A Vision-Transformer-based Ensemble Model for Multi-label Disease Classification on CT Scans and Chest Radiographs

Ojo Abayomi Fagbuagun[1,*], Ojo Femi Ajewole[2], Stephen Alaba Mogaji[1], Olaiya Folorunsho[1,3] & Samson Adebisi Akinpelu[4]

[1]Department of Computer Science, Federal University Oye-Ekiti, Nigeria
[2]Department of Computer Science, Federal University of Technology and Environmental Science, Iyin Ekiti, Nigeria
[3]Unit for Data Science and Computing, North-West University, Potchefstroom, South Africa
[4]School of Mathematics, Statistics, and Computer Science, University of Kwazulu-Natal, South Africa
*E-mail: abayomi.fagbuagun@fuoye.edu.ng

Abstract. The increase in the volume of X-rays and computed tomography (CT) images has drastically increased the workload on radiologists. As a result, a computer-aided solution with the capability to classify CT scans is needed to reduce this workload. In this paper, a vision-transformer (ViT) based model for multi-disease, ensemble-based classification is proposed. ViT, Data2Vec, and SegFormer models were fine-tuned to carry out the classification of selected diseases, namely, effusion, pneumonia, and pneumothorax. Normal cases of the selected diseases were included in the dataset. The datasets were obtained from two sources: the chest X-ray dataset from the Nigerian Institute of Health Chest Clinic and the optical coherence tomography (OCT) images dataset containing 6,621 images from University of California, San Diego. The images were preprocessed using random cropping, horizontal flipping, and normalization. The dataset was partitioned into training, validation, and testing sets. Model training was done in 10 epochs. The evaluation metrics showed a better performance from ensemble learning compared to other individual transformer models. The weighted average performance for all metrics was 90.56% precision, 90.58% recall, and 90.48% F1 score. The model is useful for classification of multiple diseases and can be used by radiologists.

Keywords: *C-T scans*; *classification*; *deep learning*; *normalization*; *vision transformer*; *stacked ensemble*.

1 Introduction

The domain of classifying chest diseases is an important area of study in medical diagnostics. This is due to restricted access to the chest region. The utilization of advanced technologies in the analysis of chest radiographs (CR) for the purpose of classification of chest radiographs has become imperative. Some of these advanced technologies include the use of computed tomography (CT) scans, and

Received January 4th, 2025, Revised May 23rd, 2026, Accepted for publication June 17th, 2026.

magnetic resonance imaging (MRI), which has proved useful in analytics of the chest region. A major challenge associated with the use of chest radiographs in disease classification is the variability associated with clinicians' interpretation of these radiographs. These variabilities could be due to differential visual perception of the images or the fact that some anatomical structures in the chest region are closely linked to the existence of chest diseases in the body. Radiologists use radiographs to confirm the presence or absence of chest diseases in MRI or CT scans. The diagnostic process that radiologists use requires intricate and meticulous procedures, which could be achieved more precisely with the assistance of software models that have the ability to detect the presence of disease in chest radiographs.

Various chest diseases have closely related symptoms, for example, tuberculosis, sore throat, and cough. This similarity in symptoms could lead radiologists to incorrect conclusions during the classification of chest diseases [1]. The availability of high-speed computers for use in computer-aided-diagnostics has aided scientists in the design of low-cost systems that could be deployed for achieving accurate and fast diagnostics outcomes [2]. Machine learning (ML) has made model development for classification of chest diseases easier. Meanwhile, advances in research have led to the emergence of deep learning (DL), which has made the development of models for disease classification more easy and feasible [3]. It is a known fact that most of the proposed works in the area of disease diagnosis using CR, MRI, and CT scans predominantly rely on convolutional neural networks (CNN), which have produced good results in this area. Nevertheless, there exist other learning techniques with several advantages over existing technologies, which are able to capture long-term dependencies in images [4] and therefore capture relevant features in data that could be used for classification purposes.

As a tool for ML, artificial neural networks (ANN) have achieved great feats in the domain of artificial intelligence, for example in facial recognition, healthcare, aerospace engineering, stock market prediction, signature verification, recognition of speech, and in processing of natural language (NLP), due to its non-linearity. In Jack *et al.* [5] it is reported that one of the obvious advantages of ANN is that less statistical training is required to solve a problem. Despite the successes, there are obvious disadvantages of using ANN that limit their application. These limitations include their black box nature, which makes it difficult to know why a certain unexpectedly predicted outcome is possible, as in the prediction of a car where "tiger" is the input. Furthermore, a large amount of data is required for the training of most ANN models, even though some exist that require little data. Moreover, ANN are computationally expensive due to data size, data depth, and the complexity of the networks [6]. With CNN, a more powerful and precise method of solving classification problems is available,

because data is available for training the network. Once trained on an image dataset, the knowledge acquired can be used for problem solving in other related areas of image processing [7]. As deep learning (DL) algorithms, CNN can analyze huge amounts of data by learning features that correctly define the data.

Transformers were initially introduced for natural language processing tasks [8] but later also applied to computer vision tasks as vision transformers (ViTs) [9]. In ViTs, an image is divided into patches and the attention mechanism allows every patch to decide the importance of other patches in the image to the task at hand. This enables ViTs to understand the global structure and the relationship between image patches. The key components of the self-attention mechanism maps the input into query (Q), key (K), and value (V) matrices, which are utilized in the computation of the self-attention features, as presented in Eq. (1):

$$\mathrm{Q} = \mathrm{X} \times \mathrm{W}^{\mathrm{q}} \tag{1}$$

$$\mathrm{K} = \mathrm{X} \times \mathrm{W}^{\mathrm{k}}$$

$$\mathrm{V} = \mathrm{X} \times \mathrm{W}^{\mathrm{v}}$$

where W^q, W^k, and W^v are the corresponding linear projections from input image to the matrices. The correlation and similarity between Q and K are normalized, thereby getting the attention distribution A in Eq. (2):

$$\mathrm{A}(\mathrm{Q}, \mathrm{K}) = \mathrm{softmax}\left(\frac{\mathrm{Q} \times \mathrm{K}^{\mathrm{T}}}{\sqrt{\mathrm{d}}}\right) \tag{2}$$

Where d is the dimension of the input image. The self-attention feature map is computed as presented in Eq. (3):

$$Z = \mathrm{SA}(Q, K, V) = A(Q, K) \times V. \tag{3}$$

The ViT architecture is presented in Figure 1.

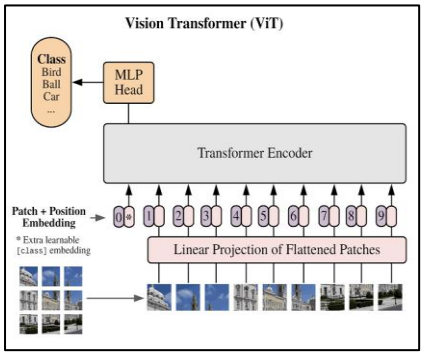

Figure 1 Vision transformer architecture.
Source: https://cameronrwolfe.substack.com/p/vision-transformers

A block figure of the self-attention mechanism is presented in Figure 2.

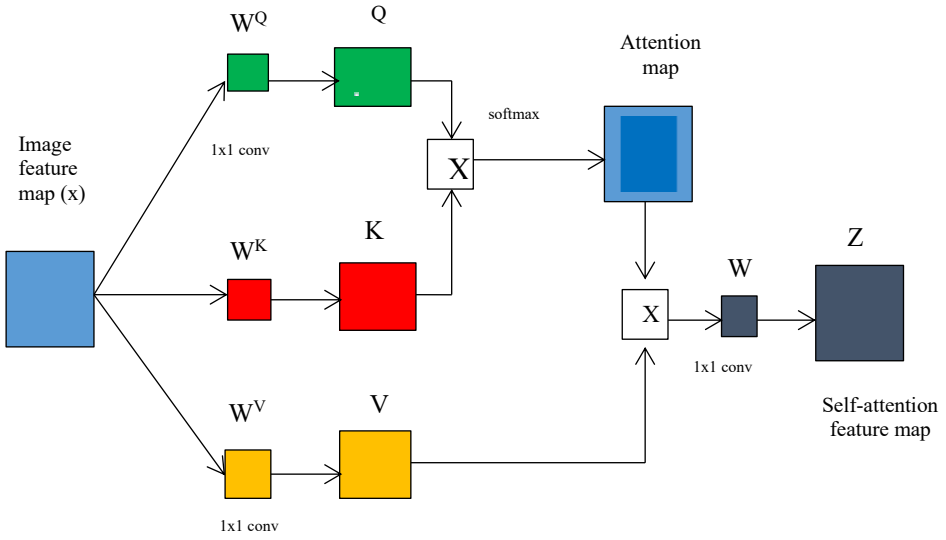

Figure 2 Self-attention mechanism of ViTs.

ViTs have been utilized in the classification of brain stroke classification on CT images, aerial image classification, and melanoma detection [10-12]. Furthermore, vision transformers have proved the ability to capture long-term dependencies that exist between input sequences, and to enable parallel processing [13]. Some benchmarks have revealed that models based on transformers outperform CNN and recurrent neural networks (RNN) due to their high performance in the absence of vision-specific inductive bias. ViTs have been applied in image classification tasks with state-of-the-art performance [15] and have been receiving attention in computer vision applications. In this proposed work, it is acknowledged that chest diseases represent a critical research area in medical diagnostics, and the utilization of advanced technologies for their identification and classification from CR images has become imperative. The first application of transformers was in NLP for tasks in machine translation with capability to track relationships between input sequences by the use of mathematical representation.

This capability, as a result of the work done by Vaswani, *et al.*, enhances the ability of transformers to handle long-range dependencies in given textual information that an ordinary ANN cannot handle in language translation. A transformer was first applied by Vaswani, *et al.* [16] as a framework that is based mainly on the use of attention mechanisms. Their application deviates from the use of recurrence networks and convolutions that are prevalent in most machine

translation tasks. The results obtained on two machine learning language translations showed that the models performed better with little time required for the training dataset as compared to traditional CNNs.

In Singh, *et al.* [17], it is stated that transformer models are the most advanced tool in language translation, a research area in NLP. Furthermore, they have been applied in several machine vision tasks with more efficiency and accuracy compared to CNNs and RNNs. One special feature of this type of model is the ability of self-attention, which equips it with the ability to model dependencies that exist between words in a text sequence. Since the breakthrough application of transformers in language translation, they have further been applied in the area of image processing and recognition. This is known as vision transformers (ViTs), which have been applied in the classification of chest diseases in the proposed method of Singh, *et al.* [17]

Transformer models are models that are based on traditional neural networks with the ability to learn context that exists in sequential data and then generate a new set of data from the existing data. The transformer architecture consists of the tokenization part, the embedding part, the positional encoding part, the transformer block, and the SoftMax part. Vision transformers are derivatives of the generic transformer architecture, where self-attention is replaced with spatial attention [18]. They were purposely designed for processing images in a 2D grid-structure. With this, vision transformers can carry out analysis and understand spatial relationships that exist in different parts of images. This architecture is useful for image classification and in computer vision tasks. ViTs works based on self-attention, exploring the global and local features of an image by focusing on different parts in it. Self-attention is used by ViTs to learn relationships between the various regions of an image while feeding the output representation of the relationships into a feedforward neural network for prediction of the outcome. As reported in the literature, ViTs have the ability to handle images with varying sizes and can train on large image datasets, like high-resolution images, without the need for downsampling or image cropping [17]. A ViT model for image recognition was introduced in a paper titled 'An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale' by Dosovitskiy *et al.* [19], published in 2021. In this work, it was disclosed that ViTs have the ability to measure relationships between input images by using an attention mechanism. With this attention mechanism, ViT accentuates some portions of a given input image while diminishing other parts of the input image by cognitive attention. The aim of doing this is to learn the important features of the input image.

2 Related Works

Berendikt, *et al.* [20] have proposed a ViT model for multi-label disease classification using CR. The objective was the generic application of vision transformers to disease classification for value addition to clinical intervention. The performance of ViT and CNN models was compared for different data sizes. It was found that ViTs models outperformed CNN when huge data is available for training. A drawback is that the performance metrics of the models varied between 62.79% and 64.93%. Despite revealing many advantages of ViTs over CNN, the proposed work could be improved in terms of performance metrics, while other metrics like specificity and AOC could also be included. Ensemble learning may further increase the performance metrics of the developed model.

Huang, *et al.* [21] have proposed a vision transformer-based multi-disease diagnosis tool using CR. The researchers aimed to explore the advantages of ViT over CNN in disease classification in order to improve the predictive capability of models for chest diseases. The authors introduced an attention module in the encoder layer to improve the models' ability to concentrate on important regions in the image. CNN was used to extract features from the images. It was discovered that the images used in the dataset were not preprocessed and this was most likely responsible for the low classification results of the developed model. Moreover, the advantages offered by the use of meta-learning were also not explored, as this has the most likelihood of improving upon the performance metrics of the model. As the metrics showed, the sensitivity was 0.863, the specificity was 0.821, while the model had an accuracy of 0.834 (83.4%). All these outcomes could be improved with better preprocessing and the introduction of ensemble learning.

Ashraf, *et al.* [22] have proposed a combination of CNN, ViT, and hybrid models for the purpose of classification of multi-label chest X-ray images. They aimed to develop a model capable of diagnosing thoracic diseases, which is challenging due to the lack of detailed information about the abnormalities associated with the diagnosis of chest diseases. In order to address this challenge, a deep learning technique was used in the identification of patterns in chest X-rays from their dataset that corresponded to different diseases. Pre-trained CNNs, vision transformer, and hybrid CNN plus transformer models were used in the experiment. The best single model achieved an AUROC of 84.2%, while the ensemble model achieved an improved AUROC value of 85.4%, outperforming other methods, as suggested by the authors. Despite achieving good results, other performance metrics that show the classification ability of the ensemble model, such as precision, accuracy, recall, and F1 score, were not used.

3 Research Method

This paper aimed to develop a ViT-based ensemble model for multi-disease classification on CR. The research methodology implemented the stacking ensemble approach with ViTs as the base of the ensemble model, because they are accurate and thus make different errors than CNNs. Variance and error rates are mostly reduced by ensemble models, while models that are individually strong but disagree on the wrong predictions made were used as base models. Furthermore, ViTs show different inductive bias than CNNs and perform stronger on large-scale data. The ensemble combined the strengths of three transformer models: VisionTransformer, Data2Vec, and SegFormer. Each transformer model underwent fine-tuning on training data to adapt to the specific classification. The ensemble was further fortified through meta-learning using a decision tree, which integrates the predictions from individual transformers to enhance the overall classification result.

The dataset was obtained from two sources: National Institute of Health Clinical Center (NIHCC) CR dataset. The NIHCC CR dataset is a rich compilation of CR images data obtained from the NIHCC [23]. For this study, four classes were selected, namely, effusion, pneumonia, pneumothorax, and normal cases,. The inclusion of a normal class ensured a balanced representation, crucial for the ability of the model to discern pathological features that are present in an otherwise healthy anatomy.

The second dataset from the OCT and CR dataset [24] contained over 5,800 CR images of both viral and bacterial pneumonia, and normal anatomy collected from University of California San Diego. A total of 6,625 images were extracted from the two datasets. The extracted images belonged to four classes: effusion, pneumonia, pneumothorax, and normal (X-rays representing a healthy state without apparent chest abnormalities). Samples from the dataset are presented in Figure 3.

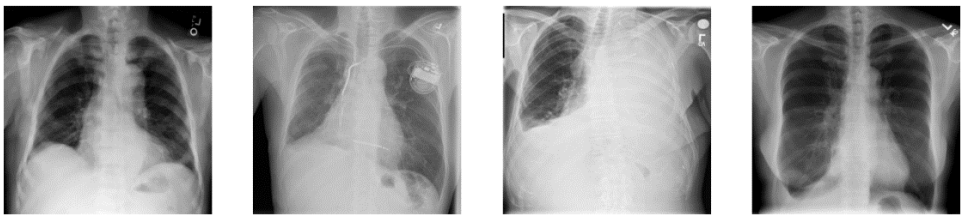

Figure 3 Samples from the dataset used in the experiment.

Data preprocessing techniques, i.e., random resize crop, random horizontal flip, and normalization, were performed on the dataset. The dataset was partitioned as presented in Table 1.

Table 1 Data splitting.

Classes	Training	Validation	Test
Effusion	1,088	158	311
Pneumonia	1,696	106	391
Pneumothorax	628	92	180
Normal	1,350	390	235

The model architecture used in the experiment is presented in Figure 4. The input were the various samples from the four classes of effusion, pneumonia, pneumothorax, and normal cases. The model was expected to classify the X-ray images into any of the four aforementioned classes.

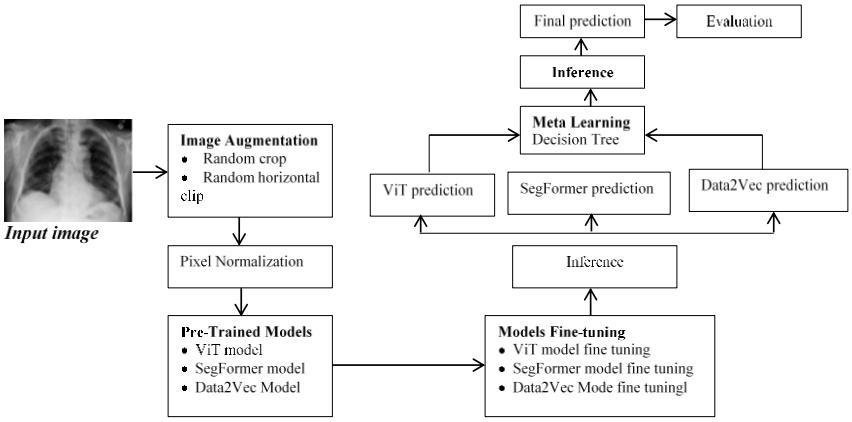

Figure 4 Model architecture.

From the architecture, each input image is augmented and normalized before being fed into the pretrained models. The models are fine-tuned by adapting them to the task and data in this work to make predictions. A decision tree meta-learner is used to combine the predictions in order to make the final prediction, after which the model performance is evaluated. The methodology consists of a model architecture that amalgamates the strengths of transformer models in an ensemble framework. Three transformer models, VisionTransformer, Data2Vec, and SegFormer, are used. In the architecture of the ViT, the input image is of the type $R^{H\times W\times C}$, with H, W, and C, representing the height, width, and the channel (RGB). The image is then split into patches of the type $R^{P\times P\times C}$, where P is the patch. This produces N patches, where $N = \frac{HW}{P^2}$ patch numbers. Next, it is

flattened into the vector X_p^0, having length $P^2 \times C$, where our $n = 1,2,\ldots N$. The sequence of embedded image patches is later derived through mapping the flattened patches into D dimensions, having linear projection E, which is trainable. Then the learnable class with embedding X_{class} is prepended to the embedded number of image patches, where X_{class} denotes the output classification Y. The patch embedding is then augmented with a single dimension positional embedding represented as E_{pos}, which adds positional information to the output. The output is then learned during the model training, which results in Eq. (4):

$$Z_0 = \left[X_{class}; X_p^1 E; \ldots; X_p^N E\right] + E_{pos} \quad (4)$$

The classification is done by feeding Z_0 into the transformer input encoder that contains a stack of L similar layers. The X_{class} value at the L-th layer of the encoder is taken and fed into the classification head. Feature extraction is done with Data2Vec, which introduces a masking strategy where certain patches are randomly replaced with a special token or other patch. The transformer encoder predicts the masked patches based on the unmasked ones and refines these predictions to generate the final latent representations of the input image. Similarly, as was done to ViT, Data2Vec underwent fine-tuning to adapt its parameters to the nuances of the CR images.

For semantic segmentation of the images, the SegFormer B4 model was used, which was designated for labelling the pixels in each image patch. It captures both local and global information or features that accommodate various resolutions without positional encoding. The SegFormer B4 model was fine-tuned on the training data, focusing on enhancing its ability to identify specific features relevant to chest disease classification. The heart of the methodology lies in the ensemble approach using decision tree, combining the strengths of three transformer models: ViT, Data2Vec, and SegFormer. Each transformer model underwent fine-tuning on training data to adapt to the specifics of the classification. The ensemble was further fortified through meta-learning using a decision tree, which integrated the predictions from individual transformers to enhance the overall classification performance. The decision tree's evaluation metrics was based on accuracy, precision, recall, and F1 score. The performance metrics of the individual models are presented, followed by the performance of the ensemble model. The Vision Transformer, after being fine-tuned on the training data, exhibited 88.14% accuracy in classifying the chest diseases. The training was carried out for 10 epochs. Each epoch lasted for 324 seconds. Figure 5 shows the evaluation of the losses on the training set and the evaluation set during training.

Figure 5 Training loss of the ViT model.

The ViT model's confusion matrix is presented in Figure 6, which represents the model's performance.

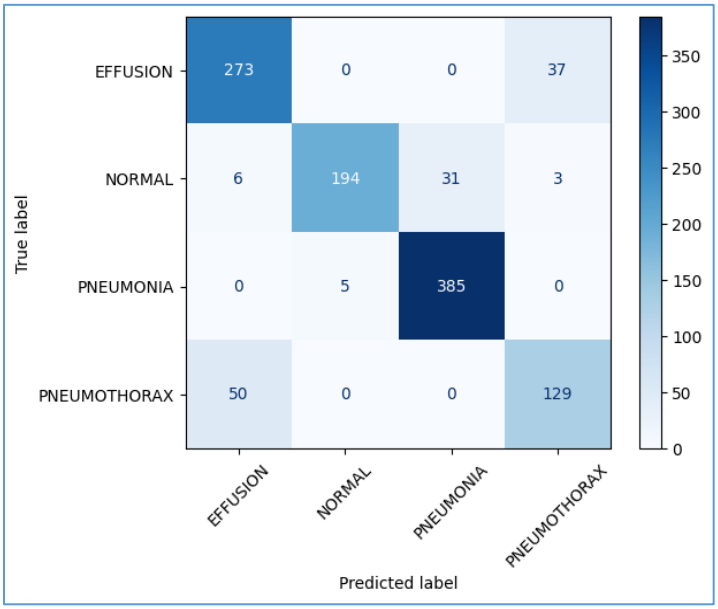

Figure 6 Confusion matrix of ViT model.

Arising from Figure 6, the evaluation metrics derived are presented in Table 2.

Table 2 ViT metrics.

Classes	Precision (%)	Recall (%)	F-1Score (%)
Effusion	82.98	88.06	85.45
Normal	97.49	82.91	89.61
Pneumonia	92.55	98.72	95.53
Pneumothorax	76.33	72.07	74.14
Macro average	87.34	85.44	86.18
Weighted average	88.31	88.14	88.04

The Data2Vec model, after fine-tuning on the training data, exhibited 88.05% accuracy in classifying the chest diseases. The training was carried out for 10 epochs. Each epoch lasted for 324 seconds. Figure 7 shows the evaluation of the losses on the training and evaluation data during the training phase.

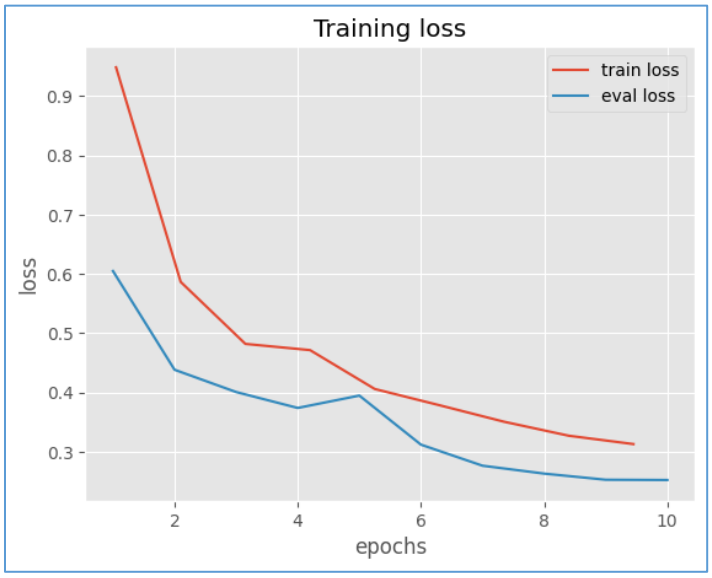

Figure 7 D2VEC training loss.

Table 3 provides a detailed insight into the model's capabilities.

Table 3 D2VEC metrics.

Classes	Precision (%)	Recall (%)	F-1Score (%)
Effusion	82.86	84.19	83.52
Normal	96.36	90.60	93.39
Pneumonia	95.01	97.69	96.33
Pneumothorax	71.19	70.39	70.79
Macro average	86.35	85.72	86.01
Weighted average	88.08	88.05	88.04

The confusion matrix of the Data2Vec model is shown in Figure 8.

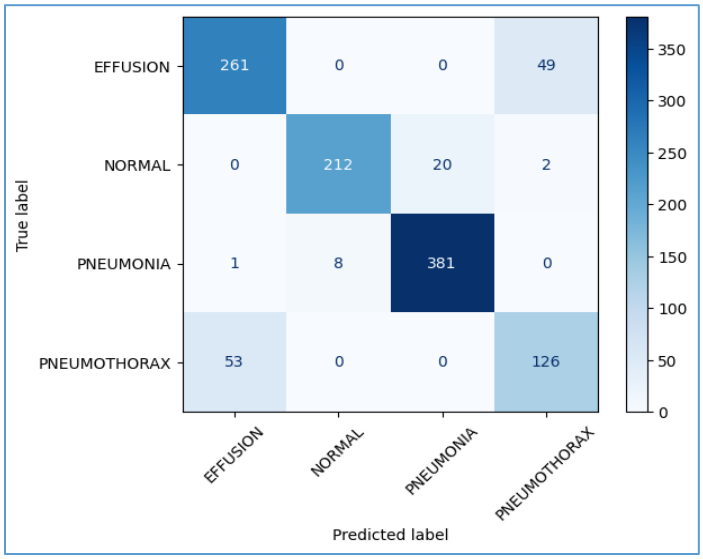

Figure 8 Data2Vec confusion matrix.

The SegFormer model, after fine-tuning on the training data, exhibited 82.75% accuracy in classifying the chest diseases. The training was carried out for 10 epochs. Each epoch lasted for 367 seconds. Figure 9 shows the evaluation of the loss on the training data and the evaluation data during the training phase. The evaluation loss of this model was less stable than that of the other two, indicating low generalization ability.

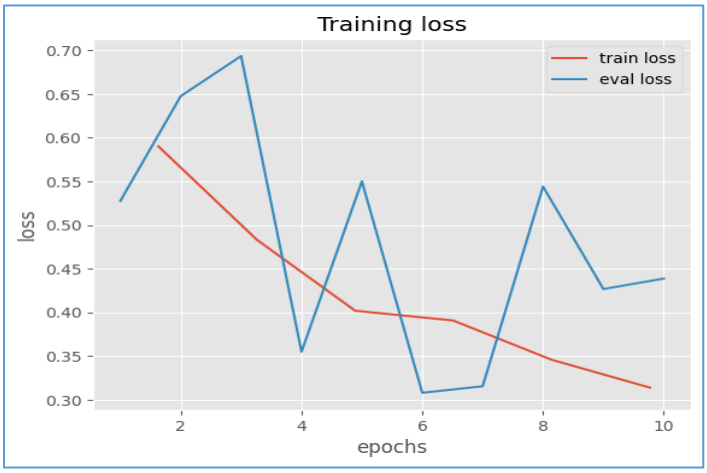

Figure 9 SFM training loss.

The confusion matrix of SegFormer model is shown in Figure 10, which shows the various components of the evaluation metrics.

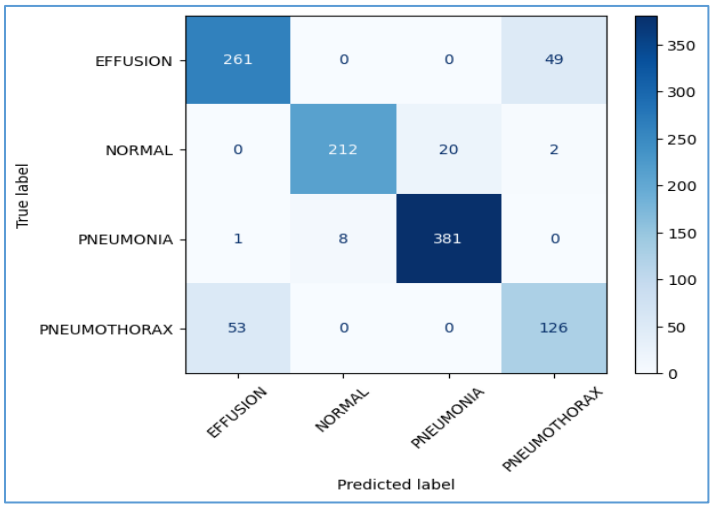

Figure 10 SegFormer confusion matrix.

Detailed insight into the model's capabilities is presented in Table 4.

Table 4 SFM metrics.

Classes	Precision (%)	Recall (%)	F-1Score (%)
Effusion	78.06	88.39	82.90
Normal	93.71	70.09	80.20
Pneumonia	86.93	97.18	91.77
Pneumothorax	68.87	58.10	63.03
Macro average	81.89	78.44	79.47
Weighted average	82.98	82.75	82.24

The ensemble model was implemented using the stacking approach. The decision tree model, after training on the output of the transformer models, exhibited 90.58% accuracy in classifying the chest diseases. The decision tree model's confusion matrix is presented in Figure 11, while the model capabilities are presented in Table 5.

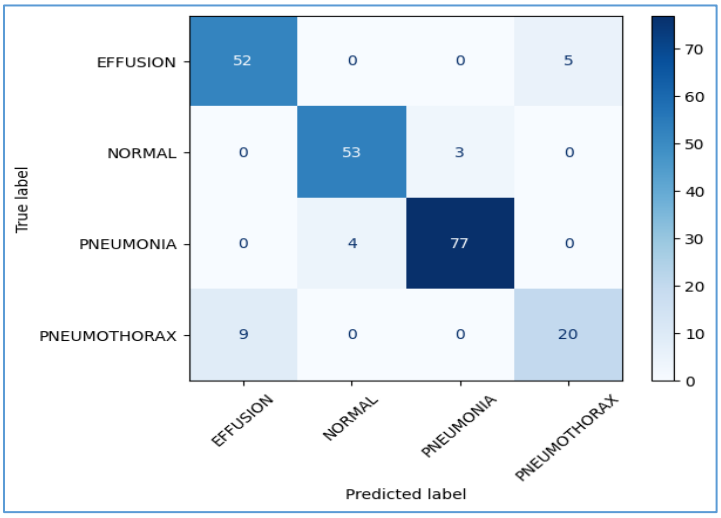

Figure 11 Decision tree confusion matrix.

Table 5 Decision tree metrics.

Classes	Precision (%)	Recall (%)	F-1Score (%)
Effusion	83.87	91.23	87.39
Normal	94.55	92.86	93.69
Pneumonia	96.30	96.30	96.30
Pneumothorax	80.00	68.97	74.07
Macro average	88.68	87.34	87.86
Weighted average	90.56	90.58	90.48

4 Comparison with Baseline Models

The results obtained from the stacking ensemble surpassed those achieved by the individual transformer models and provided a notable improvement over the baseline models. This substantiates the efficacy of the ensemble approach, showcasing its potential for advancing mordern chest disease classification. The difference between the metrics of the decision tree meta-learner and the baseline transformers are presented in Figure 12.

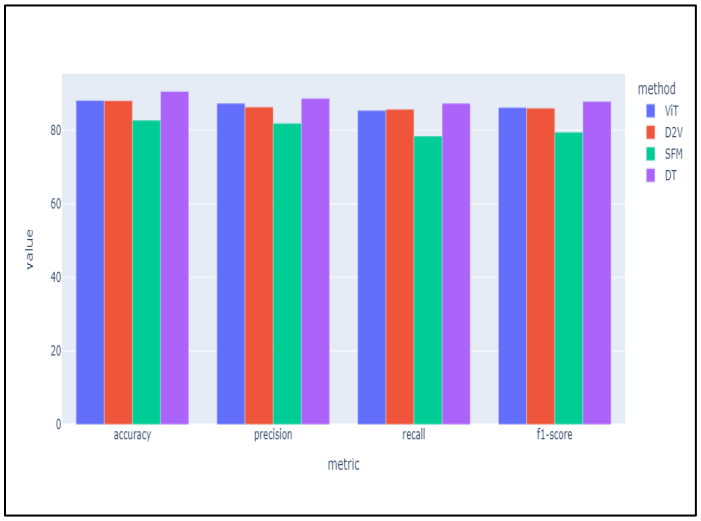

Figure 12 Comparison of the models.

The performance of decision tree was evaluated based on accuracy, precision, recall, and F1 score. The results of the ensemble model for classification achieved 83.87% precision, 91.23% for recall, 87.39% for F1 score in Effusion. It also achieved 94.55% precision, 92.86% for recall, 93.69% for F1 score in Normal CR. The classification achieved 96.3% precision, recall, and F1 score in pneumonia classification. Classification result achieved 80% precision, 68.97% for recall, and 70.07% for F1 score in pneumothorax classification. The weighted average performance of decision tree for all the metrics shows 90.56% precision, 90.58% for recall, and 90.48% for F1 score. The results obtained from the stacking ensemble surpassed those achieved by the individual transformer models and provided a notable improvement over the baseline models. This substantiates the efficacy of the ensemble approach, showcasing its potential for advancing recent chest disease classification.

5 Discussion

The results from the individual transformer models provide a nuanced understanding of their capabilities in classifying chest diseases from X-ray images. The VisionTransformer model demonstrated a high level of precision, indicating its proficiency in minimizing false positives. This characteristic is particularly valuable in medical diagnostics where the consequences of false positive predictions can be critical. The balanced F1 score further underscores the model's ability to maintain precision and recall equilibrium. The Data2Vec model showcased a well-rounded performance with a good balance between precision and recall. The accuracy and F1 score suggest its effectiveness in

capturing both positive and negative instances. The model's stability in achieving consistent results across multiple metrics positions it as a robust contributor to the ensemble.

While slightly lagging in accuracy and precision when compared to other models, Segfomer demonstrated commendable performance. The model's ability to capture true positive instances, as indicated by the recall score, ensures that it contributes valuable predictive power to the ensemble. The stacking ensemble produced notable improvements in performance metrics. Decision tree, trained on predictions from the individual transformer models, exhibited enhanced accuracy, precision, recall, and F1 score, reaching 90% across all metrics when considering the weighted average.

The drawback on the performance metrics of the models in the work of Berendikt *et al.*, which varied between 62.79% and 64.9%, was improved in this work, showing performance metrics between 76% and 97% for the various models. This is a sharp improvement on the previous work. Also, the work of Huang *et al.* had the shortcoming of not having image preprocessing and meta-learning, which were introduced in this work. The performance metrics were also improved upon because the sensitivity of our model varied between 82% and 97%.

An essential aspect of the performance of the model is its robustness and generalization to new and unseen data. The ensemble's performance on the test set and its ability to consistently achieve high metrics across different chest diseases from multiple datasets indicate its potential for real-world applicability. This paper therefore developed a single ensemble module for multi-disease classification. It also developed a model that ensures that chest X-rays images can be classified with little or no intervention from a radiologist. Furthermore, the model developed provides a better approach for disease classification not based on a traditional convoluted neural network architecture. Lastly, the model provides a tool for generating predictive outcomes of X-rays that could be used by clinicians.

6 Conclusion

The proposed ensemble model may be deployed for clinical use after due consideration for testing with private data instad of data from publicly available sources. Moreover, cross-validation needs to be done in order to monitor how well the models generalize to new and unseen data. This will help in getting a better estimate of the real-world performance of the model, detect over-fitting, reduce variance in performance estimates and ensure a more efficient use of data.

References

[1] Hassaan, M., Tayyaba, A., Ahmad, S A., Salman Z.A., Wajeeha, K. & Adnan, A., *Deep Learning-based Classification of Chest Diseases using X-rays, CT Scans, and Cough Sound Images*, Diagnostics, **13**, 2772, 2023. https://doi.org/10.3390/diagnostics13172772

[2] Liu, Z., Tong, T., Chen, L., Jiang, Z., Zhou, F. Zhang, Q., Zhang, X., Jin, Y. & Zhou, H., *Deep Learning Based Brain Tumor Segmentation: A Survey*, Complex & Intelligent Systems, **9**, pp. 1001-1026, 2023. doi: https://doi.org/10.1007/s40747-022-00815-5.

[3] Fagbuagun, O.A., Nwankwo, O., Samson S.A. & Folorunsho, O., *Model Development for Pneumonia Detection from Chest Radiographs using Transfer Learning*, Telkomnika (Telecommunication Computing Electronics and Control), **20**(3), pp. 544-550, 2022. doi: 10.12928/TELKOMNIKA.v20i3.23296.

[4] Huang, L., Ma, J., Yang, H., & Wang, Y., *Research and Implementation of Multi-disease Diagnosis on Chest X-ray based on Vision Transformer*, Quantitative Imaging in Medicine and Surgery, **14**(3), pp. 2539-2555, 2024, https://qims.amegroups.org/article/view/122245.

[5] Jack, V. & Tu, *Advantagesa Disadvantages of using Artificial Neural Networks Versus Logistic Regression for Predicting Medical Outcomes*, Journal of Clinical Epidemiology, **49**(11), pp. 1225-1231, 1996. doi: 10.1016/S0895-4356(96)00002-9.

[6] Teodoro, M.N., Félix P., Martín-Valdivia, M.T. Christine O.M. & Antonio L. Strengths, *Weaknesses, Opportunities, and Threats Analysis of Artificial Intelligence and Machine Learning Applications in Radiology*, Journal of the American College of Radiology, **16**(9), Part B, pp. 1239-1247, 2019. doi: 10.1016/j.jacr.2019.05.047.

[7] Sarker, I.H., *Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions*, SN Computer Science, **2**, 420, 2021. https://doi.org/10.1007/s42979-021-00815-1.

[8] Azad, R., Kazerouni, A, Heidari, M., Aghdam, E.K., Molaei, A., Jia, Y., Jose, A., Roy, R. & Merhof, D., *Advances in Medical Image Analysis with Vision Transformers: A Comprehensive Review*, Med. Image Anal, **91**, 103000, 2024.

[9] 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*, arXiv 2020, arXiv:2010.11929.

[10] Azhar, T., *Application of Vision Transformer for Brain Stroke Classification based on CT Images*, Eng. Technol. Appl. Sci. Res., **15**(6), pp. 29435-29439, Dec. 2025.

[11] Aboghanem, A., Abd Elfattah, M.M., Amer, H. & Tawkol, K.A.A., *Hybrid Resnet50-vision Transformer Model with an Attention Mechanism for Aerial Image Classification*, Sci Rep, **16**, 5940, 2026. doi: 10.1038/s41598-026-36492-4

[12] Garcia, A., Zhou, J., Pinero-Crespo, G., Beachkofsky, T. & Huang, X., *Clinical Application of Vision Transformers for Melanoma Classification: A Multi-dataset Evaluation Study*, Cancers (Basel), **17**(21), 3447, 2025. doi: 10.3390/cancers17213447. PMID: 41228240; PMCID: PMC12607522.

[13] Malik, H., Anees, T., Al-Shamaylehs, A.S., Alharthi, S.Z., Khalil, W. & Akhunzada, A., *Deep Learning-based Classification of Chest Diseases using X-rays, CT Scans, and Cough Sound Images*, Diagnostics (Basel), **13**(17), 2772, 2023. doi: 10.3390/diagnostics13172772. PMID: 37685310; PMCID: PMC10486427.

[14] Islam, S., Elmekki, H., Elsebai, A., Bentahar, J., Drawel, N., Rjoub, G. & Pedrycz, W., *A Comprehensive Survey on Applications of Transformers for Deep Learning Tasks*, arXiv:2306.07303, 2023.

[15] Pu, Q., Xi, Z., Yin, S., Zhao, Z. & Zhao, L., *Advantages of Transformer and its Application for Medical Image Segmentation: A Survey*. BioMedical Engineering OnLine, **23,** 14, 2024. .

[16] Vaswani, A., Shazeer, N., Parmer, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L. & Polosukhin, I., *Attention is All You Need. 31st Conference on Neural Information Processing System*, Long Beach, CA, USA, 2023.

[17] Singh, S., Kumar, M., Kumar, A., Verma, B. K., Abhishek, K., & Selvarajan, S., *Efficient Pneumonia Detection using Vision Transformers on Chest X-rays*, Scientific Reports, **14**, 2487, 2024. doi: 10.1038/s41598-024-52703-2.

[18] Chen, C., Gong, D., Wang, H., Li, Z. & Wong, K.Y.K., *Learning Spatial Attention for Face Super-resolution*, IEEE Trans. Image Process, **30**, 1219–1231, 2020.

[19] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T. & Houlsby, N., *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*, arXiv preprint arXiv, 11929. 2010.

[20] Behrendt, F., Bhattacharya, D., Kruger, J. Opfer, R. & Schlaefer, A., *Data-efficient Vision Transformers for Multi-label Disease Classification on Chest Radiographs*, Current Directions in Biomedical Engineering, **8**(1), pp. 34-37, 2022. doi: 10.1515/cdbme-2022-0009

[21] Huang, L., Ma, J., Yang, H. & Wang, Y., *Research and Implementation of Multi-disease Diagnosis on Chest X-ray based on Vision Transformer,* Quantitative Imaging in Medicine and Surgery, **14**(3), pp. 2539-2555, March 15, 2024. https://qims.amegroups.org/article/view/122245.

[22] Ashraf, S.M.N., Hasnat, A.M, & Alam, G.R., SynthEnsemble: A Fusion of CNN, Vision Transformer, and Hybrid Models for Multi-label Chest X-ray Classification, 26th International Conference on Computer and Information Technology (ICCIT), 2023.

[23] Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M. & Summers, R., *Hospital-scale Chest X-ray Database and Benchmarks on Weakly-supervised Classification and Localization of Common Thorax Diseases*, in IEEE CVPR, **7**, pp. 2097-2106, 2017.

[24] Kermany, D.S., Goldbaum, M., Cai, W., Valentim, C.C., Liang, H., Baxter, S.L. & Zhang, K., *Identifying Medical Diagnoses and Treatable Diseases by Image-based Deep Learning*, Cell, **172**(5), pp.1122-1131, 2018.