Journal of Scientific & Industrial Research
Vol. 85, January 2026, pp. 36-43
DOI: 10.56042/jsir.v85i1.5865

Explainability of Tuberculosis Diagnosis based on Chest X-Ray Images with Vision Transformer

Ramkumar Thirunavukarasu[1], Evans Kotei[2]* & Thillainayagam[3]

[1]Department of Computer Applications, School of Computer Science Engineering and Information Systems, Vellore Institute of Technology, Vellore, Tamil Nadu, India

[2]Department of Computer Science, Faculty of Applied Sciences and Technology, Kumasi Technical University, Kumasi, Ghana

[3]Department of Computer Applications, A.V.C. College of Engineering, Mayiladuthurai 609 305, India

Received 27 September 2023; revised 10 June 2024; accepted 05 January 2026

Chest X-ray radiography is a reasonably inexpensive and widely available diagnostic technology that can aid in identifying different illnesses like tuberculosis (TB), pneumonia, COVID-19 and many more. The demand for skilled personnel to evaluate X-ray radiographs is a challenge in many health facilities across the globe, particularly in underdeveloped regions. Machine Learning (ML) algorithms have enabled the automated diagnosis of TB from X-ray modalities. Aside from deep convolutional neural networks (DCNN) for vision applications, the Vision Transformer (ViT) network has also produced outstanding results in image classification. Motivated by the robustness of the transformer network on image processing tasks, the study proposes a transformer-based framework for early screening of TB disease. Three different vision transformer types ViT-Base16 (ViT-B16), ViT-Base32 (ViT-B32), and ViT-Large32(ViT-L32) were tested in the experiment to see how well they performed in identifying tuberculosis. When the transformer models' outcomes were contrasted with those of other CNNs, the VIT-B32 model performed admirably in the diagnostic procedure. The ViT-B32 model's attained accuracy, sensitivity, specificity, precision, F-1 score, and AUC scores of the ViT-B32 model were 96.96%, 96.89%, 97.01%, 96.72%, 96.80% and 0.97, respectively, on TB classification. The ViT-b32 model demonstrated superiority and generalizability. Because of its low cost and ease of use, the ViT-b32 model may provide an accurate diagnostic system to all TB patients for early screening.

Keywords: Deep learning, Machine learning, Medical image analysis, Self-attention, Transformer network

Introduction

Tb is a transmittable illness and is one of the principal causes of death globally. About 10.0 million individuals have the disease worldwide, with some suffering from drug-resistant tuberculosis (TB), which remains a public health concern. To eradicate tuberculosis (TB), the World Health Organization (WHO) has proclaimed that comprehensive health coverage for all is required.[1]

Chest X-ray (CXR) examination is a popular clinical imaging modality for detecting and diagnosing lung disorders.[2–6] Its main benefit is its low cost, accessibility, and simple operation. X-ray imaging is a method used to determine the origin of illness or an issue with one of the body's internal organs.[7] The interpretation of images relies on the radiologist, which makes the sensitivity solely dependent on their expertise. Examining the images is laborious and requires much time.

Furthermore, the need for professional radiologists might be a barrier, particularly in distant or underserved locations. Because of the above challenges, current research has concentrated on ML algorithms for automatic diagnosis and aspires to become a vital tool for physicians.[8–10]

The progress of high-performance computers with graphics processing units (GPUs) and the accessibility of clinical datasets is the driving force behind the development of artificial intelligence solutions for automatic disease identification. Increased sensitivity for results, automation of boring everyday activities, and resolving the problem of physicians not always present in distant places or underdeveloped nations are all possible benefits of automated disease diagnosis.[11,12]

Deep Learning (DL) approaches like CNN have produced outstanding results in computer vision-oriented applications in recent years, including classification, segmentation, and object recognition.[13–17] Because of this achievement, DL solutions have widely been implemented in the clinical

*Author for Correspondence
E-mail: evans.kotei@kstu.edu.gh

arena, such as for TB diagnosis.[18–20] A framework for TB prediction[21] got trained and evaluated using the Shenzhen and Montgomery County datasets.[22] Its accuracy was 90% and 80%, respectively.

A DL-based framework based on semi-supervised learning technique got proposed for TB identification using X-ray modalities. The model performed well by attaining a recall and specificity scores of 94.3%-100%, 91.1%-100%, and 99.1%-100%, respectively.[23]

The Gou *et al*.[24] proposed three ways to diagnose TB using X-ray radiography. Step 1 required modifying the physical features of the CNN model, while steps 2 and 3 optimized the model using an artificial bee colony approach. The system detected seven TB-related symptoms after training with the Shenzhen dataset (SZ)and Montgomery County (MC) datasets. A CNN model trained on the MC and SZ datasets, based on Bayesian theory, was introduced for optimal performance[25] in order to address the SoftMax inference limitation

ViTs[26] with built-in self-attention as an alternative to CNNs in computer vision applications produces cutting-edge results in vision tasks like segmentation, classification, and detection.[2,27–29] In its operations, the input image gets split into patch embeddings like tokens in natural language processing models. Position embedding added to the image patches transcends through a series of encoding layers for feature extraction. Following the ViT model's success, other variations with improved performance have arisen.

The data-efficient image transformers (DeiT)[30] model, which works on a teacher-student process, was introduced to help transformer models train on smaller datasets to produce better outcomes. DeiT performed better on ImageNet with an accuracy of 85.2%, making it better than the original ViT model. Instead of the direct tokenization used in the ViT model, a layer-wise Tokens-To-Token Vision Transformer (T2T-ViT) can encode each token.[31] The model was successful on ImageNet, with a Top-1 accuracy of 82.3%. On the other hand CrossViT, a to-way ViT model extracts multi-scale characteristics for optimal performance.[32] To quickly share information, a fusion approach based on cross-attention was devised to integrate image patch tokens of various sizes. The model performed well by achieving 82.8% Top-1 accuracy on ImageNet.

Notwithstanding the transformer network's capacity to simulate global interaction between token embeddings via self-attention, it suffers in modelling local representations, which is critical in image processing. The drawback gets addressed by introducing locality to the original ViT framework by including depth-wise convolution and non-linear activation functions into the model.[33] The model achieved 94.2% Top-5 accuracy on ImageNet.

In contrast, the Pyramid Vision Transformer (PVT) addresses the limitations of porting the transformer to diverse dense prediction workloads. On ImageNet, it had a Top-1 error rate of 18.3%.[34] A hierarchical Transformer is a general-purpose framework for computer vision where representations get calculated with Shifted windows.[35] The framework addresses the difficulties of converting the transformer from language to visual. Among the issues are variances amid the two spheres, such as notable changes in the size of visual elements and the higher pixel quality of photos than words in text.

Motivated by the performance of the transformer network on image classification, this study proposes an enhanced transformer model for TB identification from X-ray radiography. The contributions of the study are below.

- The study presents a transformer-based model with exceptional performance for TB diagnosis based on X-ray modalities.
- The model's resilience and generalization are compared to existing ViT-based frameworks, and the proposed model performed better.
- The study offers detailed performance evaluation with visualization to aid physicians and doctors offer correct decisions on image interpretations.

Materials and Methods

This section describes the vision transformer-based [26] architecture proposed for TB identification. The experimental dataset employed for the study is X-ray radiographs containing TB and no TB. The model's premise is to slice the input set equal patch-sizes with position embeddings.[36] In vision tasks, two dimensional (2D) images get reshaped as $X \in R^{H \times W \times C}$ into a linearly flattened 2D patches $X_p \in R^{N(P \times P \times C)}$, where (W, H) is the main image resolution, C is the channel, (P, P) is each pixel density of the patch, $N = HW/P^2$ is the resultant amount of patches, which also serves as the Transformer's operative input sequence length. A patch embedding is constructed by flattening the patches and projecting them to D dimensions using a trainable linear projection, as indicated in Eq. (1). The

framework has a Multi-head Self-Attention (MHSA) and Multilayer Perceptron (MLP), where each module employs a residual connection with normalization. The MLP has the Gaussian Error Linear Unit (GELU) activation function and two dense layers. The computational process is in Eq. (1).

$$Z = \left(Z_{class}\,; X_P^1 E; X_P^1 E; \ldots X_P^N E) + E_{pos}\right) \quad \ldots (1)$$
$$E = R^{(P \times P \times C) \times D}, E_{pos} \in R^{(N+1) \times D}$$
$$Z_I = MSA\left(LNZ(Zl-1)\right) + Zl - 11 = 1 \ldots L$$
$$Z_I = MLP\left(LN(zl)\right) + zl1 = 1 \ldots L$$
$$y = LN(z_k^0)$$

Dataset

The Shenzhen and Montgomery County[22], and Chest X-ray images for tuberculosis datasets[37] got pooled to establish a database for the experiment. The breakdown of each dataset is presented in Table 1. A sample of images from the database used in the experiment is shown in Fig. 1.

Model Architecture

Compared with CNNs that can extract local features and have translation equivariance, ViT has substantially a lesser amount inductive bias. Only the ViT structure's MLP levels are local and transnational equivariant, whilst the self-attention layers are global. At the start of the model, position embeddings are attached to the patches for spatial information extraction.

Table 1 — Database

Dataset	TB positive	TB negative	Total
Montgomery county	58	80	138
Shenzhen	336	326	662
Chest X-ray images for tuberculosis	700	800	1500
Total	1094	1206	2300

Three varieties of ViT models (ViT-16, ViT-32, and ViT-32) performed the feature selection and classification. The idea is to put these models to the test for TB diagnosis. The input patch size and the model size are denoted with a notation: for example, ViT-L32 denotes the "Large" variation with a 32×32 input patch size, and ViT-B32 denotes the "Base" variation with a 32 × 32 input patch size. The specifics of the ViT variant configurations based on the BERT model are presented in Table 2.

The suggested model consists of stacked transformer layers that transmit the flattened output from the encoder layer to batch normalization and dense layers constituting the MLP block. There are three layers in the encoder. The MSA layer amalgamates all attention outputs to the relevant linear dimensions. The second layer is the MLP with two dense layers and activation function. There is the normalization layer which assists the model in learning unique properties for improved performance. Because of its deterministic nonlinearity with a stochastic regularization effect, the GELU gets utilized in the first dense layer. In the final dense layer, a SoftMax activating function with L2 regularization gets employed to reduce over-fitting. A simplified block structure of the ViT model for TB identification with X-ray radiographs is illustrated in Fig. 2.

Experimental Details

Using TensorFlow with the Keras library, a 12 GB NVIDIA Tesla K80 GPU, Python programming, and online-based Google collaborative infrastructure, the model was developed. To reduce over-fitting during training, data augmentation techniques such as re-scaling, zooming, flipping, shearing, and rotating helped to provide variation to the dataset.

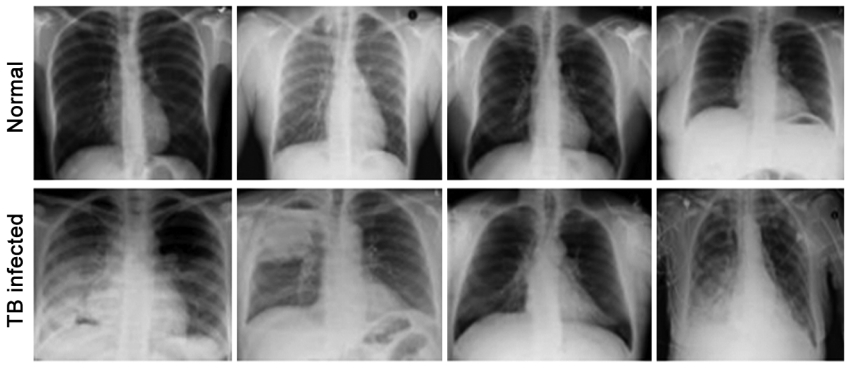

Fig. 1 — Sample of images used in the experimentation

The following hyperparameters got used to training the model: RectifiedAdam optimizer, a learning rate of 1×10^{-4}, a dropout rate of 0.2, and 30 epochs. The suggested transformer-based TB diagnosis system will be a valuable addition to the healthcare system to increase global health coverage. In Fig. 3 the workflow of the proposed framework is presented from a clinical perspective. The system will help radiologists to offer screening and diagnosis of the disease in areas where radiologists are in short supply due to the rapid spread of infections and inadequate screening facilities.

Evaluation Metrics

Conventional metrics like accuracy, sensitivity, specificity, Area Under the Curve (AUC) and Confidence Interval (CI), were used to evaluate the model efficacy. The AUC, which highlights the model's balance between excellent and subpar classifications, is computed using the ROC curve and is a widely used performance metric for medical classification problems. These metrics are as follows:

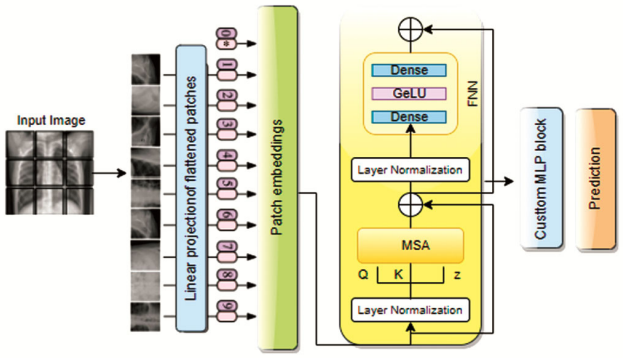

Fig. 2 — Conceptual framework of the proposed framework

$$Accuracy = \left(\frac{\sum Number\ of\ predictions}{\sum Total\ number\ of\ predictions}\right) \quad \ldots (2)$$

$$Sensitivity = \left(\frac{\sum True\ positives\ predictions}{\sum True\ Positives\ and\ False\ Negative}\right) \quad \ldots (3)$$

$$Specificity = \left(\frac{\sum Correct\ prediction}{\sum Total\ input\ samples}\right) \quad \ldots (4)$$

$$Precision = \frac{TP}{TP+FP} \quad \ldots (5)$$

$$F-1 = \frac{2 \text{ x } SEN \text{ x } PREC}{SEN + PREC} \quad \ldots (6)$$

$$Error \pm const * \sqrt{((\text{error} * (1 - \text{error}))/\text{n})} \quad \ldots (7)$$

where, $Error = \frac{incorrect\ prediction}{total\ prediction}$

$n = total\ number\ of\ observations$

Accuracy is a standard measure of accurately classified (True Positive and True Negative) samples divisible by total data samples, as shown in Eq. 2. To calculate sensitivity or recall, divide the number of truly detected positive occurrences by the total number of all actual positive (True Positive and False Negative) cases, as shown in Eq. 3. Specificity determines the model's ability to identify non-infected occurrences.

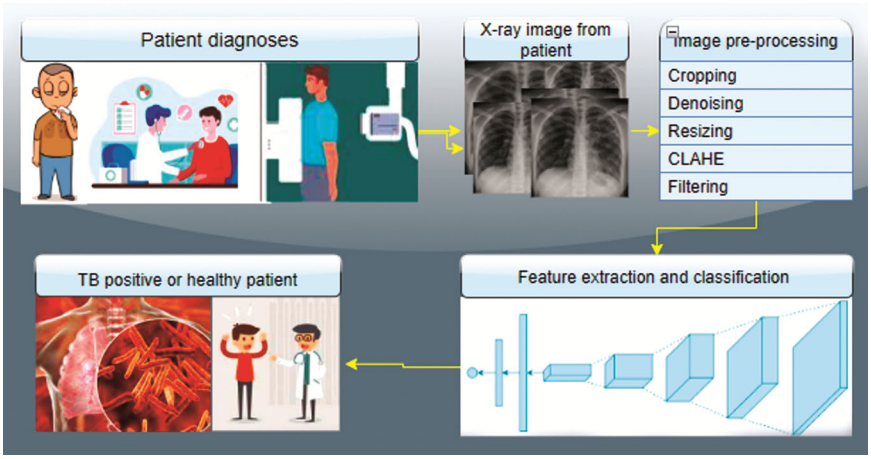

Fig. 3 — Workflow of the proposed automatic TB diagnosis system from a clinical perspective

Table 2 — Specifics of the various Vision Transformer models

Model	Image size	Layers	size	Perceptron size	Heads	Parameters
ViT-Base	224 × 224	12	768	3072	12	86 M
ViT-Large	224 × 224	32	1024	4096	16	307 M

Table 3 — Experimental outcomes from the ViT models

Model	Acc (%)	Sen (%)	Spec (%)	Prec (%)	F-1 Score (%)	AUC	CI
ViT-B16	96.00	95.98	96.02	95.63	96.04	0.964	0.04 ± 0.0080
ViT-B32	96.96	96.89	97.01	96.72	96.80	0.972	0.03 ± 0.00702
ViT-L32	94.91	94.61	95.19	85.82	90.00	0.941	0.05 ± 0.00898

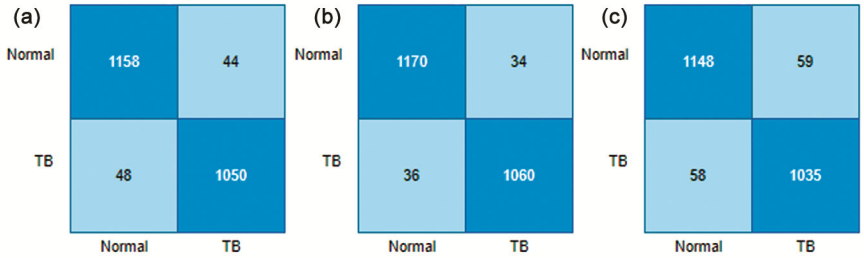

Fig. 4 — Confusion matrix for normal and tuberculosis (TB) classification: (a) ViT-B16 model, (b) ViT-B32 model, (c) ViT-L32 model

It is computed by dividing the total number of genuine negative (TN and FP) instances by the number of cleanly classified negative cases (TN) as shown in Eq. 4. Precision is used as one of the standard measures. Its calculation is in Eq. 5. The F-measure was used to compare the sensitivity and precision scores. When the sensitivity equals the precision, then the F-score is maximized. Its calculation is in Eq. 6. The confidence Interval is another standard metric utilized in this study for performance assessment. Its computation is in Eq. (7).

Results and Discussion

This work examines the potential of pretrained and optimized ViT models for automated TB identification from X-ray radiography. ViT-B32 had the best performance, with an accuracy 96.96% and an AUC of 0.972. The accuracy score of the ViT-B16 was 96.00%, and the AUC was 0.964. The ViT-L32 model, on the other hand, had the lowest accuracy and AUC ratings of 94.91 and 0.941. The Confidence interval metric measured the degree of uncertainty or certainty in the model's predictions, where a constant value of 1.6 (95%) was used to compute the CI, and the results are in Table 3.

The smaller model (ViT-B16, ViT-B32) worked well because of a condensed parameter size, according to the intuition drawn from the findings. The larger model (ViT-L32) performed the worst due to its enormous parameter size, which had minimal effect on the model's overall effectiveness. The confusion matrix in Fig. 4 confirms this assertion. The outcomes of all models based on the specified standard assessment measures employed for the study are given in Table 3.

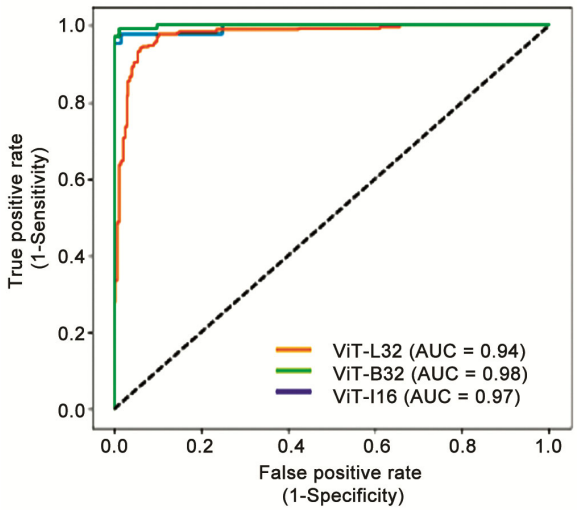

Fig. 5 — ROC Curves for TB and non-TB classification

In Fig. 4 the confusion matrix for the models is presented. The ViT-B32 model misclassified 34 of 1094 TB positive images, and 36 non-TB out of 1206 got misclassified, making it the best model among the three.

The huge parameter size of the vision transformer model is one of its limitations since it makes training it computationally costly. Moreover, it has trouble identifying local patterns essential for achieving the best results in image classification tasks. This had an impact on the models' performance. The ROC curves for the models are in Fig. 5.

Visualization

The proposed work intends to provide a system that can aid radiologists in the automated screening of TB

using X-ray radiographs. Deep neural networks are black boxes from the radiologist's point of view, as there is no way to determine which portion of the input picture to the model was responsible for the judgment or what the model learnt from the input. There is no indication when this model fails and misclassifies the input image. This study presents interpretations from the prediction by creating heat maps to display the portion of the input X-ray most predictive of the TB manifestation as shown in Fig. 6.

The attention map approach got employed to visualize the predictions and misclassifications made by the best model (ViT-32), which gives comprehension based on visual diagnosis for radiologists to infer. The resulting heat map indicates the area of the picture that is most likely to include TB markers for categorization purposes. The model's self-attention score is used to visualize the input image, with red regions contributing the most.

Performance Comparison

The results of the ViT-B32 transformer model got compared to that of other CNN models like EfficientNet-B5, ReNet50, DenseNet-121, and MobileNet. The same dataset used in training the ViT models got employed to pretrain the CNN models to evaluate their performance. The results are in Table 4.

The confusion Metrix as in Fig. 7 obtained from the various CNN models employed for TB detection has compared their results to the transformer network. It is evident from Table 3 that the ViT-B32 model

Table 4 — Comparison of pretrained CNN models to the ViT-B32 model for TB classification

Model	Acc (%)	Sen (%)	Spec (%)	Prec (%)	F-1 Score (%)	AUC
EfficientNet-B5	94.87	94.52	95.27	94.78	94.65	0.95
ReNet50	94.34	93.69	94.94	94.38	94.03	0.95
DenseNet-121	94.00	93.24	94.69	94.01	93.62	0.94
MobileNet	94.69	94.14	95.19	85.41	89.56	0.95
ViT-B32	96.96	96.89	97.01	96.72	96.80	0.97

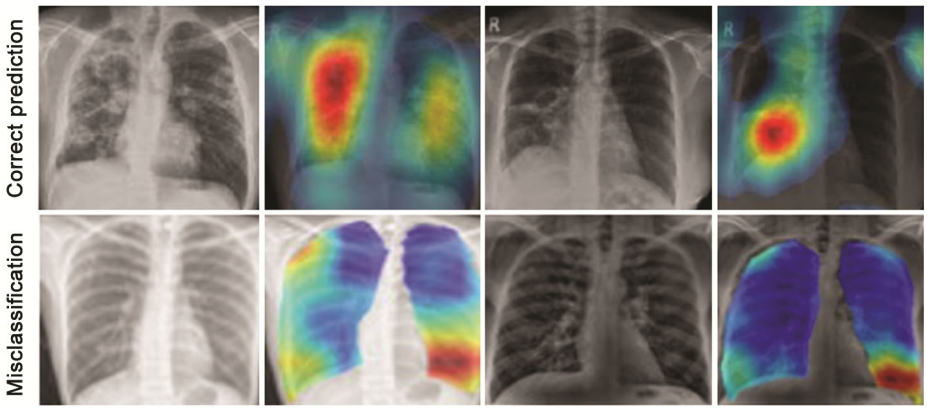

Fig. 6 — Visualization of TB prediction made by ViT-B32 model using images and associated heatmaps: correctly predicted (top), misclassified (bottom)

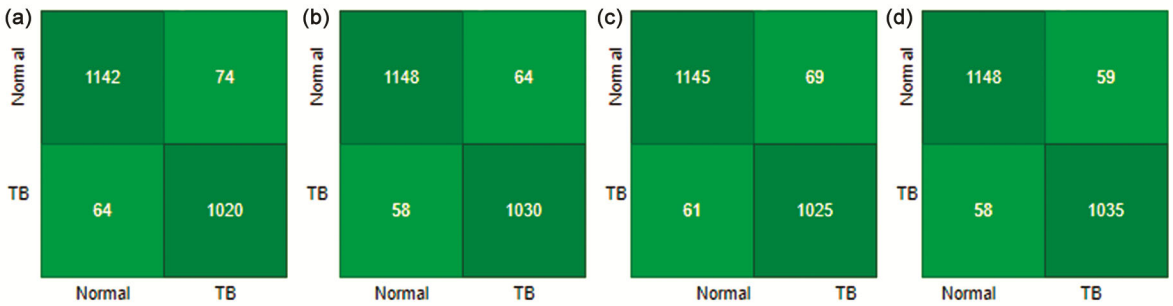

Fig. 7 — Confusion matrix of some outstanding CNN models pretrained for TB detection: (a) DenseNet-121, (b) efficientNet-B5, (c) ReNet50, (d) MobileNet

outperformed all the pretrained CNN models in all aspects of the evaluation criteria.

The ViT model's performance in TB detection got evaluated in this study. As a result, the ViT-B32 model offered the best quantitative and statistical measurements in classifying the images. The capacity of the model to receive global input from the early layers and the deep self-attention mechanism, which enables patch representation analysis for decision-making, are the main factors contributing to the improved performance. For TB detection, the ViT-B32 model got tested against CNN-based models. The ViT-B32 beat CNN-based transfer learning in all assessment measures examined for the study, as depicted in Table 3. These findings imply that the ViT-B32 model outperforms existing designs in TB classification challenges. The attention map depicts a highly accurate visualization of the predictions, making it appropriate for clinical adaptation.

Conclusions

This research provided a DL-based framework for early identification of TB illness. This research compares and analyzes different cutting-edge CNN and transformer models to see which works better using established measures that assess performance visually and statistically. The study focused on diagnosis of TB based on X-ray radiographs with three variants of ViT original model, (ViT-B16, ViT-B32, and ViT-L32). The results indicate that ViT-B32 surpasses current deep CNN networks for TB detection. The ViT model's attention mechanism makes it effective at detecting the most impacted regions in the images. The suggested approach contributes to DL solutions offered for the early detection of the disease. It is also cost-effective, making it economical and accessible to patients, especially in areas with fewer radiologists. The cutting-edge detection performance of this model might be employed as a rapid and efficient diagnostic tool, reducing TB casualties caused by wrong or delayed screening. Future research will investigate the application of CNN and transformer models to identify TB utilizing chest X-ray modalities.

References

1 Annabel B, Boon S den, Dean A, Dias H M, Dennis F & Floyd K, *Global Tuberculosis Report 2022*, Geneva 2022, doi:cc bY-Nc-sa 3.0 iGo.

2 Rajaraman S, Zamzmi G, Folio L R & Antani S, Detecting tuberculosis-consistent findings in lateral chest x-rays using an ensemble of CNNs and vision transformers, *Front Genet*, **13** (2022) 1–13, doi:10.3389/fgene.2022.864724.

3 Das D, Santosh K C & Pal U, Inception-based deep learning architecture for tuberculosis screening using chest x-rays, in *Int Conf Pattern Recognit (ICPR)*, 2021, 3612–3619, doi:10.1109/icpr48806.2021.9412748.

4 Ghaderzadeh M & Asadi F, Deep learning in the detection and diagnosis of Covid-19 using radiology modalities: A systematic review, *J Healthc Eng*, (2021) 6677314, doi:10.1155/2021/6677314.

5 Iqbal A, Latief J & Mudasir M, CoroNet: A deep neural network for detection and diagnosis of COVID-19 from chest x-ray images, *Comput Methods Programs Biomed*, **196** (2020), doi:10.1016/j.cmpb.2020.105581.

6 Verma D, Bose C, Tufchi N, Pant K, Tripathi V & Thapliyal A, An efficient framework for identification of tuberculosis and pneumonia in chest x-ray images using neural network, *Proc Comput Sci*, Vol 171 (Elsevier B.V.) (2020) 217–224, doi:10.1016/j.procs.2020.04.023.

7 Singh J, Tripathy A, Garg P & Kumar A, Lung tuberculosis detection using anti-aliased convolutional networks, *Proc Comp Sci*, Vol 173, (Elsevier B.V) (2020) 281–290, doi:10.1016/j.procs.2020.06.033.

8 Verenich E, Martin T, Velasquez A, Khan N & Hussain F, Pulmonary disease classification using globally correlated maximum likelihood: an auxiliary attention mechanism for convolutional neural networks, (2021) 1–13, http://arxiv.org/abs/2109.00573.

9 Meng J, Tan Z, Yu Y, Wang P & Liu S, TL-med: A two-stage transfer learning recognition model for medical images of COVID-19, *Biocybern Biomed Eng*, 42 (2022) 842–855, doi:10.1016/j.bbe.2022.04.005.

10 Kotei E & Thirunavukarasu R. Ensemble technique coupled with deep transfer learning framework for automatic detection of tuberculosis from chest X-ray radiographs, *Healthcare*, 10 (2022) 2335, doi:10.3390/healthcare 10112335

11 Shi L, Liu W, Zhang H, Xie Y & Wang D, A survey of GPU-based medical image computing techniques Lin, *Quant Imaging Med Surg*, **2(3)** (2012) 188–206, doi:10.3978/j.issn.2223-4292.2012.08.02.

12 Jos Escorcia-Gutierrez, Gamarra M, Soto-Diaz R, Alsafari S, Yafoz A & F Mansour R, Optimal synergic deep learning for COVID-19 classification using chest X-Ray images, *Comput Mater Contin*, **75(3)** (2023) 5255–5270. doi:10.32604/ cmc.2023.033731.

13 Mukherjee P, Roy CK & Roy SK. OCFormer: One-class transformer network for image classification. *arXiv:220411449v1*. 2022. http://arxiv.org/abs/2204.11449.

14 Chen J, Frey EC, He Y, Segars WP, Li Y & Du Y, TransMorph: Transformer for unsupervised medical image registration, *Med Image Anal*, 2022, 82, doi:10.1016/ j.media.2022.102615.

15 Meedeniya D, Kumarasinghe H, Kolonne S & Fernando C, Chest X-ray analysis empowered with deep learning : A systematic review, *Appl Soft Comput*, 126 (2022) 109319, doi:10.1016/j.asoc.2022.109319.

16 Lin A, Chen B, Xu J, Zhang Z, Lu G & Zhang D, DS-TransUNet: Dual swin transformer U-Net for medical image segmentation, *IEEE Trans Instrum Meas*, **71(8)** (2022) 1–13, doi:10.1109/TIM.2022.3178991.

17 Desai M & Shah M, An anatomization on breast cancer detection and diagnosis employing multi-layer perceptron neural network (MLP) and Convolutional neural network (CNN), *Clin eHealth*, **4** (2021) 1–11, doi:10.1016/ j.ceh.2020.11.002.

18 Duong L T, Le N H, Tran T B, Ngo V M & Nguyen P T, Detection of tuberculosis from chest X-ray images: Boosting the performance with vision transformer and transfer learning, *Expert Syst Appl* (2021) 184 (July), doi:10.1016/j.eswa.2021.115519.

19 Puttagunta M K & Ravi S, Detection of tuberculosis based on deep learning based methods, *J Phys Conf Ser*, 1767(1) (2021) doi:10.1088/1742-6596/1767/1/012004.

20 Ayaz M, Shaukat F & Raja G. Ensemble learning based automatic detection of tuberculosis in chest X-ray images using hybrid feature descriptors, *Phys Eng Sci Med*, **44(1)** (2021) 183–194, doi:10.1007/s13246-020-00966-0.

21 Akbar S, GhaniHaider N, Tariq H, Tuberculosis diagnosis using x-ray images. *Int J Adv Res*, 7(4) (2019) 689–696, doi:10.21474/ijar01/8872.

22 Jaeger S, Candemir S, Antani S, Wáng Y-X J, Lu P-X & Thoma G, *Two Public Chest X-Ray Datasets for Computer-Aided Screening of Pulmonary Diseases*, Vol 4 (2014) doi:10.3978/j.issn.2223-4292.2014.11.20.

23 Hooda R, Sofat S, Kaur S, Mittal A & Meriaudeau F, Deep-learning: A potential method for tuberculosis detection using chest radiography, *Proc 2017 IEEE Int Conf Signal Image Process, Appl ICSIPA 2017*, (2017) 497–502, doi:10.1109/ICSIPA.2017.8120663.

24 Guo R, Passi K & Jain C K, Tuberculosis diagnostics and localization in chest x-rays via deep learning models, *Front Artif Intell*, **3** (2020), doi:10.3389/ frai.2020.583427.

25 Ul Abideen Z, Ghafoor M, Munir K, Saqib M, Ullah A, Zia T, Tariq S A, Ahmed G & Zahra A, Uncertainty assisted robust tuberculosis identification with bayesian convolutional neural networks, *IEEE Access*, **8** (2020) 22812–22825, doi:10.1109/ACCESS.2020. 2970023.

26 Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J & Houlsby N, An image is worth 16x16 words: Transformers for image recognition at scale, http://arxiv.org/abs/2010.11929.

27 Xie E, Wang W, Yu Z, Anandkumar A, Alvarez J M & Luo P, SegFormer: Simple and efficient design for semantic segmentation with transformers, (2021) 12077–12090, *arXiv:210515203v3*.

28 Azizi S, Mustafa B, Ryan F, Beaver Z, Freyberg J, Deaton J, Loh A, Karthikesalingam A, Kornblith S, Chen T, Natarajan V & Norouzi M, Big self-supervised models advance medical image classification, in *ICCV*, ; 2021, 3478–3488, http://arxiv.org/abs/2101.05224.

29 Hou B, Kaissis G, Summers R M & Kainz B, RATCHET: Medical Transformer for Chest X-ray Diagnosis and Reporting, *Medical Image Comput Computer Assist Interven* – MICCAI 2021, 24th Int Conf (Strasbourg, France), September 27 – October 1, 2021, Proceedings Part VII, 293 – 303, doi:10.1007/978-3-030-87234-2_28

30 Touvron H, Cord M, Douze M, Massa F, Sablayrolles A & Jégou H, Training data-efficient image transformers & distillation through attention, (2021) 1–22, http://arxiv.org/abs/2012.12877.

31 Yuan L, Chen Y, Wang T, Yu W, Shi Y, Jiang Z *et al*., Tokens-to-Token ViT: Training vision transformers from scratch on ImageNet, *Proc IEEE Int Conf Comput Vis*, 2021, 538–547, doi:10.1109/ICCV48922.2021.00060.

32 Chen C F, Fan Q & Panda R, CrossViT: Cross-attention multi-scale vision transformer for image classification, (2021) 347–356, doi:10.1109/ICCV48922.2021.00041.

33 Li Y, Zhang K, Cao J, Timofte R & Van Gool L, LocalViT: Bringing locality to vision transformers, 2021, http://arxiv.org/abs/2104.05707.

34 Wang W, Xie E, Li X, Fan D, Song K, Liang D, *et al*., Pyramid vision transformer: A versatile backbone for dense prediction without convolutions, *Proc IEEE Int Conf Comput Vis*, 2021, 548–558. doi:10.1109/ICCV4 8922.2021.00061.

35 Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, Lin S & Guo B, Swin transformer: Hierarchical vision transformer using shifted windows, *Proc IEEE Int Conf Comput Vis*, 2021, 9992–10002, doi:10.1109/ICCV48922.2021.00986.

36 Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, Kaiser L, Polosukhin I, Attention is all you need, *Adv Neural Inf Process Syst*, (2017) 5999–6009.

37 Lambert Z, Petitjean C, Dubray B M, Kuan S. SegTHOR: Segmentation of THoracic Organs at Risk in CT images, *Proc 10th Int Conf Image Proces Theory Tools Appl (IPTA)* (Paris, France) 2020, 1–6, doi:10.1109/IPTA50 016.2020.9286453.