IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 15 May 2024, accepted 9 July 2024, date of publication 15 July 2024, date of current version 23 July 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3428446

RESEARCH ARTICLE

Tuberculosis Detection From Chest X-Ray Image Modalities Based on Transformer and Convolutional Neural Network

EVANS KOTEI[1] AND RAMKUMAR THIRUNAVUKARASU[2]

[1]Department of Computer Science, Faculty of Applied Science and Technology, Kumasi Technical University, Kumasi, Ghana
[2]School of Computer Science Engineering and Information Systems, Vellore Institute of Technology, Vellore 632014, India

Corresponding author: Ramkumar Thirunavukarasu (ramkumar.thirunavukarasu@vit.ac.in)

This work was supported by Vellore Institute of Technology, Vellore, India, [Article Processing Charge (APC)].

ABSTRACT Tuberculosis (TB) is an airborne disease with a high fatality rate that often affects people's lungs. Early detection of the disease can guarantee a cure, but that is not the case due to the lack of experts and screening facilities. With the advancement in Artificial Intelligence (AI) techniques, publicly available clinical datasets and high computer systems, several solutions have been proposed to automate the diagnosing process based on deep learning (DL) algorithms. This study presents a DL technique based on a combined Data efficient image transformer and the Residual Network-16 (ResNet-16) model for effective TB diagnosis from X-ray images. The TBX11K dataset got utilized for this investigation. The dataset has three categories—Healthy, Sick but non-TB and TB. Relying on only healthy X-rays as the negative category in clinical settings where there are many unwell but non-TB samples can cause considerable false positives in the model prediction, hence the adoption of the sick but non-TB class. The dataset got divided into training, validation, and testing sets for the experiment. The self-attention mechanism within the transformer part of the proposed model learns crucial information and constructs the relationship between image tokens. The ResNet-16 part uses depth-wise convolution to gather local representations and reduce computing costs while increasing the diagnosing accuracy. The global average pooling is applied to the feature maps at the final convolution layer to generate heatmaps based on the class activation map to show exactly where the prediction happened. The suggested model obtained TB diagnostic accuracy, sensitivity, specificity and precision rates of 99.38%, 99.49%, 99.26% and 99.24%, respectively. Additionally, it is lightweight (6.9M) and detects more quickly (4.79ms) than the other comparative cutting-edge versions. The suggested model's exceptional results demonstrate its effectiveness in TB detection, allowing for practical use.

INDEX TERMS Convolutional neural network (CNN), deep learning, vision transformer network, tuberculosis, X-ray.

I. INTRODUCTION

TB is a transmittable illness triggered by Mycobacterium that affects human lungs but may also distress other regions of the human anatomy. There are ways to avoid and typically recover from tuberculosis (TB). Nonetheless, TB accounted for nearly twice as many fatalities as HIV/AIDS in 2022 and was the second most common infectious agent worldwide cause of death after coronavirus disease (COVID-19). TB continues to claim the lives of almost 10 million people annually. In 2022, there were 7.5 million newly diagnosed cases of tuberculosis worldwide. With 7.1 million in 2019, above the pre-COVID baseline (and prior historical peak) and up from 5.8 million in 2020 and 6.4 million in 2021, representing the highest figure since the World Health Organization (WHO) started global TB monitoring in 1995. Eight countries accounted for 87% of the world's TB cases. The countries include India (27%), Indonesia (10%), China (7.1%), the Philippines (7.0%), Pakistan (5.7%), Nigeria (4.5%), Bangladesh (3.6%) and the Democratic Republic

The associate editor coordinating the review of this manuscript and approving it for publication was Mohammad Shorif Uddin.

© 2024 The Authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License.
For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

of the Congo (3.0%) [1]. Early detection can guarantee treatment, but this is not the case owing to a shortage of professionals and screening facilities. Due to a global lack of qualified radiologists, there is a rising demand for artificial intelligence-based medical image processing systems for successful healthcare delivery. Various artificial intelligence-based systems for medical image analysis got developed over the years for modern healthcare delivery [2], [3], [4]. Even the World Health Organization (WHO) criteria for illness screening have been amended in response to developments in artificial intelligence, allowing Computer Aided Diagnosis (CAD) systems to complement human interpreters for medical imagery [1].

Deep learning approaches are mainly for performing various tasks [5], [6]. For instance, a CNN model was suggested in [7] and showed remarkable performance in recognizing and detecting large-scale fish classes. Artificial intelligence in clinical dispensation has evolved into a critical component of modern healthcare delivery, such as disease identification from medical images [8]. The implementation of Computer-Assisted Diagnosis Systems [9], [10], [11], [12] improve physicians' and radiologists' decision-making to provide appropriate health care to patients.

CNN is a popular DL technique for medical image analysis due to its capacity to extract spatial representations from images, removing the drawback of hand-crafted feature engineering associated with standard machine learning algorithms [13], [14], [15]. Because of freely obtainable large-scale datasets [16], [17], [18] with annotations and the obtainability of high-speed computer resources like graphical processing units (GPUs), studies on identifying lung diseases in X-rays based on DL approaches have shown higher performance [19].

In [15], a deep CNN model with an SVM classifier is used to diagnose pneumoconiosis using data from the National Institute for Occupational Safety and Health (NIOSH) [20]. The experimental findings indicated that the suggested framework outperformed previous preceding models. SVM performed admirably, with an accuracy of 92.68%. MobileNet and the Artificial Ecosystem-based Optimization (AEO) algorithm were employed to extract features from X-ray images for tuberculosis diagnosis [21]. The approach increased the model's classification performance, resulting in 90.2% and 94.1% accuracy rates.

Disease diagnosis using medical imaging is a vital tool for optimal healthcare delivery. Varied image acquisition techniques and resolutions sometimes influence the procedure. The authors of [22] suggested a model for tuberculosis detection that uses three image-enhancing approaches to increase image quality. The methods used were Unsharp Masking (UM), High-Frequency Emphasis Filtering (HEF), and Contrast Limited Adaptive Histogram Equalization (CLAHE). This increased the model's classification performance by 89.92% accuracy and 94.8% area under the curve.

In [23], a CNN model based on X-ray modality got built for TB diagnosis. The feature extraction and classification were based on a semi-supervised localization technique with 27 layers and 12 residual connections. The final layer included a lesion-based and image-based classifier. The model's rating on accuracy was 94.73%. Based on datasets from Shenzhen (SZ) and Montgomery County (MC) [24], a model for tuberculosis diagnosis described in [25] got trained and evaluated using the CNN method, achieving superior performance than standard machine learning techniques presented for comparable tasks. Its accuracy was 90% and 80%, respectively. The authors of [26] proposed three-way tuberculosis diagnosing model using X-ray radiography. Step 1 required modifying the structural features of the CNN model, while steps 2 and 3 optimized the model using an artificial bee colony approach. The algorithm detected seven symptoms linked with TB after being trained on the Shenzhen dataset. It achieved an accuracy, sensitivity and AUC scores of 98.46%, 98.76% and 99, respectively. A Bayesian-based CNN got proposed in [27] to overcome the SoftMax inference problem. The model improved accuracy for both datasets MC and SZ to 96.42% and 86.46%, respectively.

In [28], the authors developed a system for diagnosing tuberculosis from X-ray images that included pretrained models and handcrafted feature engineering techniques. The Montgomery County and Shenzhen datasets were used to train and validate the model using a k-fold cross-validation approach. It has an Area Under Curve score of 0.99 and 0.97.

Despite the overwhelming effectiveness of CNN models in medical image processing, the restricted number of datasets available for training causes overfitting and poor generalization. Transfer learning is one method researchers employ to deal with the scarcity of medical images to train DL models whose success partly depends on larger datasets. Transfer learning is a technique that involves transferring information learned in a pretrained model (source domain) to another domain (target domain) [29].

Pretrained CNN networks that include Residual Network (ResNet), ChexNet, InceptionV3 and SqueezeNet extracted the vital features from the dataset for classification. The proposed framework increased the model's ability to diagnose with higher standards.

Deep learning methods developed for medical image classification occasionally show variations in prediction due to the unpredictable nature of medical images produced by multiple machines with varied settings and resolutions. The variance has an impact on the model's performance. Eliminating the variance in DL models, some researchers employed methods like ensemble learning, where predictions from different models get merged for final classification to improve the models' performance [30]. Based on the Shenzhen and Montgomery datasets, a stacked ensemble model with feature descriptors and pre-trained CNN models was suggested for TB detection [31]. The results of three pre-trained networks, VGGNet, Resnet, and GoogleNet, were integrated for further

categorization by a support vector machine (SVM) classifier. The ensemble classifier attained accuracy scores of 82.6% on the Montgomery dataset and 84.7% on the Shenzhen dataset.

Recently, Vision Transformers (ViTs) [32] with in-built self-attention phenomenon have been an alternative to CNNs in computer vision tasks. Some investigations employed vision transformers for image segmentation, classification and detection [33], [34], [35]. The authors in [33] proposed a transformer-based deep neural network for Corona Virus (COVID) diagnosis through X-ray modalities. The framework realized outstanding results with a classification accuracy of 99.7%, while a similar model proposed in [36] achieved 98% accuracy along with an Area Under the Curve (AUC) score of 99%.

Transformer networks rely on a larger dataset to attain optimal performance. As a result, the authors in [37] merged several publicly accessible datasets for the experiment on TB detection. The feature extraction and classification were based on the ViT network. The model attained an accuracy of 97.72% and an AUC of 100%. Despite the ViT network's exceptional performance, optimal outcomes require a large dataset. The authors in [38] defused this notion by proposing the Data Efficient Image Transformer (DeiT) model that achieved optimal performance without requiring a large training dataset, i.e., with Imagenet1k only. The transformer network is amassing success by following on the same tangent as CNNs by going through a series of transformations for maximum performance.

The transformer model has produced outstanding results in sequential tasks [39] and vision problems.

In vision tasks, the ViT [40] model has produced excellent performance in clinical image processing tasks. Despite this, it is most often affected by scarce data accessibility, which prevents it from learning innate biases such as translation and equivariance. As a result, the capacity to examine the relationship between nearby image pixels get affected. Furthermore, the execution complexity of ViT models grows with increased resolution, necessitating more processing power.

Even though CNN models, on the other hand, have demonstrated promising results even with low data based on their intrinsic inductive bias properties, they still have some drawbacks. CNNs fail to translate the relative location of multiple image characteristics, necessitating huge receptive fields to encode the combination of these representations and capture far-off relationships. Another drawback is large convolutional kernels that require higher computational resources for processing [35].

Based on the limitations identified within these image classification benchmark models, we propose a CNN-Transformer-based model to utilize the intrinsic advantages of these two models to increase detection performance and reduce model complexity. The proposed model employs the DeiT transformer and the ResNet-16 models for TB diagnosis based on X-ray image modalities. The transformer part of the framework extracts the global information, while the CNN part extracts the local patterns, which boosts the detection performance. The DeiT model is pretrained on a smaller dataset (ImageNet), unlike the conventional ViT model trained on a much larger dataset. The CNN part further reduces the model's complexity through the depth-wise convolution operation, making the model lighter and faster for diagnosis.

Below are the study's unique contributions:

- The proposed framework can extract global and local representations from the input set, but stand-alone CNN and Transformer models do not have both capabilities. Having both global and local representation features enhance the model's diagnosing efficacy.
- Compared to the traditional ViT model pretrained on a larger dataset, the model's minimal complexity yields excellent results. Depth-wise convolution in the proposed framework further reduces complexity without compromising model performance.
- The model is lightweight because of its few parameter sizes, making it lighter and faster in prediction than other cutting-edge models, ensuring more accurate diagnosis per second.

The remainder of the study flows as follows. Section II goes through the specifics of the suggested technique and the datasets employed for the experimentation. Section III is the Results and analysis. Section IV discusses the study's results, and Section V concludes it.

II. METHOD

The study describes a combined Transformer-CNN deep learning model for detecting TB from X-ray radiographs. The proposed model is a combined DeiT, which is a variant of the original vision transformer and the ResNet-16 CNN model. ViT's architecture depends on the basic Transformer [33] with self-attention phenomena, which has sparked much interest because of its outstanding result in natural language processing tasks. The CNN part of the model extracts local information, and the DeiT part extracts global relations, which provides enhanced classification performance.

A. PROPOSED MODEL

Given an input set of X-ray images, $I = \{a_i, b_i\}_{i=1}^{q}$, where a_i are images in the dataset and b_i are the respective class labels, where $b_i \in \{1, 2, \ldots c\}$ and c represents the number of classes. The suggested technique aims to learn a mapping between the input chest image and the right class label. The input image α with size $h \times w \times c$ (where h, w, and c are the height, width, and number of channels) goes through a linear embedding layer where it is converted into a sequence of length n by splitting the image into tiny blocks b as shown in "(1)":

$$\alpha = \left(\alpha_b^1 \alpha_b^2 \cdots \alpha_b^n\right) \tag{1}$$

where $n = h \times w / b^2$.

The series of blocks or patches is linearly projected onto a vector of model dimension M_{dim} using a learnt embedding

matrix E before being fed into the encoder.Position Embedding applies a linear transformation to each 2-dimensional (2D) categorization and converts it into a 1-dimensional (1D) feature vector. The position-aware embeddings go with a learnable class token X_{class} and a distillation token X_{distil} as in "(2)". The patch-embedded data, class and distillation tokens interact with one another via a self-attention method. Next, in order to complete the classification job, the embedded representations are concatenated with a learnable class token called X_{class}.

$$Z = \left(\alpha_{class}^{\cdot}; \alpha_{dist}^{\cdot}; \alpha_b^1 E; \alpha_d^2 E \ldots; \alpha_d^n E\right) + E_{pos}$$
$$E \in R^{(d \times d \times c)} \times_{dim}^{M} \text{ and } E_{pos} \in R^{(n+2)} \times_{dim}^{M} \quad (2)$$

The suggested framework has the DeiT transformer backbone, which can train on fewer datasets and deliver excellent results, unlike the classic ViT model. The encoder part of the model consists of L layers, each having a multi-head self-attention (MSA) and a feed-forward network (FFN). The MSA utilizes the self-attention (SA) mechanism to identify correlations between various input image blocks. The MSA uses multiple self-attention heads, for instance (SA_1, SA_2, SA_3,...SA_t), where t is the number of attention heads needed to capture n distinct relationships between the visual blocks or patches. The outputs from individual heads are then concatenated and projected to the final dimension by a linear layer. The MSA computation is in "(3)"

$$MSA = Concat(SA_1, SA_2, \ldots SA_t)W^n, \quad (3)$$

where $W^n \in R^{t.d_K \times M_{dim}}$

Because CNN can extract local characteristics from images, it is combined with the transformer model to learn local and global representations, which are necessary for efficient detection. The suggested framework is in Fig. 1.

The CNN section sorts the image characteristics into distinct semantic categories to get a small number of visual tokens as shown in "(4)". The clustered semantic representation allows the model to focus on vital portions while ignoring irrelevant background details.

$$T = SoftMax_{xy}(I \text{ x } W_A)^T I \quad (4)$$

$W_A \in R^{C \times L}$ is the semantic grouping from I,SoftMax(·) and I is the feature map. For any feature map I, a pointwise convolution is performed to map each pixel i_p, $\in R^c$ *to* L number of groups and spatial pooling is applied to get tokens. The spatial pooling technique generates weight tokens via the SoftMax activation function. The resultant weights get multiplied by the original feature, which produces new attention maps. The recurrent tokenizer weights [36] depend on the previous layer's visual tokens. T_{in} is employed to concatenate all the layers to lessen the sparseness of semantic information hence the computing cost gets reduced as shown in "(5)".

$$W_R = T_{in} W_{T \to R}$$
$$T = SoftMax_{xy}(IW_R)^T \quad (5)$$

where $W_{T \to R} \to R^{C \times CG}$

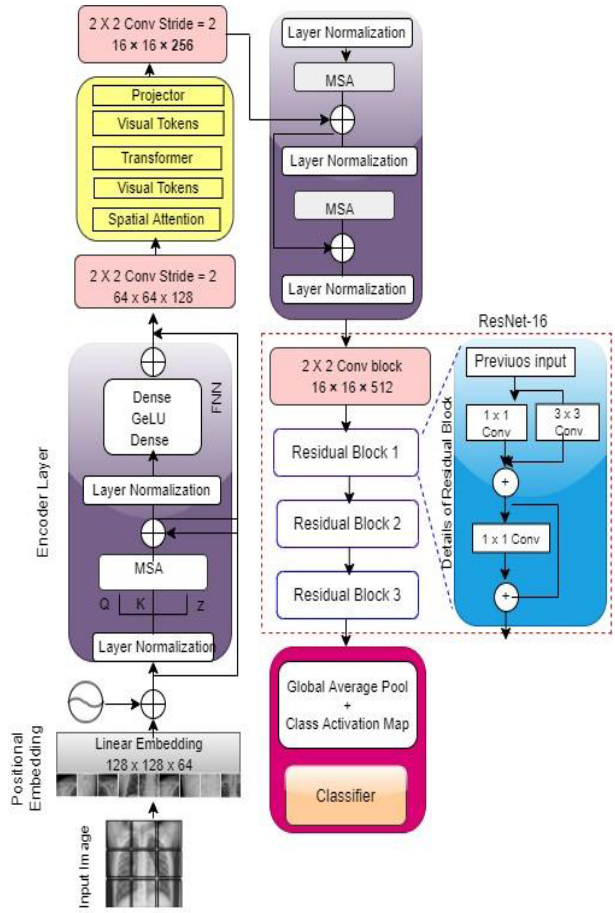

FIGURE 1. Proposed CNN-Transformer based framework.

The recurrent tokenizer begins with the second layer because it requires tokens from a previous layer and gradually refines it to the previously processed ideas. The proposed system leverages the transformer framework, which models interactions between these visual tokens after tokenization using input-dependent weights. With fewer tokens, the transformer architecture covers more conceivable notions. Below is a modified transformer

$$T'_{out} = T_{in} + SoftMax_L(T_{in}K)(T_{in}Q)^T T_{in} \quad (6)$$
$$T_{out} = T'_{out} + \sigma(T'_{out}F_1)F_2 \quad (7)$$

where T_{in}, T'_{out}, T_{out} $\in R^{L \text{ x } n}$ are the visual tokens calculated within the transformer as a key-query inner product: $(T_{in}\ K)(T_{in}\ Q)^T \in R^{L \times L}$ is K and Q in transformers, F_1, F_2 $\in R^{L \times n}$, which is two pointwise convolution and σ (·) is the Relu activation function.

An improved feature map is created by projecting these visual tokens into the pixel space as depicted in "(8)".

$$X_{out} = X_{in} SoftMax_L((X_{in}W_Q)(TW_K)^T)T, \quad (8)$$

X_{out}, X_{in} $\in R^{X \times Y \times C}$ denotes the feature map's output and input,X_{in} $W_Q \in R^{X \times Y \times C}$, indicates the Q value derived by the feature map, $(TW_K) \in R^{L \times C}$, denotes the K value

TABLE 1. Split of the TBX11K dataset.

Class	Train	Validation	Testing	Total
Normal images	960	120	120	1200
Sick but non-TB	960	120	120	1200
TB	960	120	120	1200
Total	2880	360	360	3600

computed out of the tokens. The Q and K weights are; $W_Q \in R^{C\times C}$ and W_K $R^{C\times C}$ and their product governs the transformation of visual tokens into the feature map. Focusing on valuable semantic data, VT may rebalance the input feature map according to the semantic relevance and offer the foundation for further categorization. The Multi-head Self-attention layer employs depth-wise with a kernel size, $k \times k$ and a stride S to decrease K and V. The approach uses a learnable bias B during the computation of attention. The expression is in "(9)".

$$\text{Attention}(Q, K, V) = \text{SoftMax}\left(\frac{QK^T}{\sqrt{d_{\dot{k}}}} + B\right)v \quad (9)$$

where bias $B \in R^{nx\frac{n}{k^2}}$

B. DATASET

The TBX11K dataset [37], which consists of three categories (Healthy, Sick but Not TB and TB), was employed for this investigation. There are 11200 X-ray images in the collection with corresponding bounding box annotations indicating the impacted areas. All images have a 512-pixel size. The class distributions are as follows: Healthy class = 5000 images, Sick but Not TB = 5000 and TB = 1200. Because of the imbalanced nature of the dataset, sampling from the larger classes was performed to balance the TB class. The sampled dataset was divided into training, validation, and testing sets, each with 2880, 360, and 360 X-ray images in the ratio of 80:10:10. The details of the split are in Table 1.

The majority of the TB datasets that are currently available include two categories, which are TB and non-TB, also referred to as healthy cases. In reality, abnormalities in chest X-rays like tuberculosis, atelectasis, cardiomegaly, effusion, infiltration, mass and nodule show aberrant patterns such as hazy and irregular lesions that differ from healthy X-rays. Therefore, in clinical circumstances where there are numerous ill but non-TB patients, utilizing just "healthy" X-rays as the "negative" category tends to lead to significant false positives in the model prediction, hence the introduction of the third category (sick but non-TB) [38]. The sample image from the dataset is in Fig. 2.

C. EXPERIMENTAL SETTINGS

The implementation was done in the cloud using Google collaboration infrastructure with 12GB NVIDIA Tesla K80 GPU, TensorFlow with the Keras library, and Python programming language. The dataset was split into three for

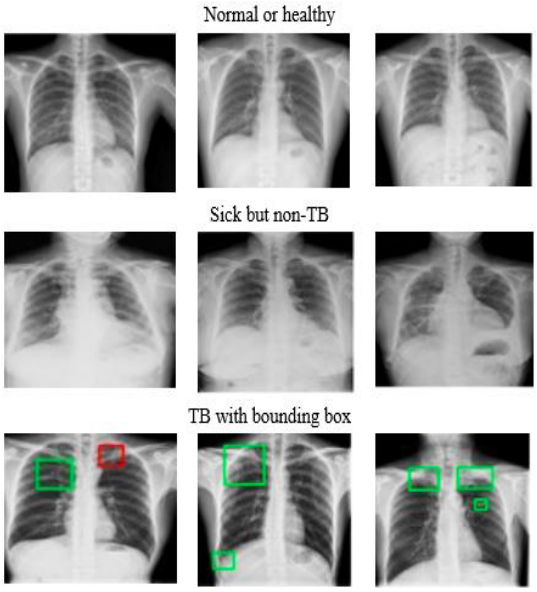

FIGURE 2. Sample images from the TBX11K dataset.

training, validation, and testing in the ratio of 80:10:10. The following data augmentation techniques got applied to the training dataset to prevent overfitting and improve classification performance, rescale (1/225), zooming (0.2), horizontal flipping (yes), shear (0.25), rotation (300) and width and height shift (0.2). The augmentation techniques employed in the study introduced diversity in the dataset, preventing the model from overfitting. For optimal performance, robustness and generalization of the model for chest X-ray analysis, all chest radiographs got screened and preprocessed to exclude low-quality or unreadable images.

Cropping and scaling were applied to the dataset to eliminate dark spots and markings, which provide little to no information and confuse the model during prediction. Training and hyperparameter tuning of the proposed model for optimal performance was achieved based on the following settings: Adam optimizer, the learning rate of 0.001, a dropout rate of 0.2 and 15 training epochs. The experiment employed three types of the DeiT transformer network as the transformer backbone: DeiT-Tiny (DeiT-T), DeiT-Small (DeiT-S) and DeiT-Base (DeiT-B). ResNet-16, on the other hand, is the CNN component of the proposed model. The suggested model augments the decision-making process of doctors and radiologists for early diagnosis, particularly in TB-endemic areas. The model's effectiveness is evaluated and compared to other cutting-edge automated TB detection methods.

D. PERFORMANCE EVALUATION

As indicated in "(10)", accuracy is a standard measure of properly categorized (True Positive and True Negative) samples divisible by total data samples. The number of correctly identified positive instances is divided by the total number of all actual positive (True Positive and False Negative) cases, as stated in "(11)", to measure sensitivity or recall.

Specificity determines the model's capacity to recognize non-infected instances (i.e., comparable to recall for positive cases). It is calculated by dividing the number of cleanly categorized negative cases (TN) by the total number of actual negative (TN and FP) cases, as described in "(12)". Additionally, we calculated the precision of the model by computing the percentage of genuine positive predictions among all positive predictions as shown in "(13)". The model was further assessed by calculating the F1-Score, which is the harmonic mean of Precision and Recall. In "(14)" is the computation.

$$Accuracy = \frac{(TP + TN)}{(TP + FN) + (FP + TN)} \tag{10}$$

$$Sensitivity = \frac{(TP)}{(TP + FN)} \tag{11}$$

$$Specificity = \frac{(TN)}{(TN + FP)} \tag{12}$$

$$Precision = \frac{(TP)}{(TP + FP)} \tag{13}$$

$$\text{F1 Score} = \frac{2 \times Sensitivity \times Precision}{Sensitivity + Precision} \tag{14}$$

III. RESULTS AND ANALYSIS

In this section, we present and analyze the results of the experiments to evaluate the proposed framework. The performance got assessed using the benchmark evaluation criteria utilized in this study. We examine the model's accuracy, sensitivity, specificity and diagnosing speed based on the employed dataset. The class activation map (CAM) [46] algorithm got utilized to visualize the predictions to aid doctors with further understanding in the final forecast. The visualization part clears the doubt of physicians on the notion that DL models are black boxes because the prediction process is not known. The attention maps provide information on which part of the image the model focused on making a prediction. Finally, we contrast the model's findings with other cutting-edge deep learning models proposed for TB diagnosis.

A. RESULTS ON CHEST X-RAY

The section displays the experimental findings from the suggested model as well as the other benchmark models used in the study for comparison. The experiment explored training and evaluating the suggested model by combining the ResNet-16 model with the DeiT variants: DeiT-Tiny (DeiT-T), DeiT-Small (DeiT-S), and DeiT-Base (DeiT-B), which formed the transformer network's backbone. To ascertain the robustness and efficacy of the proposed model, four other deep CNN models, namely, ResNet50, EfficientNet-B3, MobileNetV2 and Visual Geometry Group-16 (VGG-16), got trained on the same dataset for comparison. Table 2 contains different variants of the DeiT transformer model with varying parameters and heads considered for experimentation.

The parameters of DeiT-B for the experiment were computed based on Embedding Dimension (d) = 768 and Number of Heads (h) = 12. The formula is Dimension/head = 64. The other DeiT variations (DeiT-S and DeiT-Ti) had different head values, but the d remained the same. The number of heads and embedding dimension are the only variables that differ between models; we maintain a consistent dimension per head of 64. Table 3 shows the results of the evaluation of the models.

TABLE 2. Variants of the DeiT transformer neural network.

DeiT variant	Embedding size	Head	Layers	Param (M)	Image resolution
DeiT-T	192	3	12	5	224
DeiT-S	384	6	12	22	224
DeiT-B	768	12	12	86	224

TABLE 3. Results obtained by the proposed framework.

Dataset	Model	Param (M)	Acc (%)	Sen (%)	Spec (%)	Diagnosis time per image (ms)
TBX11K dataset	DeiT-T + ResNet-16	6.9	98.65	99.79	99.71	4.79
	DeiT-S + ResNet-16	23.2	98.61	98.0	98.12	6.9
	DeiT-B + ResNet-16	87.31	95.92	95.70	96.89	9.4

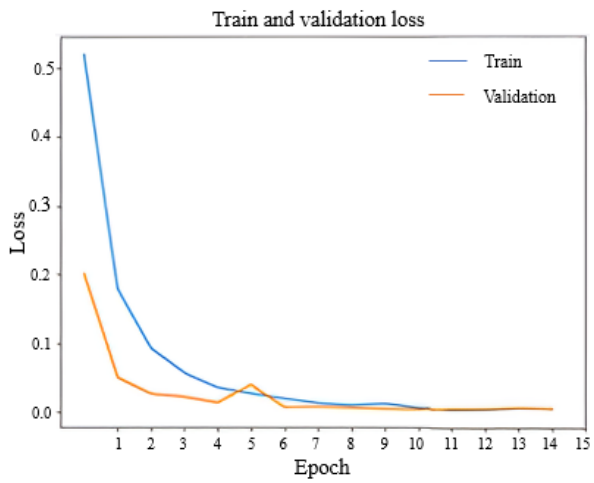

FIGURE 3. Visualization of training and validation loss versus epoch.

The experimental results show that the DeiT-T + ResNet-16 attained the best accuracy, sensitivity and specificity scores, beating the other two models. It also attained the best diagnosing time. The results including average time taken per image prediction is in Table 3. Based on these results, the DeiT-T + ResNet-16 proved superior over the other two models for TB diagnosis. The training and validation loss per epoch graph of the DeiT-T + ResNet-16 model is in Fig. 3.

A second experiment was carried out with the aforementioned cutting-edge CNN models for performance comparison. Due to the DeiT-T + ResNet-16's exceptional performance, its results got compared to other benchmark pretrained CNN models using the same assessment measures. The CNN models employed were ResNet50, EfficientNet-B3, MobileNetV2, and VGG-16. The experiment is conducted under the same conditions (Google Collab)

TABLE 4. Comparison of proposed models results to other four deep CNN models with same dataset.

Dataset	Model	Param (M)	Acc (%)	Sen (%)	Spec (%)	Diagnosis time per image (ms)
TBX11K dataset	MobileNetV2	6.9	96.93	98.99	99.17	4.81
	EfficientNet-B3	12	97.98	99.0	98.65	5.30
	VGG-16	138	94.79	95.91	96.23	12.32
	ResNet50	26	94.55	94.91	94.89	7.0
	DeiT-T+ ResNet-16	6.9	98.65	99.79	99.71	4.79

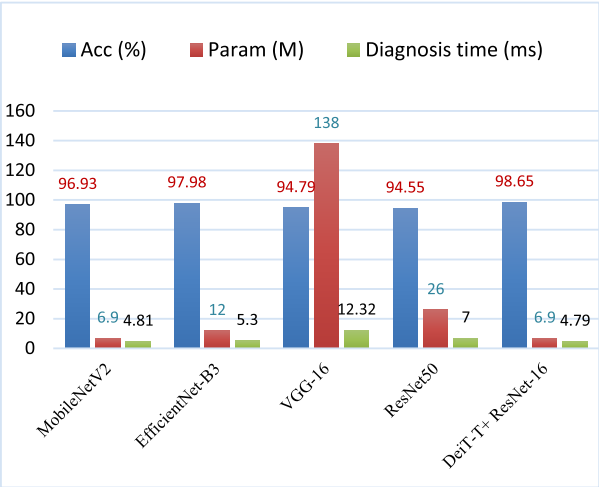

FIGURE 4. Performance comparison of the proposed model and benchmark CNN models for TB detection.

TABLE 5. Performance comparison of proposed model to other DL models for TB detection.

Reference	Model	Dataset	Metrics (%)
Akbar et al. [25]	CNN	MC and SZ datasets [24]	Acc = 90
Abideen et al., [27]	AlexNet, VGG16, VGG19, CNN	MC and SZ datasets [24]	Acc = 96.42
Hooda et al. [23]	CNN	MC and SZ datasets	Acc = 94.73
Guo et al. [26]	CNN	SZ [24]	Acc = 98.46, Sens = 98.76 AUC = 99
Ayaz et al., [28]	VGG-Net, MobileNet, ResNet50, Inceptionv3	MC and SZ datasets [24]	AUC = 99
Proposed model	DeiT-T + ResNet-16	TBX11K dataset	Acc = 98.65 Sen = 99.79 Spec = 99.71

TABLE 6. Results of the proposed model and four benchmark deep CNN models on new test dataset.

Dataset	Model	Acc (%)	Sen (%)	Spec (%)	Prec (%`)	F1-Score
TBX11K dataset	MobileNetV2	98.25	98.22	98.28	98.22	98.22
	EfficientNet-B3	98.63	98.73	98.52	98.48	98.11
	VGG-16	95.75	95.43	96.06	95.92	95.67
	ResNet50	95	94.42	95.57	95.38	94.90
	DeiT-T+ ResNet-16	99.38	99.49	99.26	99.24	99.36

with identical Graphical Processing Unit (GPU) and memory parameters for fairer model evaluation. The DeiT-T + ResNet-16 implementation's training datasets were employed to train the other four benchmark models for TB detection. The results indicate that the DeiT-T + ResNet-16 model performed better than the CNN models in this scenario. The assessment indicators include diagnosis accuracy, sensitivity, specificity, and model size (parameters). Table 4 displays the comparative experimental results. The outstanding performance shown by the model confirms its ability to diagnose the disease from the images.

The suggested model was tested against four deep CNN benchmark models for complexity and performance evaluation. The results revealed that our model had the best accuracy with fewer parameter sizes and the shortest diagnosis time. The model's complexity got measured based on the parameter size and the diagnosis time.

The model has a parameter size of 6.9M, which is 53.97% lesser to the MobileNetV2, 180.95% fewer to EfficientNet-B3, and 116.11% less to ResNet50. It also offers the best diagnosis time compared to all the bench mark models employed in the study for comparison as shown in Table 4.

In Fig. 4 is comparison of the proposed model's complexity based on the parameter size and diagnosing time against benchmark CNN models for TB diagnosis.

Table 5 compares the performance of the proposed model to other existing CNN models presented for automatic TB detection based on X-ray images. Again, the performance of the proposed model was outstanding, proving its efficacy in TB diagnosis.

An additional trial was conducted using an alternative dataset from a different population set to prove the efficacy and generalizability of the model. The trained DeiT-T + ResNet-16 model and the other four benchmark CNN models were tested using the Shenzhen dataset (SZ) and the Montgomery County (MC) chest X-ray dataset [24].The Montgomery dataset was collected by the Department of Health and Human Services, Montgomery County, Maryland, USA. The Montgomery dataset was collected by the Department of Health and Human Services, Montgomery County, Maryland, USA. The dataset consists of 138 frontal chest X-rays, of which 80 are Normal and 58 have TB manifestations. The Shenzhen dataset was collected by Shenzhen No.3 People's Hospital, Guangdong Medical College, Shenzhen, China. The dataset contains 662 frontal chest X-rays, of which 326 are un-infected, whereas 336 are TB-infected. The combined dataset consists of 800 images, of which 394 are TB-infested and 406 are Normal. The obtained results are presented in Table 6.

Interestingly, all the models tested on the new dataset gained some level of improvement in performance. The proposed model again stood tall by achieving the best results compared to the other CNN models based on the evaluation

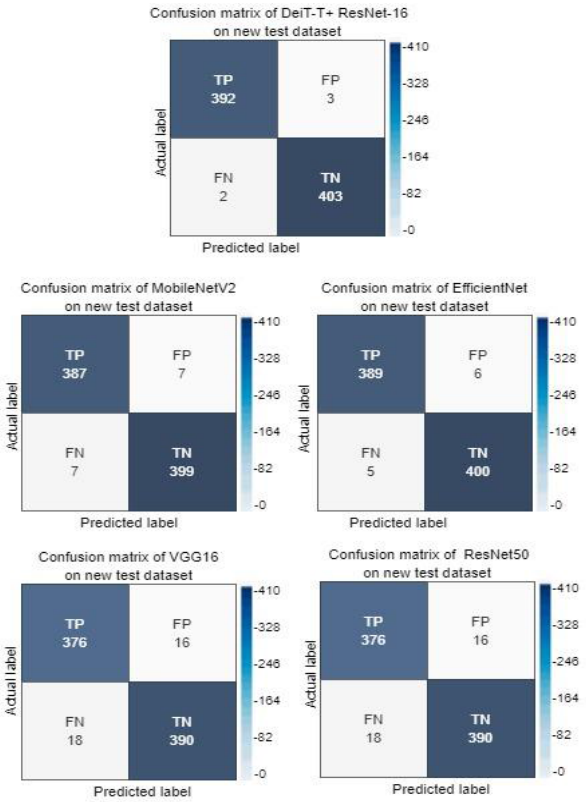

FIGURE 5. Confusion metrix of the proposed DeiT-T+ ResNet-16 and benchmark CNN models.

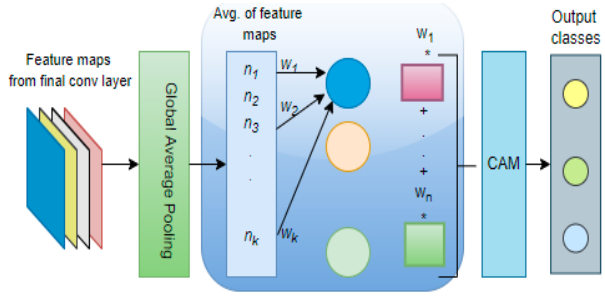

FIGURE 6. Weighted combinations of feature maps for CAM computation.

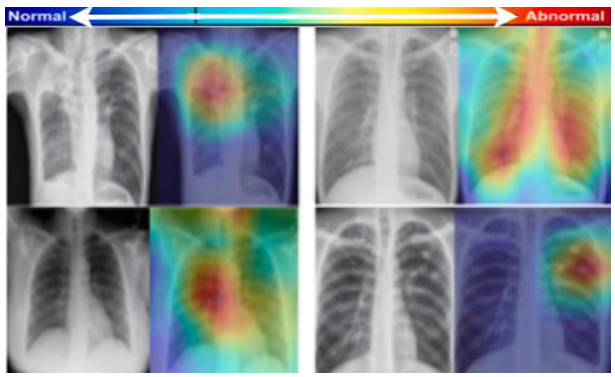

FIGURE 7. Diagnosis visualization.

metrics employed for the study. The confusion metrix of the models based on the new test dataset is in Fig. 5.

B. VISUALIZATION

In deep learning applications, seeing where the model is learning from to make a prediction is vital for its acceptance for mass screening. In this study, class activation mapping gets employed to visualize the prediction process by generating heatmaps at the affected area in the image. Activation maps show the vital regions for prediction. The Global average pooling (GAP) algorithm got utilized in creating the Class Activation Map [46] by applying it to the feature maps at the final convolution layer.

The intuition is that in Fig. 6 is that if the last convolutional layer has n filters, there will also be n feature maps. A weighted combination of n feature maps makes up the activation map for an output class. A feature map's global average is the computed average value of its pixels. The GAP outputs will consist of n scalars, for example, $k_1, k_2 \ldots, k_n$, after calculating the global average for each of the feature maps. After that, a linear model is learned from these GAP outputs and applied to the class labels (TB, sick but non-TB, and healthy). The model learns using weights $w_1, w_2, \ldots, w_n$ for each N output class. As a result, N linear models are all trained. After obtaining the n weights for each N class, the feature maps get weighted to create the class activation maps. The mathematical formula for calculating the class score for an output class c, is in "(15)".

$$y^C = \sum_k w_k^C \frac{1}{Z} \sum_i \sum_j A_{ij} k \tag{15}$$

where:

A_{ij}^k is the pixel location at (i, j) on the kth feature map. The feature map's overall pixel count is Z. The weight of the kth feature map for class c is w_k^c.

The localization outcomes for TB-related symptoms on several test instances based on CAM are in Fig. 7. The impacted regions are marked for vision, allowing doctors and radiologists to make more informed judgments throughout the examination.

C. ABLATION STUDY

We performed ablation tests on the DeiT transformer versions to illustrate the influence of transformers in the proposed paradigm. Changing the basic structure from DeiT-T to DeiT-S leads to a tiny percentage gain in efficiency at the expense of a substantially greater computational cost for the proposed model. As a result, given the computational expense, all quantitative comparisons in this article are performed with DeiT-T. In the experiment, the CNN component and transformers were removed one at a time while maintaining other conditions to determine the impact of each component on the prediction process. The findings in Table 7 show that introducing the transformer network improves the extraction of far-reaching relations with fewer parameters and computing

TABLE 7. Ablation research on the effect of CNN and transformer networks on model efficacy.

Model	Param (M)	TFLOPs	Acc (%)
DeiT-T+ResNet-16	7	0.04	98.65
DeiT-T	5	0.02	93.11
ResNet-16	2	0.05	89.68

resources. However, because of a lack of medical image datasets, the performance of the pure transformer structure is weak.

By modifying the Key (K) values while preserving other conditions, the DeiT-T-ResNet-16 model got investigated with varied patch sizes using image serialization. The experimental results reveal that model's performance diminishes when the Key value is big. In the experiment, the Key values were 2, 4, 8, and 16. The highest accuracy was obtained on Key value 2 (98.45%), while the least was on Key value 16 (90.11%). One probable explanation is that too tiny image patches degrade the image's semantic content. The custom block of the model got altered by varying the number of levels and activations. First was one dense layer, batch normalization, and Relu activation function, and the accuracy reduced to 92%. By excluding batch normalization, the accuracy dropped to 90%.

The Gaussian Error Linear Unit (Gelu) [47] activation function was introduced in place of Relu for further assessment. Gelu is an optimal neural network activation function, $x\Phi(x)$, where $\Phi(x)$ is the standard Gaussian cumulative distribution function. Unlike ReLUs that gate inputs by their sign ($X1_{x>0}$), GELU weights inputs according to their value.

However, when the Relu activation function gets substituted with the Gaussian Error Linear Unit (Gelu) activation function, with a single dense layer and batch normalization, the model obtains an accuracy of 94% and 92.5% without Batch normalization. This demonstrates that Gelu activation is marginally more successful than Relu activation in processing the outputs of the stacked transformer encoders. In the same experiment, a second dense layer got introduced while maintaining the Gelu activation function and batch normalization. This architecture yielded the best accuracy result of 98.65%. Meanwhile, with a third dense layer and the same conditions, the model's performance dropped to 95.21%. The intuition is that the model tends to over-fit beyond two dense layers while all scenarios remain constant.

IV. DISCUSSIONS

The data-efficient transformer (DeiT) and ResNet-16 formed the Backbone of the proposed model because of their efficiency and low parameter size. Compared to the original vision transformer, DeiT has fewer parameters and less training time and can produce better results even on small datasets. ResNet-16 framework has similar properties as the data-efficient transformer. The above properties contributed to the fast convergence of our model at epoch 15 with outstanding performance.

The ResNet-16 and DeiT-T transformer networks were coupled in this study to identify low and far-range feature relations from image sequences. The suggested model's remarkable results put it in the spotlight compared to prior cutting-edge models in TB diagnosis. This study demonstrates that a hybrid design based on CNN and transformers improves diagnostic effectiveness.

The model was exposed to three ablation experiments to determine the influence of each component on the classification process. The first ablation investigation demonstrates a performance difference between a pure transformer and a CNN with fewer datasets. The standard ViT's low performance is due to the lack of an inductive bias in the self-attention mechanism. Even while the transformer network gets equivalent performance with large amounts of data, this is not the same in the medical domain where the dataset is scarce.

The second ablation trial demonstrated that the model performance drops with increased Key value. According to intuition, image serialization works perfectly for natural image classification but degrades model performance in medical image classification. The reason is that image serialization separates the affected or sick region, which distorts the semantic information needed for prediction. The third ablation investigation demonstrated that the Gelu activation function performs better with transformer networks than the Relu activation function. The suggested model is fewer in parameter size (6.9M) than the benchmark models used for comparison. The model attained better accuracy, sensitivity, and specificity values of 98.65%, 99.79%, and 99.71% in classification.

The detection speed of the proposed model was also compared to the other four models employed for experimentation (ResNet50, EfficientNet-B3, MobileNetV2, and VGG-16), as shown in Table 4. According to Table 4, the suggested model diagnosed the illness from the X-ray image at a speed of 4.79 ms, 0.42% quicker than MobileNetV2, 10.11% quicker than EfficeientNet-B3, 37.49% faster than ResNet50 and 88.02% faster than VGG-16, respectively as presented in Fig. 4. The efficacy of the proposed model was further examined using a new dataset, which the model has not seen before. This new dataset was presented to the benchmark CNN models to measure their performance on the unseen dataset. The results from the CNN models were compared to the proposed model.

To further prove the efficacy of the proposed model, its performance was compared to other deep learning models proposed for TB diagnosis. The results of our proposed combined CNN-Transformer model are superior to the other deep learning techniques because those models were CNN-based and hence suffer from extracting global or long-range representations from the images. Despite those models' performances being inferior to our model, the results attained in [26] and [28] are remarkable, as indicated in Table 5.

Based on the results presented in Table 6, the proposed model (DeiT-T+ ResNet-16) outperformed all the other

models based on the assessment criteria selected for this study. The outstanding performance of the proposed model further strengthens its efficacy and potential for mass screening of TB disease, especially in underprivileged countries where the disease is more prevalent.

One factor that contributed to the suggested model's speed in terms of TB diagnosis is the use of the DeiT-T model, which is a lightweight self-attention structure that creates the micro-design. The framework differs from the typical Vision transformer. The model's pyramid shape improves the algorithm's computing efficiency by gradually reducing the spatial dimension. The depth-wise convolution in the model's final stage further decreases the model parameter, making it lightweight. All of these contributed to the proposed model's outstanding results.

V. CONCLUSION

This study provided a deep CNN transformer-based framework that could categorize the study's X-ray dataset as healthy, sick but not TB or TB to aid clinicians in making an accurate diagnosis. The experiment was conducted based on the TBX11K dataset containing three classes. Data preprocessing methods such as rescaling and cropping helped to improve data quality. The lack of medical image datasets necessitated the implementation of data augmentation techniques, including rescale (1/225), zooming, horizontal flipped, shear, rotation, and width and height shift to provide variety to the dataset and prevent the model from overfitting. We created a system for TB screening that combines the DeiT-T model and the ResNet-16 model.

The transformer networks' unique ability to extract long-range dependencies from an image and CNN's advantage to learn local information both aid in the model's efficacy. The experimental findings indicate that the DeiT-T + ResNet-16 performed better in terms of classification and diagnosis time. Other cutting-edge CNN models, including ResNet50, EfficientNet-B3, MobileNetV2, and VGG-16, were compared to the DeiT-T + ResNet-16 model. The suggested model attained classification accuracy, sensitivity, specificity precision and F1- values of 99.38%, 99.49%, 99.26%, 99.24 and 99.33. The DeiT-T + ResNet-16 model, which has fewer parameters, can diagnose the disease from the X-ray images at a speed of 4.79 ms, which is faster than all the compared benchmark models. Based on the model's effectiveness, it can help with TB illness screening in less affluent areas where the disease is more prevalent.

Future work will focus on other pretrained deep CNN and Transformer models based on transfer learning TB diagnosis for improved performance. We anticipate using neural architecture search methods to find the best parameters for the combined CNN-Transformer model that can run on edge devices in the future.

REFERENCES

[1] *Global Tuberculosis Report*, World Health Org., Geneva, Switzerland, 2023.

[2] M. Momeny, A. A. Neshat, A. Gholizadeh, A. Jafarnezhad, E. Rahmanzadeh, M. Marhamati, B. Moradi, A. Ghafoorifar, and Y.-D. Zhang, "Greedy autoaugment for classification of mycobacterium tuberculosis image via generalized deep CNN using mixed pooling based on minimum square rough entropy," *Comput. Biol. Med.*, vol. 141, Feb. 2022, Art. no. 105175, doi: 10.1016/j.compbiomed.2021.105175.

[3] B. Aljaddouh and D. Malathi, "Trends of using machine learning for detection and classification of respiratory diseases: Investigation and analysis," *Mater. Today, Proc.*, vol. 62, no. 7, pp. 4651–4658, 2022, doi: 10.1016/j.matpr.2022.03.120.

[4] I. D. Apostolopoulos and T. A. Mpesiana, "Covid-19: Automatic detection from X-ray images utilizing transfer learning with convolutional neural networks," *Phys. Eng. Sci. Med.*, vol. 43, no. 2, pp. 635–640, Jun. 2020, doi: 10.1007/s13246-020-00865-4.

[5] S. Wang, J. Cao, and P. S. Yu, "Deep learning for spatio-temporal data mining: A survey," *IEEE Trans. Knowl. Data Eng.*, vol. 34, no. 8, pp. 3681–3700, Aug. 2022.

[6] Y. Xiao, J. Wu, Z. Lin, and X. Zhao, "A semi-supervised deep learning method based on stacked sparse auto-encoder for cancer prediction using RNA-seq data," *Comput. Methods Programs Biomed.*, vol. 166, pp. 99–105, Nov. 2018, doi: 10.1016/j.cmpb.2018.10.004.

[7] U. Ahmad, M. J. Ali, F. A. Khan, A. A. Khan, A. Ur Rehman, M. M. A. Shahid, M. A. Haq, I. Khan, Z. S. Alzamil, and A. Alhussen, "Large scale fish images classification and localization using transfer learning and localization aware CNN architecture," *Comput. Syst. Sci. Eng.*, vol. 45, no. 2, pp. 2125–2140, 2023, doi: 10.32604/csse.2023.031008.

[8] J. Ao, X. Shao, Z. Liu, Q. Liu, J. Xia, Y. Shi, L. Qi, J. Pan, and M. Ji, "Stimulated Raman scattering microscopy enables Gleason scoring of prostate core needle biopsy by a convolutional neural network," *Cancer Res.*, vol. 83, no. 4, pp. 641–651, Feb. 2023, doi: 10.1158/0008-5472.can-22-2146.

[9] K. Murphy, S. S. Habib, S. M. A. Zaidi, S. Khowaja, A. Khan, J. Melendez, E. T. Scholten, F. Amad, S. Schalekamp, M. Verhagen, R. H. H. M. Philipsen, A. Meijers, and B. van Ginneken, "Computer aided detection of tuberculosis on chest radiographs: An evaluation of the CAD4TB v6 system," *Sci. Rep.*, vol. 10, no. 1, pp. 1–11, Mar. 2020, doi: 10.1038/s41598-020-62148-y.

[10] A. Abbas, M. M. Abdelsamea, and M. M. Gaber, "DeTrac: Transfer learning of class decomposed medical images in convolutional neural networks," *IEEE Access*, vol. 8, pp. 74901–74913, 2020, doi: 10.1109/ACCESS.2020.2989273.

[11] R. Elakkiya, P. Vijayakumar, and M. Karuppiah, "COVID_SCREENET: COVID-19 screening in chest radiography images using deep transfer stacking," *Inf. Syst. Frontiers*, vol. 23, no. 6, pp. 1369–1383, Dec. 2021, doi: 10.1007/s10796-021-10123-x.

[12] R. Elakkiya, K. S. S. Teja, L. J. Deborah, C. Bisogni, and C. Medaglia, "Imaging based cervical cancer diagnostics using small object detection–generative adversarial networks," *Multimedia Tools Appl.*, vol. 81, no. 1, pp. 191–207, Jan. 2022, doi: 10.1007/s11042-021-10627-3.

[13] W. Ren, A. H. Bashkandi, J. A. Jahanshahi, A. Q. M. AlHamad, D. Javaheri, and M. Mohammadi, "Brain tumor diagnosis using a step-by-step methodology based on courtship learning-based water strider algorithm," *Biomed. Signal Process. Control*, vol. 83, May 2023, Art. no. 104614.

[14] M. K. Puttagunta and S. Ravi, "Detection of tuberculosis based on deep learning based methods," *J. Phys., Conf. Ser.*, vol. 1767, no. 1, Feb. 2021, Art. no. 012004, doi: 10.1088/1742-6596/1767/1/012004.

[15] L. Devnath, S. Luo, P. Summons, and D. Wang, "Automated detection of pneumoconiosis with multilevel deep features learned from chest X-ray radiographs," *Comput. Biol. Med.*, vol. 129, Feb. 2021, Art. no. 104125, doi: 10.1016/j.compbiomed.2020.104125.

[16] K. Jin, Z. Gao, X. Jiang, Y. Wang, X. Ma, Y. Li, and J. Ye, "MSHF: A multi-source heterogeneous fundus (MSHF) dataset for image quality assessment," *Sci. Data*, vol. 10, no. 1, pp. 1–6, May 2023, doi: 10.1038/s41597-023-02188-x.

[17] P. Lakhani and B. Sundaram, "Deep learning at chest radiography: Automated classification of pulmonary tuberculosis by using convolutional neural networks," *Radiology*, vol. 284, no. 2, pp. 574–582, Aug. 2017, doi: 10.1148/radiol.2017162326.

[18] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. Ball, K. Shpanskaya, J. Seekins, D. A. Mong, S. S. Halabi, J. K. Sandberg, R. Jones, D. B. Larson, C. P. Langlotz, B. N. Patel, M. P. Lungren, and A. Y. Ng, "CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison," in *Proc. 31st Innov. Appl. Artif. Intell. Conf., 9th AAAI Symp. Educ. Adv. Artif. Intell.*, 2019, vol. 33, no. 1, pp. 590–597, doi: 10.1609/aaai.v33i01.3301590.

[19] J. Escorcia-Gutierrez, M. Gamarra, R. Soto-Diaz, S. Alsafari, A. Yafoz, and R. F. Mansour, "Optimal synergic deep learning for COVID-19 classification using chest X-ray images," *Comput., Mater. Continua*, vol. 75, no. 3, pp. 5255–5270, 2023, doi: 10.32604/cmc.2023.033731.

[20] Wikimedia Found. (2023). *Donatella Versace*. Accessed: Mar. 10, 2022. [Online]. Available: https://www.cdc.gov/niosh/topics/cwhsp/cwhsp-xray.html

[21] A. T. Sahlol, M. A. Elaziz, A. T. Jamal, R. Damaševičius, and O. F. Hassan, "A novel method for detection of tuberculosis in chest radiographs using artificial ecosystem-based optimisation of deep neural network features," *Symmetry*, vol. 12, no. 7, p. 1146, Jul. 2020, doi: 10.3390/sym12071146.

[22] K. Munadi, K. Muchtar, N. Maulina, and B. Pradhan, "Image enhancement for tuberculosis detection using deep learning," *IEEE Access*, vol. 8, pp. 217897–217907, 2020, doi: 10.1109/ACCESS.2020.3041867.

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

[24] S. Jaeger, S. Candemir, S. Antani, Y.-X. J. Wáng, P.-X. Lu, and G. Thoma, "Two public chest X-ray datasets for computer-aided screening of pulmonary diseases," *Quant Imag. Med. Surg.*, vol. 4, no. 6, pp. 475–477, 2014.

[25] S. Akbar, N. GhaniHaider, and H. Tariq, "Tuberculosis diagnosis using X-ray images," *Int. J. Adv. Res.*, vol. 7, no. 4, pp. 689–696, Apr. 2019, doi: 10.21474/ijar01/8872.

[26] R. Guo, K. Passi, and C. K. Jain, "Tuberculosis diagnostics (don't short) and localization in chest X-rays via deep learning models," *Frontiers Artif. Intell.*, vol. 3, Oct. 2020, Art. no. 583427, doi: 10.3389/frai.2020.583427.

[27] Z. Ul Abideen, M. Ghafoor, K. Munir, M. Saqib, A. Ullah, T. Zia, S. A. Tariq, G. Ahmed, and A. Zahra, "Uncertainty assisted robust tuberculosis identification with Bayesian convolutional neural networks," *IEEE Access*, vol. 8, pp. 22812–22825, 2020, doi: 10.1109/ACCESS.2020.2970023.

[28] M. Ayaz, F. Shaukat, and G. Raja, "Ensemble learning based automatic detection of tuberculosis in chest X-ray images using hybrid feature descriptors," *Phys. Eng. Sci. Med.*, vol. 44, no. 1, pp. 183–194, Mar. 2021, doi: 10.1007/s13246-020-00966-0.

[29] V. Chouhan, S. K. Singh, A. Khamparia, D. Gupta, P. Tiwari, C. Moreira, R. Damaševičius, and V. H. C. de Albuquerque, "A novel transfer learning based approach for pneumonia detection in chest X-ray images," *Appl. Sci.*, vol. 10, no. 2, p. 559, Jan. 2020, doi: 10.3390/app10020559.

[30] E. Kotei and R. Thirunavukarasu, "Ensemble technique coupled with deep transfer learning framework for automatic detection of tuberculosis from chest X-ray radiographs," *Healthcare*, vol. 10, no. 11, p. 2335, Nov. 2022, doi: 10.3390/healthcare10112335.

[31] S. Rajaraman, S. Candemir, Z. Xue, P. O. Alderson, M. Kohli, J. Abuya, G. R. Thoma, and S. Antani, "A novel stacked generalization of models for improved TB detection in chest radiographs," in *Proc. 40th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2018, pp. 718–721, doi: 10.1109/EMBC.2018.8512337.

[32] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16 × 16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[33] C. Liu and Q. Yin, "Automatic diagnosis of COVID-19 using a tailored transformer-like network," *J. Phys., Conf. Ser.*, vol. 2010, no. 1, Sep. 2021, Art. no. 012175, doi: 10.1088/1742-6596/2010/1/012175.

[34] Y. Su, Q. Liu, W. Xie, and P. Hu, "YOLO-LOGO: A transformer-based YOLO segmentation model for breast mass detection and segmentation in digital mammograms," *Comput. Methods Programs Biomed.*, vol. 221, Jun. 2022, Art. no. 106903, doi: 10.1016/j.cmpb.2022.106903.

[35] E. Kipkogei, G. A. A. Argoty, I. Kagiampakis, A. Patra, and E. Jacob, "Explainable transformer-based neural network for the prediction of survival outcomes in non-small cell lung cancer (NSCLC)," *MedRxiv*, Oct. 2021, doi: 10.1101/2021.10.11.21264761.

[36] D. Shome, T. Kar, S. Mohanty, P. Tiwari, K. Muhammad, A. AlTameem, Y. Zhang, and A. Saudagar, "COVID-transformer: Interpretable COVID-19 detection using vision transformer for healthcare," *Int. J. Environ. Res. Public Health*, vol. 18, no. 21, p. 11086, Oct. 2021, doi: 10.3390/ijerph182111086.

[37] 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, Dec. 2021, Art. no. 115519, doi: 10.1016/j.eswa.2021.115519.

[38] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," 2020, *arXiv:2012.12877*.

[39] E. Kotei and R. Thirunavukarasu, "A systematic review of transformer-based pre-trained language models through self-supervised learning," *Information*, vol. 14, no. 3, p. 187, Mar. 2023, doi: 10.3390/info14030187.

[40] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," 2020, *arXiv:2005.12872*.

[41] L. Alzubaidi, J. Zhang, A. J. Humaidi, A. Al-Dujaili, Y. Duan, O. Al-Shamma, J. Santamaría, M. A. Fadhel, M. Al-Amidie, and L. Farhan, "Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions," *J. Big Data*, vol. 8, no. 1, pp. 1–74, Mar. 2021.

[42] A. Vaswani, "Attention is all you need," in *Proc. 31st Conf. Neural Inf. Process. Syst.*, vol. 30, Dec. 2017, pp. 5999–6009.

[43] B. Wu, C. Xu, X. Dai, A. Wan, P. Zhang, Z. Yan, M. Tomizuka, J. Gonzalez, K. Keutzer, and P. Vajda, "Visual transformers: Where do transformers really belong in vision models?" in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 579–589, doi: 10.1109/ICCV48922.2021.00064.

[44] Y. Liu, Yu-huan, Y. Ban, and H. Wang. (2023). *TBX 11K Large Scale Tuberculosis Dataset*. [Online]. Available: https://www.kaggle.com/datasets/usmanshams/tbx-11

[45] Y. Liu, Y. Wu, Y. Ban, and H. Wang, "Rethinking computer-aided tuberculosis diagnosis," in *Proc. CVPR*, Jun. 2020, pp. 2646–2655.

[46] M. Mortensen, "Antecedents and consequences of team boundary disagreement.," *Acad. Manage. Proc.*, vol. 2004, no. 1, pp. M1–M6, Aug. 2004, doi: 10.5465/ambpp.2004.13862426.

[47] D. Hendrycks and K. Gimpel, "Gaussian error linear units (GELUs)," 2016, *arXiv:1606.08415*.

EVANS KOTEI received the B.Sc. and M.Sc. degrees in computer science from the Kwame Nkrumah University of Science and Technology, Kumasi, Ghana, and the Ph.D. degree in computer science from Vellore Institute of Technology, Vellore, India. He is currently a Lecturer with the Department of Computer Science, Faculty of Applied Science and Technology, Kumasi Technical University, Kumasi. His research interests include machine learning, deep learning, computational techniques for computer vision and optimized computer-aided diagnosis system development. He also has skills in big data analytics, data analysis, and data mining.

RAMKUMAR THIRUNAVUKARASU received the Ph.D. degree from Anna University, Chennai, in 2010 for his thesis titled "Synthesizing Global Association Rules in Multi-Database Mining." He is currently working as a Professor with the School of Computer Science Engineering and Information Systems, Vellore Institute of Technology, Vellore, India. Having 22 years of experience in Higher education and research, regularly he contributes papers in various reputed international Journals. Some of his research works have been published in journals from Springer, Elsevier, Wiley, World-Scientific, and other publishers. He has functioned as resource person in more than 20 Faculty Development Program and organized seminars and workshops which are funded by ISTE, AICTE, and others. He has successfully guided four research scholars to lead to Ph.D. degree.