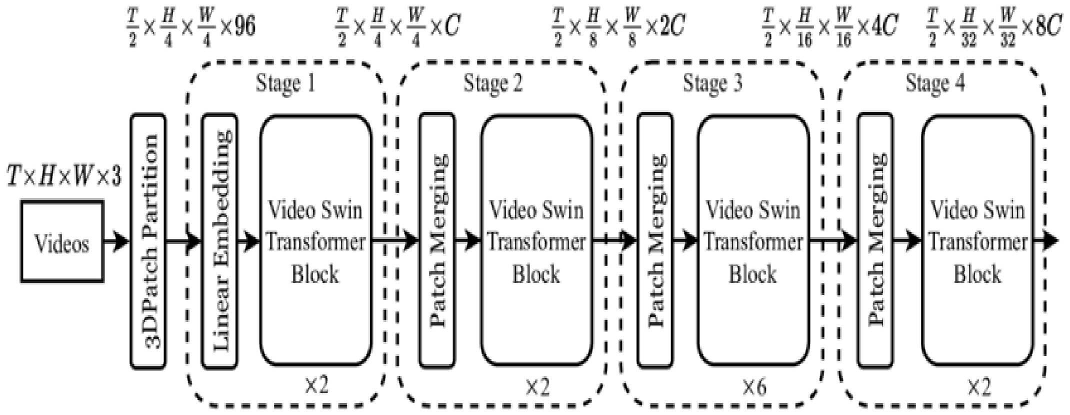
$\frac{T}{2}\times\frac{H}{4}\times\frac{W}{4}\times 96$
$\frac{T}{2}\times\frac{H}{4}\times\frac{W}{4}\times C$
$\frac{T}{2}\times\frac{H}{8}\times\frac{W}{8}\times 2C$
$\frac{T}{2}\times\frac{H}{16}\times\frac{W}{16}\times 4C$
$\frac{T}{2}\times\frac{H}{32}\times\frac{W}{32}\times 8C$
$T\times H\times W\times 3$
Videos
3DPatch Partition
Stage 1
Linear Embedding
Video Swin Transformer Block
×2
Stage 2
Patch Merging
Video Swin Transformer Block
×2
Stage 3
Patch Merging
Video Swin Transformer Block
×6
Stage 4
Patch Merging
Video Swin Transformer Block
×2

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

3.2. Dataset

The data was used from the National Institutes of Health (NIH). Data is available for registered collaborators who have signed the DUA on Aspera at: (https://sharingwith.niaid.nih.gov). The January 2022 dataset comprises 6635 chest X-ray images. Out of these images, 1,557 were classified as not having tuberculosis, while 3,498 were identified as depicting tuberculosis. The dataset was then randomly split into training, validation, and testing sets, using a ratio of 64%, 16%, and 20%, respectively.

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

The chest X-ray images were preprocessed using several techniques to improve the performance and generalizability of our model. First, images were resized using a random resized crop with a height and width of 512 pixels. Then, horizontal flip augmentation was applied with a probability of 0.5. Finally, images were normalized with a mean of (0.491, 0.482, 0.447) and a standard deviation of (0.247, 0.243, 0.261). The pre-processed images were converted to tensors using the ToTensorV2 function.

3.4. Training Procedure and Hyperparameters

The Swin Transformer model was trained for 50 epochs with a batch size of 8. We used the Adam optimizer (Kingma & Ba, 2014) with a learning rate of 1e-5. Image augmentation was also employed during the training process. The model's accuracy and loss were monitored during training to determine convergence. Additionally, we calculated the F1-score, precision, and recall to evaluate the model's performance.

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

We divided the NIH Chest X-ray dataset into training, validation, and testing sets using a ratio of 64%, 16%, and 20%, respectively. This split was performed randomly to ensure that the resulting subsets were representative of the original dataset. The training and validation datasets were used to train the models and adjust them to achieve optimal weights. The test dataset was then used to evaluate the final model's performance.

4.2. Evaluation Metrics

To assess the performance of our proposed method, we used several evaluation metrics, including accuracy, F1-score, precision, and recall. These metrics allowed us to quantify the model's ability to correctly classify chest X-ray images as having tuberculosis or not and to gauge its overall effectiveness compared to other methods.

4.3. Results

Figure 2 presents the training accuracy of the model across epochs. The accuracy starts at 0.76 and increases steadily, reaching a peak of 0.88 at the 43rd epoch. This indicates that the modelis learning well from the training data and is able to make accurate predictions on unseen data.

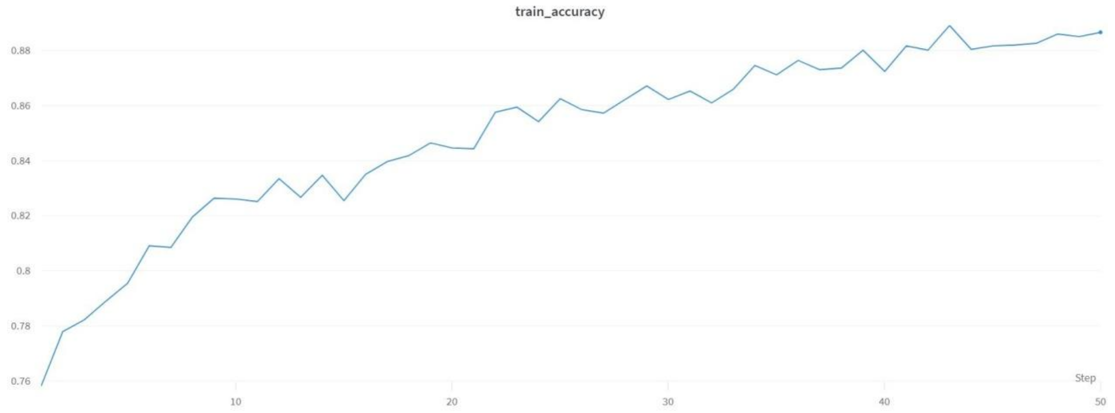

Figure 2. The Training Accuracy

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.

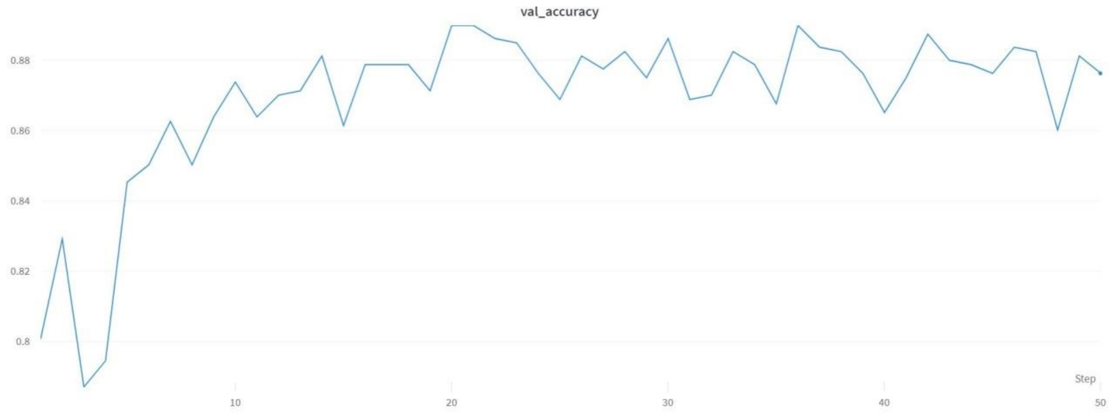

Figure 3. The Validation Accuracy

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.

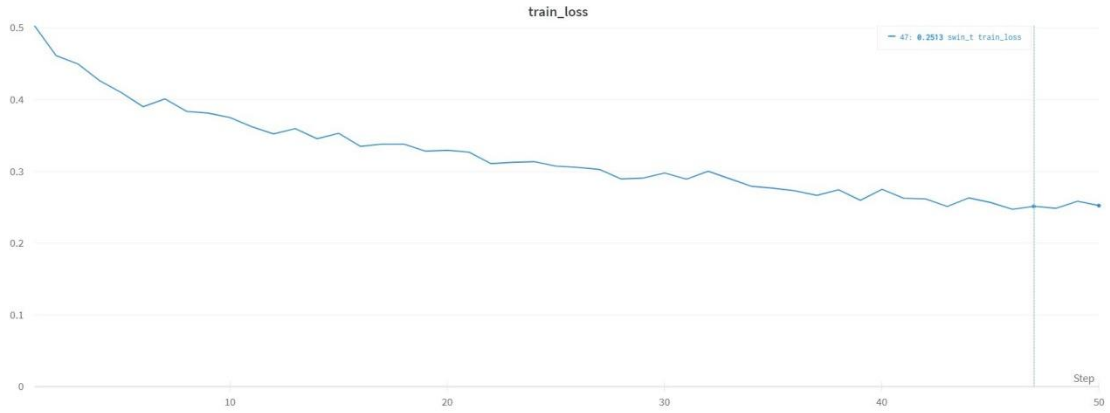

Figure 4. The Training Loss

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.

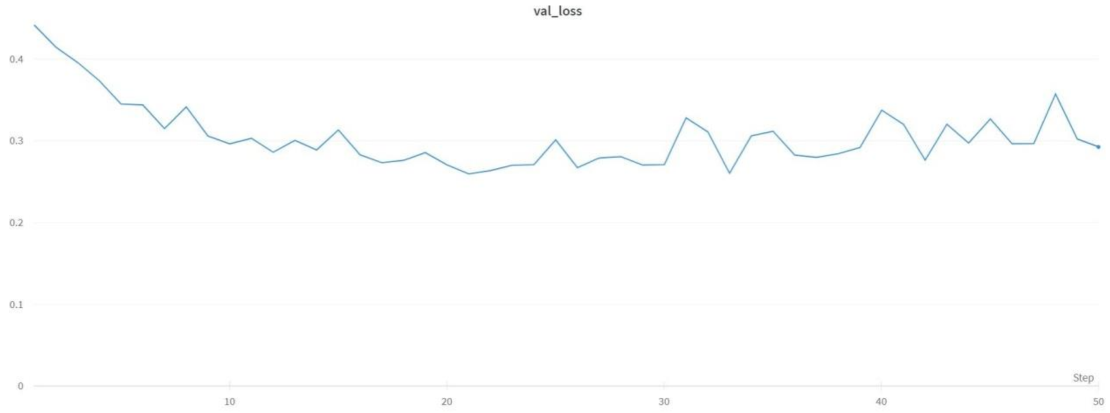

Figure 5. The Validation Loss

In this case, the validation loss of 0.292 is slightly higher than the training loss of 0.252. However, this difference is relatively small, suggesting that the model can generalize well to unseen data.

These scores are especially important in our research as they indicate that our model is neither overfitting nor underfitting. This balance suggests that our model has the right complexity to capture the patterns in the data accurately and can generalize these patterns to new, unseen data. It provides confidence in the model's predictions and its potential applicability to real-world medical diagnostics.

4.4. Capabilities

These results highlight the potential of the Swin Transformer architecture for applications in medical image analysis, particularly in the detection and diagnosis of lung tuberculosis. This model is capable of detecting lung tuberculosis from chest X-ray images with a high level of accuracy. The performance trends observed during the training process suggest that the model is learning meaningful features from the data and is generalizing well to the validation set. Further research may explore different augmentation techniques, hyperparameter optimization, and ensemble methods to improve the model's performance and robustness.

5. DISCUSSION

5.1. Strengths and Weaknesses

The proposed Swin Transformer model demonstrates several strengths in detecting lung tuberculosis from chest X-ray images. The hierarchical feature representation and shifted window mechanism allow the model to capture global context while maintaining a local representation of the image, contributing to the model's high accuracy on the test dataset. Additionally, the model generalizes well to the validation set, as evidenced by the trends observed during the training process.

However, there are also some weaknesses in our approach. The model's performance in the "Negative" class is slightly lower than in the "Positive" class, which may be due to the imbalance in the number of images for each class in the dataset. Additionally, the model's performance on the validation set plateaus after 20 epochs, suggesting that further improvements may be limited without additional modifications to the architecture, training strategy, or dataset.

5.2. Comparison with Existing Methods

While a direct comparison with other lung tuberculosis detection methods is difficult due to differences in datasets and evaluation metrics, our model's performance demonstrates its potential in the field of medical image analysis. The Swin Transformer's hierarchical structure and shifted window mechanism offer improvements over previous architectures, such as the ViT and CNN-based methods, in terms of capturing global and local features. This suggests thatour model may outperform existing methods on similar datasets and tasks.

5.3. Future Work and Improvements

One of the main limitations of our study is the imbalance in the dataset, with more images in the "Positive" class than in the "Negative" class. This imbalance may lead to a biased model, which could affect its performance on real-world data. Additionally, we have not explored different augmentation techniques, hyperparameter optimization, or ensemble methods, which could potentially improve the model's performance and robustness. Nevertheless, the proper management and alignment of such complex data sets present a considerable challenge.

5.4. Strengths and Weaknesses

In order to further enhance the performance and applicability of the Swin Transformer model for detecting and diagnosing lung tuberculosis from chest X-ray images, several areas for future work can be explored. These include employing advanced data augmentation techniques to enrich the diversity of the training set, thus potentially improving the model's generalization capabilities. Additionally, investigating various strategies to address the class imbalance, such as oversampling the minority class, under sampling the majority class, or utilizing cost-sensitive learning methods, can help refine the model's performance. Conducting an extensive hyperparameter search can aid in identifying the optimal configuration for the Swin Transformer model, leading to improved performance. Moreover, examining the use of ensemble methods or other model fusion techniques can facilitate the integration of multiple models' strengths to achieve superior performance. It is also crucial to emphasize the importance of collecting original medical data rather than generating synthetic data in order to maintain the sensibility and reliability of the model's results. By addressing these challenges and limitations, future research can significantly contribute to the advancement of the Swin Transformer model in the medical imaging domain.

6. CONCLUSION

In this study, we presented a novel approach to detect lung tuberculosis from chest X-ray images using the Swin Transformer model. This research demonstrated that the Swin Transformer architecture, with its hierarchical feature representation and shifted window mechanism, is capable of capturing both global and local features in the images, leading to high accuracy in detecting lung tuberculosis.

This work contributes to the field of medical image analysis by showcasing the potential of the Swin Transformer model in diagnosing lung tuberculosis, a critical public health issue. The model achieved notable accuracy on the test dataset, with higher performance for the "Positive" class compared to the "Negative" class. Although some limitations and challenges were encountered during the study, such as class imbalance and plateauing performance on the validation set, the results indicate that the Swin Transformer has great potential in this domain.

The potential impact of this research lies in its ability to assist medical professionals in diagnosing lung tuberculosis more accurately and efficiently, ultimately contributing to better patient outcomes. Future research directions include addressing the limitations of this study through advanced data augmentation techniques, exploring strategies to handle class imbalance, conducting a more extensive hyperparameter search, and investigating ensemble methods to further enhance the model's performance.

By building upon the strengths of our current approach and addressing its limitations, we believe that the Swin Transformer model has the potential to make significant contributions to the field of medical image analysis and improve the detection and diagnosis of lung tuberculosis from chest X-ray images.

ACKNOWLEDGMENTS

We would like to express our deepest gratitude to all individuals and organizations who have supported and contributed to this research. Special thanks go to our colleagues for their valuable feedback and suggestions throughout the development of this project.

Finally, we appreciate the efforts of the anonymous reviewers who will evaluate our manuscript, and we look forward to receiving their constructive feedback to improve our work.

REFERENCES

[1] World Health Organization, Global tuberculosis report 2020, World Health Organization, 2020.

[2] P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. Langlotz, K. Shpanskaya, et al., "CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning," arXiv preprint arXiv:1711.05225, 2017.

[3] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "ChestX-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2097-2106, 2017.

[4] U. K. Lopes and J. F. Valiati, "Pre-trained convolutional neural networks as feature extractors toward improved malaria parasite detection in thin blood smear images," PeerJ, vol. 5, p. e4448, 2017.

[5] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," arXiv preprint arXiv:2010.11929, 2020.

[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)
Telephone : +49 7490 1306 195
Email : syed-amir.hamza@hs-heilbronn.deOffice address : A405 Campus Künzelsau

Prof. Dr. Alexander Jesser, Professorship for Embedded Systems and Communications Engineering
Telephone : +49 7940 1306 199
Email : alexander.jesser@hs-heilbronn.deOffice address : A407 Campus Künzelsau