Triple Transformer Ensemble Fusion Method for Pox Virus Classification

K.P. Haripriya
Department of Computer Science, Periyar University, India
priya22prakasam@gmail.com

H. Hannah Inbarani
Department of Computer Science, Periyar University, India
hhinba@periyaruniversity.ac.in

Abstract: *Background and objectives: pox viruses are infectious agents that affect both humans and animals, often presenting similar skin lesions, making accurate diagnosis a medical challenge. Early detection and classification are crucial for outbreak control and timely clinical intervention. Automated diagnosis is essential, particularly for accurate multi-class classification. Methods: the novel ensemble method was developed to address the multi-class-wise prediction by using the Triple Transformer Ensemble Fusion Method (TTEFM). The TTEFM method was compared with existing pre-trained transformer methods, including the Vision Transformer (ViT), Mobile_ViT, and Data-Efficient Image Transformer (DEiT). The model was trained and tested using Monkeypox Skin Lesion Dataset (MSLD), which includes four classes: chickenpox, measles, monkeypox and normal. Results: the TTEFM methods outperform other state-of-the-art works. Based on the evaluation metrics, the methods are compared with other pre-trained transformers. The TTEFM method attains 99% accuracy for all the classes. The ensemble techniques were proven using the one-way Analysis of Variance (ANOVA) technique. Conclusion: the automated identification of skin lesions is crucial for clinical diagnosis, enabling dermatologists to identify and treat pox virus infections effectively. The presented TTEFM model provides a highly accurate and reliable solution for medical image classification.*

Keywords: *Vision transformer, DEiT, mobile_ViT transformer, ensemble methods, classification, medical images*.

Received April 23, 2025; accepted July 27, 2025
https://doi.org/10.34028/iajit/22/6/11

1. Introduction

Every year, the world experiences new outbreaks of viral diseases. While some have their impact, others do not. After the COVID-19 outbreak, another lethal viral outbreak emerged Monkeypox virus (Mpoxv), an endemic pathogen, still it was not a new disease. In 2024, several viruses, including the Nipah virus [18], monkeypox, and the "zombie virus," were in the news, which gained more attention due to public health risks. Mpoxv [14] is a virus belonging to the Orthopoxvirus family, with significant impacts on human health. First detected in monkeys at a Denmark research laboratory in 1958 [37], the virus subsequently spread to Central and West Africa, where it became endemic. However, in 2023, reported cases started to increase in Europe and North America, raising global concerns [5, 11]. By September 15, 2024, 122 nations had reported more than one lakh cases, a significant global health issue as the virus spread outside of traditionally endemic areas [26]. Monkeys are not the only species that can spread the virus; other species, such as squirrels and rats, can also do so. The virus will be exposed within two to four weeks and carry symptoms like fever, swollen lymph nodes, blistering rashes, and muscle and headache [22]. The rashes will begin to appear on the face, palms, and other parts of the body. Occasionally, if the condition becomes severe, the virus may cause the person to die. The primary diagnosis of monkeypox is Polymerase Chain Reaction (PCR), which is not available in remote areas, and the cost of the testing is too high.

The statistical rate of the disease is rising year by year. So, the automatic identification of the disease is required based on the computer design. Nowadays, the automatic system has evolved in all domains, including object detection, plant disease identification, and the healthcare sector. In the beginning, Convolutional Neural Networks (CNNs) performed well in the classification of diseases, but they could not handle overlapping feature differentiation in images, which created difficulties in multi-class classification. To address this, pre-trained deep learning models have been proposed, enabling improved feature extraction, the detection of complex patterns, and reduced computational costs. These models have proven highly promising in medical image analysis, enhancing the accuracy and efficiency of disease detection.

In contrast to CNNs and other pre-trained CNN variants, which are based on local receptive fields, transformers have changed image classification by using self-attention mechanisms to store long-range dependencies and global context. Special attention is given to all the affected red bump areas in the human body. Especially, Vision Transformer (ViTs) segment images into patches and process them as sequences, which facilitates efficient feature extraction without convolutional biases. Likewise, Mobile_ViT and Data-Efficient Image Transformer (DEiT) are some examples of pre-trained transformer models that improve performance by learning hierarchical

representations, especially in multi-class classification. Although computationally demanding, transformers are great at processing small datasets and, therefore, are a strong candidate to replace CNNs for deep image classification.

1.1. Problem Statement

Among all the skin lesions that are the consequence of different viral infections, the categorization of these lesions presents the greatest problem. This is because the lesions found within and between classes tend to exhibit visual characteristics that overlap. The differentiation between one infection and another is made more difficult by the presence of red bumps and similarities in texture patterns that are shared by several types of viruses. The visual resemblance among viral infections makes it difficult to accurately identify them, particularly in situations when the lesions are not easily distinguishable from one another. This might result in a potential misdiagnosis and a delay in treatment.

To address the challenges of inter-class and intra-class similarities of the pox virus classification, the proposed of Triple Transformer Ensemble Fusion Method (TTEFM) was introduced to reduce the bias and to provide balanced treatment for all the classes. These methods integrate the strengths of three pre-trained transformers, including ViT, DEiT, and Mobile_ViT, with a hard voting strategy. In this method, all three models independently predict the same input image, and the majority of the three predictions is chosen as the final result. This ensemble system enhances stability, decreases variance, and increases the model's ability to generalize between visually similar skin lesion classes.

1.2. Major Contributions

The primary contributions of this work are evident in the initial stage, where preprocessing has been finalized using one-hot label encoding to convert all labels into a machine-readable format, and normalization of the transformer has been executed. The ensemble method of the TTEFM model was introduced to tackle classification on a per-class basis. Various pre-trained transformers, such as the ViT, DEiT, and Mobile_ViT methods, are utilized to train and test the images. A one-way Analysis of Variance (ANOVA) test was conducted to demonstrate that the ensemble model is distinct from other methods. The evaluation metrics indicated that the superior and more reliable model successfully classified the pox virus images.

The passage describes the remaining parts of this paper. In section 2, different authors' proposed work and the different benchmark methods are discussed. Section 3 discusses the methods used in this work, which include the pre-trained transformer and the ensemble methods of TTEFM. Section 4 presents a discussion of all the results, including the evaluation metrics. Finally, the conclusion of the work was discussed with its limitations and future directions.

2. Related Works

In this area, classical machine learning and deep learning approaches are used in a wide variety of real-world settings to recognize objects and classify photographs, particularly in medical and normal images. Numerous medical situations call for the utilization of automatic detection. Previous work applied to different datasets is discussed with their techniques and measures.

2.1. Classification for Pox virus and Monkeypox Virus Using ML and DL Techniques

In this investigation, Luong *et al*. [17] utilised a collection of images, including monkeypox skin lesion images and a monkeypox image dataset. Using deep learning methods like ResNet50, VGG16, and MobileNet, these characteristics were retrieved. Following this categorization, machine learning techniques such as the AdaBoost method, decision trees, logistic trees, random forests, K-Nearest Neighbors (KNN), and Gaussian naive bayes were used. With a 97% success rate, the combination of MobileNet with logistic regression produces a progressive outcome. Maqsood *et al*. [19] reported using the Monkeypox Skin Lesion Datasets (MSLDs). Using deep learning methods like ViT, swin Transformer, ResNet 50, ResNet 101, EfficientNetV2, and ConvexNet V2, deep models are employed for feature extraction. Then, feature fusion and selection are performed using optimisation techniques, such as the entropy-controlled firefly approach. Finally, the classification was done based on a multi-class support vector machine with 98.65% accuracy. The proposed application for mobile devices with human monkeypox detection capabilities, utilizes an advanced deep-learning techniques to achieve a successful classification. To maintain the robustness of performance, the study used models such as ResNet18, GoogleNet, EfficientNetB0, NasNet Mobile, ShuffleNet, and MobileNetV2 in disease identification. Surprisingly, the proposed system worked by achieving a high level of accuracy (91.11%) in binary classification, indicating that it could be used as a useful tool for finding and diagnosing Monkeypox early on [25].

Bala *et al*. [2] took a MSID from the Kaggle repository. The author performed data augmentation to increase the dataset count and then separated the training and testing data. The proposed CNN model then employs machine learning and deep learning approaches. Sitaula and Shahi [30] tune all types of hyperparameters to identify the optimal model. The proposed monkeypox net method achieved a multi-class

classification rate of 98.91%. In this investigation, the collected images are from publicly available Google Images. A deep learning method was applied to classify multi-class pox virus disease. Various deep learning methods were applied, including VGG16, VGG19, ResNet50, ResNet101, IncepResNetv2, MobileNetV2, InceptionV3, Xception, EfficientNet-B0, EfficientNet-B2, DenseNet-121, and DenseNet-169, and were selected the top two accuracies. Xception and DenseNet-169 were selected to generalize the new images for the majority voting ensemble method. Muthulakshmi *et al*. [21] achieved 85.44% accuracy in this ensemble method.

2.2. Classification Using Vision Transformer and DEiT Transformer

In 2024, Hussain *et al*. [8] utilized the poles of the prismatic cell LIBs dataset for the laser welding photographs. The photos were enhanced to increase their quantity, subsequently extracting features using the VGG16 and MobileNet approaches. These features are now integrated into a single vector. The ViT methodology employs all these attributes for classification purposes. All eight categories are highly classified, and the model's accuracy is 97%. In 2025, the method [29] utilizes three distinct iris datasets. The three different feature maps employed to extract image features: Central Local Adaptive Binary Patterns (CLABP), Left Local Adaptive Binary Patterns (LLABP), and Right Local Adaptive Binary Patterns (RLABP). The ViT technique now utilizes each of these three methods independently for classification purposes. They introduced a novel model that combines three components into a cohesive feature. The features as patches were allocated and subsequently employed the ViT model to reduce the model's error rate. In 2024, Ulukaya and Deari [35] utilized the annotated rice disease data for classification purposes. They used data from five distinct rice disease categories to construct a model. The foundational model is the ViT base 32, which implements various factors to achieve optimal accuracy. The model incorporates a fine-tuning parameter with data augmentation and employs the categorical focal loss entropy method. The proposed approaches are compared with other leading publications, resulting in an accuracy of 88.57% across five distinct classifications.

2.3. Classification Using Mobile_ViT Method

According to Ding and Yang [9], they collected an apple leaf dataset from the Ningxian Modern Agricultural Industrial Park in Qingyang City, Gansu Province, and the Haisheng Apple Planting Base in Yulinzi Town, Zhengning County. Five different classes were token and used deep learning approaches to perform the classification. These methods include the ViT, the Mobile_ViT transformer, and the swim transformer. The improved Mobile_ViT method produces a better result when compared with state-of-the-art works. They achieved an accuracy of 98.54%. Zhu *et al*. [38] stated that the corneal ulcer, located in the human eye, they employed deep learning techniques. Initially, they applied the Mobile_ViT method, which improved the extraction of local and global features. The proposed method produces classification accuracy in the range of 88.7% to 91.5%, respectively. Gradient-weighted class activation mapping visualizes all the extracted features.

Table 1. Summary of the pre-trained CNN and Transformers in detail.

Author	Year	Dataset	Feature extraction	Feature selection	Classification	Best techniques with accuracy
Luong *et al*. [17]	2023	Monkeypox Skin Lesion images	ResNet50, VGG16, and MobileNet	X	AdaBoost method, decision trees, decision trees, logistic regression, random forests, KNN, and Gaussian naive bayes	MobileNet with logistic regression-97%
Maqsood *et al* [19]	2024	Monkeypox skin lesion images	ViT, Swin transformer, ResNet 50, ResNet 101, EfficientNetV2, and ConvexNet V2, deep models	Entropy-controlled firefly	Multiple SVM	98.65%
Sahin *et al*. [25]	2022	MSLD	X	X	ResNet18, GoogleNet, EfficientNetB0, NasNet Mobile, ShuffleNet, and proposed (MobileNetV2)	MobileNetV2-91.11%
Bala *et al*. [2]	2023	MSLD	VGG16, ResNet50, MobileNetV1, Inception V3, Xception	X	LR RF SVM K-NN XGBoost	MobileNetV1+LR-90.64%
			X	X	VGG16, ResNet50, MobileNetV1, Inception V3, Xception, and modified CNN(MOXNet)	MOXNEt-98.91%
Sitaula and Shahi [30]	2022	Monkey skin lesion images dataset	X	X	VGG16, VGG19, ResNet50, ResNet101, IncepResNetv2, MobileNetV2, InceptionV3, Xception, EfficientNet-B0, EfficientNet-B2, DenseNet-121, and DenseNet-169	Xception and DenseNet-169 along with hard voting techniques -87.13%
Din *et al*. [8]	2024	Laser welding images	VGG16 and MobileNet	X	Hybrid ViT model	97%
Ulukaya and Deari [35]	2025	Rice image dataset	X	X	Mobile Net V2, Efficient Net B7, VGG 19, Inception V3, ResNet, Mobile Net V2+FT, Efficient Net B7+FT, VGG 19+FT, Inception V3 +FT, ResNet 152+FT, ViT B16, ViT B32, ViT B16 +FT, ViT B32+FT	88.57%
ALkahla *et al*. [1]	2024	Ovarian cancer	X	X	Uni-Swin_T, Parallel_Swim_T	96.05%
Haripriya and Inbarani [12]	2025	Monkeypox skin lesions	X	X	CNN, VGG16, VGG19, Resnet50, Hybrid Fuzzy PCA VGG16 Method	91.42%

Table 1 summarise the pre-trained models of CNN and transformer methods used in their applications. Monkeypox, apple plant disease, human iris, brain, and lung images are studied using distinct sets of images in both controlled and uncontrolled contexts. These analyses are based on the studies discussed above. For the most part, illness categorization is accomplished through the utilization of pre-trained methods such as CNNs, transformers, and other models. Transformer models are the foundation for defining monkeypox pictures because they are trained on a small dataset, which yields great accuracy at all stages. The pre-trained transformer techniques are applied to the dataset. Having said that, not every class is given the same amount of attention.

3. Materials and Methods

This section talks about the preprocessing techniques and the transformer methods. To import the various categories of images as input, along with their relevant dimensions. Following the import of the images, the next step is the preprocessing stage, where the data augmentation and label encoding processes are carried out. Fivefold cross-validation was used as the basis for the assessment procedure that was carried out on this model. Following the completion of this, other deep learning algorithms for transformers are implemented. When it came to determining the final prediction classes, the ensemble techniques were successfully utilized. Based on the results, the assessment metrics were carried out to determine which approach was superior. Finally, we classified each approach based on the classes it belonged to. Figure 1 provides a graphical representation of the processes for the approaches.

3.1. Image Pre-processing

All the images in this section undergo image preprocessing techniques to enhance and reduce noise. The different techniques can be applied to images.

In this case, label encoding and data augmentation are performed.

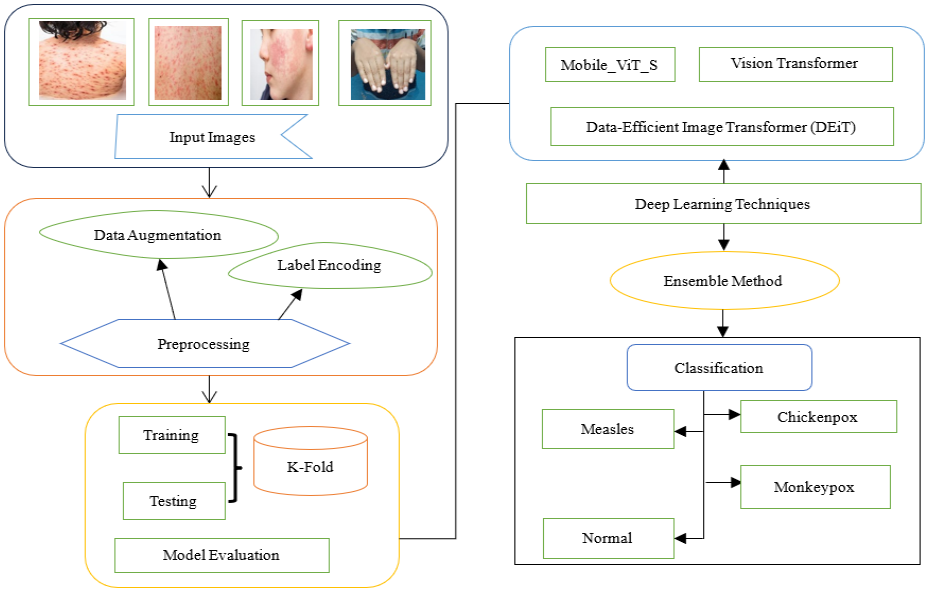

Figure 1. Workflow of the methodology.

3.1.1. Label Encoding

Generally, the class labels are in categorical form. Machines cannot directly process categorical values. Consequently, we performed one-hot label encoding on those categorical labels. Now the labels are encoded into a machine-readable format. The class labels are now formatted as follows: for chickenpox 0, for measles 1, for monkeypox 2, and normal 3. Likewise, the label encoding was performed on the image labels.

3.1.2. Data Augmentation

A small number of images is not enough to train the deep learning techniques. To traditionally increase the count of images, data augmentation was performed. The parameters used for image augmentation are described in Table 2. The rotation, shear, zoom, width, and horizontal shift fill modes were performed. In each folder, the count of the images is typically increased. An evaluation of this model was performed.

3.2. Model Evaluation

The model evaluation was done based on K-fold cross-validation. It works better in generalization for machine and deep learning methods [32]. In general, if the validation is applied to the model, then it will start its training for k-1 times as per the given k. The data will take k-1 for training the model, and the last part will be for testing. It will just start iterating the models until the K value is attained. The cross-fold validation will work well for unseen images [36].

3.3. Vision Transformer

Convolution and the pooling layer are often the foundations upon which CNN operates, extracting the characteristics from the picture as raw data. The dense layer will generate the final result once the picture features have converged. The ViT approach was created to address CNN's drawbacks. It was created by Dosovitskiy [10] to use Natural Language Processing (NLP) to handle audio and text. Later, it appears in medical image processing, which uses the self-attention mechanism to provide outstanding outcomes. The self-attention mechanism, which pays particular attention to more significant characteristics, is a key component of deep learning.

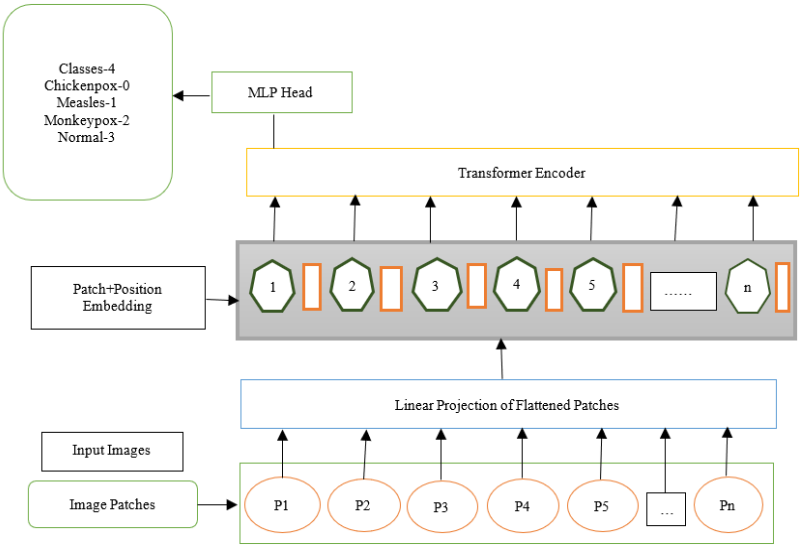

Figure 2. The architecture of the ViT model.

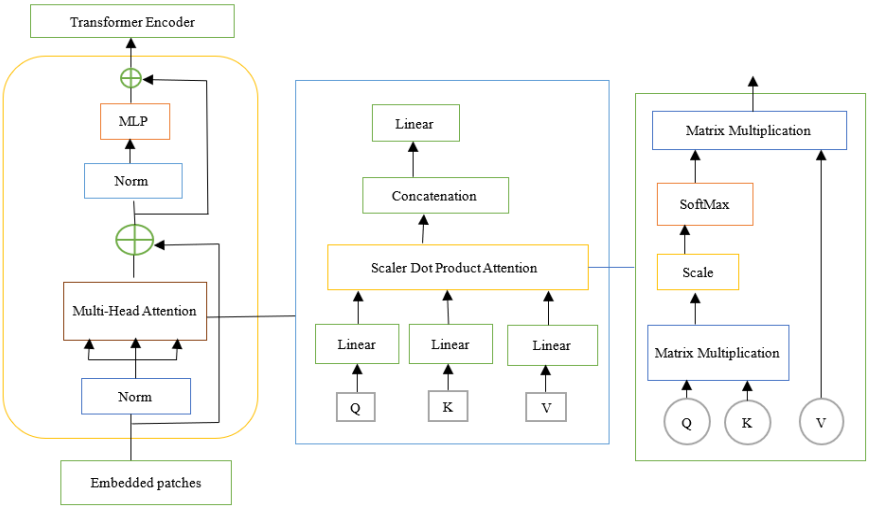

Figure 3. Details of the transformer encoder.

The input pictures are separated into patches according to H×W×C, where C represents the number of channels and H and W represent the image resolution. N*(P^2*C), where N is (H*W/P^2) and (P, P) represents the resolution of split patches, results from sequencing the patches into a 2D flattened form [15]. The linear projection transforms all these 2D patches into 1D flattened data. Figure 2 illustrates the position embedding steps employed to organise the patch information, thereby preventing data from being mixed up [27]. The hexagon shape indicates the position of the pixel, and the cylinder shape represents Pixel values in 1D format. The transformer encoder now receives the embedded patches. Within this Multi-Head Self-Attention (MSA) mechanism, a separate one was formed for each patch. The 1D information is divided into three types of matrices: Query (Q), Key (K), and Value (V). These work together in the self-attention mechanism to figure out how different parts of an image relate to each other. The Query (Q) shows what a specific part is searching for in other parts, while the Key (K) explains what each part contains and how relevant it is to the others. After applying attention, the Value (V) transmits the actual information. By calculating attention scores using Q and K, the model identifies which parts are most important and uses V to update their representation. This technique allows the model to effectively capture long-range dependencies in images. The attention score is calculated by using the Scaler dot product function. Equation (1) shows the

functionality of the dot product. Figure 3 explains the working procedure from the embedded patches to the transformer encoder.

$$Attention = Softmax\left(QK^T|\sqrt{d_k}\right)V \quad (1)$$

Where QK^T computes the similarity between the patches of images. The square root of d_k is used to scale the values so that the negative value can be ignored. Now the values are multiplied by V, deciding how much information can be taken for the final decision. Due to the small size of the datasets, this work used a pre-trained ViT. If the size of the dataset is small, then a trained transformer can be used to improve the generalization and reduce the computational time for training the model. Despite this, the model yields superior accuracy as it specifically targets the affected lesions of bed bumps for analysis. After extracting the features from the MSA mechanism, a Multi-Layer Perceptron (MLP) was used to classify the pox virus disease.

3.4. Mobile_ViT Transformer

The Mobile_ViT is a compact ViT model specifically engineered for mobile devices. It incorporates multiple convolutional and transformer blocks to improve both the speed and accuracy of image classification and various computer vision tasks.

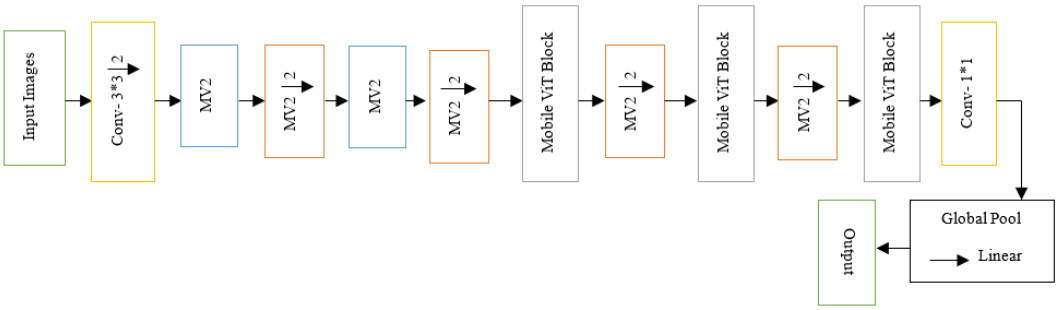

Figure 4. The architecture of Mobile_ViT_S transformer

The process initiates with a convolutional layer designed to extract local processing features from the images. MV2 facilitates the down-sampling of these features by a factor of two. This procedure persists through multiple blocks, culminating in the mobile ViT Block, which maintains compliance with Conv 1x1. A global pooling layer is applied before the output layer [20], as depicted in Figure 4.

Inside the Mobile_ViT block, each local and global representation of the features is extracted. In the initial stage, the input is in the dimension of C*H*W which extracts the local features based on the convolutional layer starting from n*n dimension to 1*1 dimension. With the help of local representation, the global features are extracted by using three different techniques such as fold, unfold, and transformer block. By concatenating the local and global features again the MV2 blocks perform again. This process is explained in Figure 5.

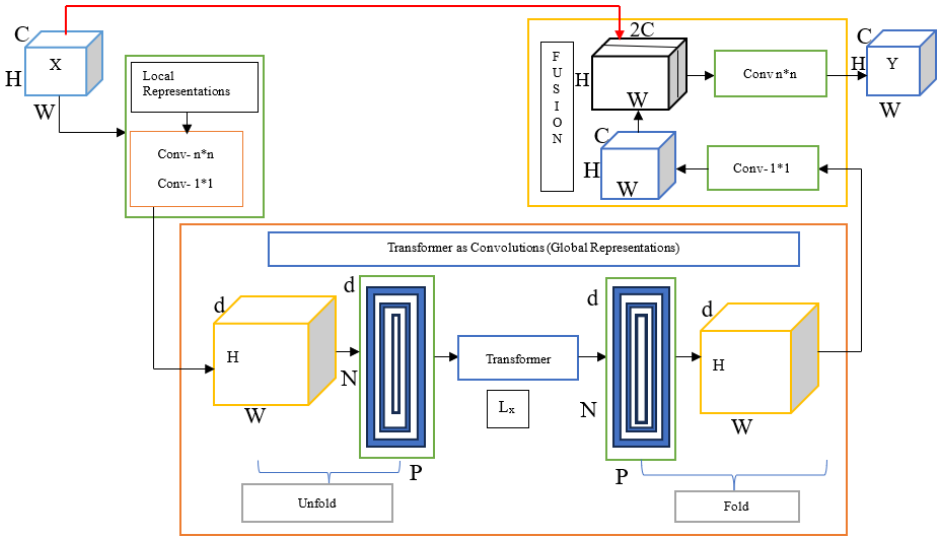

Figure 5. Workflow of Mobile_ViT block

The difference between the traditional ViT method and the Mobile_ViT method is in traditional ViT all the input images are divided into patches and then positional embedding is performed with encoding. Since the VIT methods need a large dataset and computational time. The mobile _vit method uses the unfold, transformer, and fold method to extract the features effectively. If handles the small dataset and

computational time effectively [6].

3.5. Data Efficient Image Transformer (DEit)

The DEiT is a ViT model proposed by Touvron *et al.* [34] to make training transformers more efficient without needing large-scale datasets. Classic ViT depend on huge labelled data, e.g., ImageNet-21k, which makes them expensive to train computationally and out of the question for small datasets [10]. DEiT overcomes this challenge by including a distillation token, through which the model can learn not just from the input data but also from a teacher network, often a CNN such as ResNet [34]. This improves learning efficiency, and the model can generalize well even with small datasets such as ImageNet-1k.

One of the most important innovations of DEiT is its distillation process, which enhances data efficiency without compromising accuracy [16]. In contrast to conventional ViTs that need enormous amounts of labelled data, DEiT learns efficiently by distilling knowledge from a pre-trained CNN, minimizing the requirement for large datasets [34]. In addition, DEiT is computationally efficient and optimized for speed, which makes it a feasible solution for real-world applications with limited resources. Through the use of both self-attention and knowledge distillation, DEiT provides similar performance to baseline ViTs while reducing data and computational needs substantially

3.6. Ensemble Transformer

The ensemble learning technique constructs a model by integrating multiple algorithms to achieve enhanced outcomes [28]. Nowadays the transformer methods are used for the classification of medical images. Three different transformation models such as ViT, DEiT, and Mobile_ViT method. Vision transformers (ViT, DEiT, and Mobile_ViT) revolutionise image processing by treating images as patch sequences. It uses self-attention to capture global context and performs well on large datasets. DEiT is especially improved in efficiency on smaller datasets with optimized training and distillation techniques. Mobile_ViT combines transformers and convolutions for better results in resource-constrained environments. These three prediction outcomes are combined using the hard voting ensemble method, and the respective equations are described from Equations (2) to (5)

$$X = ViT\ Transformer( ) \tag{2}$$

$$Y = DeiT\ Transformer\ ( ) \tag{3}$$

$$Z = Mobile\_ViT\ Transformer\ ( ) \tag{4}$$

For each test sample X in the $D^k_{\ test}$, we will compute predictions $P(X)$, $P(Y)$, $P(Z)$, where $P(X, Y, Z)$ $\mathcal{E}\{1, 2, 3, 4\ldots\ldots..., C\}$ and C is the number of classes. Combine predictions using hard voting:

$$P_{ensemble}(X) = argmax_c \sum_{m \in \{ViT,DeiT,Mobile\_VIT\}} \Pi(P_m(X) = C) \tag{5}$$

Where $\Pi(.)$ is the indicator function.

Algorithm 1: TTEFM method.

Input: Pox Virus Images
Output: Classification based on Classes
Steps:

1. *Initiate*
2. *Define the dataset $D = \{X_i, Y_i\}^N$ i=1, where X_i denotes the input images and Y_i represents the corresponding labels*
3. *Load the Dataset D. Ensure X_i $\mathcal{E}^{R\,H\,*\,W\,*\,C}$, where H, W, and C are the image's height, width, and channels.*
4. *Use the K-Fold cross-validation:*
   - *4.1 Split D into k folds $D= \{D_1, D_2, D_3, \ldots\ldots\ldots, D_k)$ where $D^k=(D^k_{\ train}, D^k_{\ Val}, D^k_{\ test})$.*
   - *4.2 For each fold K:*
     - *$D^k_{\ train}$ is used for training.*
     - *$D^k_{\ val}$ is used for validation.*
     - *$D^k_{\ test}$ is used for testing.*
5. *Load the three transformers: Let M_{ViT}, M_{DEitT}, and M_{Mobile_ViT}.*
   - *Train each model on D_K Train for all K.*
   - *Save the predictions P_{ViT}, P_{DEiT}, and P_{Mobile_ViT}*
6. *Perform Ensemble with Hard Voting using Equation (5).*
7. *Utilize the performance metric to assess the effectiveness of the model.*
8. *Classifications are conducted according to class.*
9. *Halt.*

The model was designed utilising the aforementioned algorithm. The model is an ensemble comprising ViT, DEiT, and Mobile_ViT, utilising a hard voting method for decision-making. The process involves generating independent class labels for each model based on the input image. Subsequently, it aggregates the votes and selects the majority class as the final output, effectively utilising all three methods. In the event of a tie, the final decision is determined by selecting the model that exhibits the highest validation accuracy as the best prediction. Alternatively, a predefined priority order may be employed to resolve ties in instances of comparable model performances. This ensemble method ensures enhanced robustness and accuracy by leveraging the complementary strengths of ViT, DEiT, and Mobile_ViT. Figure 6 delineates the methodologies employed.

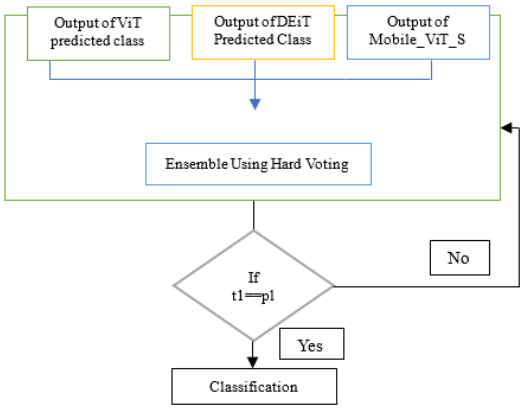

Figure 6. Ensemble method of TTEFM model.

The architecture style of the ensemble model varies when the components are trained independently. The

predicted model exhibited non-uniform calibration throughout its entirety. Certain models may rely on overconfidence and under confidence for specific classifications, leading to biases. The hard voting ensemble method was employed to enhance model robustness and mitigate bias, offering a straightforward approach for integrating heterogeneous models.

4. Result and Discussion

This section presents the outcomes of each transformation model and the ensemble methods, accompanied by their corresponding evaluation metrics. The initial subsection addresses the dataset and the simulation of system information. This subsequent sub-section presents the results along with their corresponding metrics.

4.1. Simulation and Information about Dataset

The Kaggle website hosts the publicly available Monkeypox Skin Image Dataset (MSID) [3]. All the images are in colour with 3 channels: red, blue, and green. The image format is .png, 224 x 224. The dataset contains 4 different classes: chickenpox, measles, monkeypox, and normal. The total number of images for originals is 770. Since we use deep learning techniques, we need more than just a few images to train the model and achieve better generalization results. To virtually increase the count of datasets, we performed image augmentation. Table 2 describes the original and virtually enhanced image count details.

All these transformer models were performed on Windows 11, with a 13th Gen Intel (R) Core (TM) i7-13620H processor operating at 2.40 GHz and 16 GB of RAM, and supported by a GPU RTX 4050 system. The algorithm was developed using Anaconda Navigator with Jupyter Notebook, along with supporting packages such as torch, scikit-learn, timm, NumPy, Pandas, OS, CSV, Seaborn, and Matplotlib.

Table 2. Details of the dataset.

Class label name	Original image	Augmented image
Chickenpox	107	508
Measles	91	591
Monkeypox	279	779
Normal	293	793
Total	770	2671

Table 3 lists the parameters used for image augmentation. We enhance the images using rotation, shearing, zooming, horizontal flip, and vertical flip. We gather new dimensions from the enhanced images to improve the model's performance for unseen images.

Table 3. Parameter of image augmentation.

Parameter name	Parameter size
Rotation Range	40
The range for width shift	0.3
The range for height shift	0.3
Shear Range	0.3
Zoom Range	0.3
horizontal flip	True
vertical flip	True
fill mode	Nearest

4.2. Result Analysis and Measures

4.2.1. Performance Metrics

To compare the proposed method to existing benchmark methods, the statistical metrics used here are accuracy, precision, recall, F1-score, balanced accuracy, specificity, and geometric mean. The computation is done based on a confusion matrix. A confusion matrix for multi-class classification has N labels with N actual and expected values. Based on the confusion matrix, evaluation metrics of the transformer methods and their description are described in Table 4. Based on these metrics the evaluation was done for the transformer models and the proposed work of TTEFM.

Table 4. Evaluation metrics for transformer methods.

Method name	Formula	Description
Accuracy [13]	$\frac{TP+TN}{TP+TN+FP+FN}$	Finds the overall correctness of the model.
Precision [4]	$\frac{TP}{(TP+FP)}$	Finds the proportion of true positive out of all positive predictions
Recall (or) Sensitivity [31]	$\frac{TP}{(TP+FN)}$	The model finds the correctly identified positive samples
F1- Score [24]	$\frac{2*(Precision*Recall)}{(Precision+Recall)}$	Combines both precision and recall and stores them as a single value
Balanced Accuracy [23]	$\frac{Sensitivity+Specificity}{2}$	It will give equal weight priority to both minority and majority classes. This measure is suitable for imbalanced datasets.
Specificity (or) true negative rate [7]	$\frac{TN}{TN+FP}$	TNR measures the model's ability to identify the negative samples correctly.
Geometric Mean [33]	$\sqrt{Sensitivity*Specificity}$	It finds the balance between sensitivity and specificity and is useful for imbalanced datasets.
False Positive Rate (FPR)	$\frac{FP}{FP+TN}$	FPR measures the proportion of negative instances incorrectly classified as positive.

4.2.2. Result Analysis

Based on the above metrics, different transformation models were applied to both the original and augmented images.

The 5-fold cross-validation method was used for the training and testing phase. The mean value is used to find each transformation model's training and testing accuracy in each fold. This value is then used as the

accuracy for each model, and the values are shown in Table 5. This table clearly states that the training phase of the original images performs well. However, in general, it provides less accuracy compared to training. Introducing new images led to a further decline in the model's performance, likely due to the limited size of the training dataset.

Image augmentation techniques were employed as a key aspect to address this issue and improve generalisation. By increasing the count of the images in the training phases, which means different directions in angle and the dimension of the images and other transformations are introduced to a diverse dataset during training.

The augmented images now perform better in terms of generalisation. So, augmentation techniques were useful when we had small datasets.

Table 5. Training and testing accuracy of different transformer methods for both original and augmented images.

Evaluation/Methods	Original images			Augmented images		
	DEiT	Mobile_ViT	Vision transformer	DEiT	Mobile_ViT	Vision transformer
Fold 1	100	99.51	100	99.81	99.11	99.39
Fold 2	100	98.38	97.08	100	99.34	98.36
Fold 3	100	99.03	99.68	100	99.44	99.67
Fold 4	100	98.54	99.68	100	99.44	99.67
Fold 5	100	98.86	100	100	99.67	99.25
Average (training)	100	98.86	99.29	99.96	99.40	99.27
Average (testing)	92.73	91.95	91.04	98.46	98.24	95.92
Standard deviation				0.0008	0.0020	0.0054
Confidence level				±0.0007	±0.0018	±0.0047

Table 6. Comparative analysis of different transformers with their measures.

Methods	Accuracy	Precision	Recall	F1-score	Balanced accuracy	Specificity	Geometric mean
ViT	95.92	95.89	95.54	95.69	95.54	98.62	97.06
DEiT	98.46	98.41	98.41	98.41	98.41	99.49	98.95
Mobile_ViT	98.24	98.20	98.12	98.16	98.12	99.41	98.76
TTEFM	**99.03**	**98.98**	**98.99**	**98.98**	**98.99**	**99.68**	**99.33**

Table 6 shows a comparative analysis of different transformation models with their respective values. The tale contains accuracy, precision, recall, F1-score, balanced accuracy, specificity, and the geometric mean value. Remarkably, within 10 epochs, the model produces a strong result for the benchmark method.

The ViT method attains 95.92 for accuracy, 95.89 for precision, 95.54 for recall, 95.69 for F1-score, 95.54 for balanced accuracy, 98.62 for specificity, and 97.06 for the geometric mean, respectively. The DEiT method yields superior results, with scores of 98.46, 98.41, 98.41, 98.41, 99.49, and 98.95, respectively. The Mobile_ViT method has 98.24, 98.20, 98.12, 98.16, 98.12, 99.41, and 98.76, respectively.

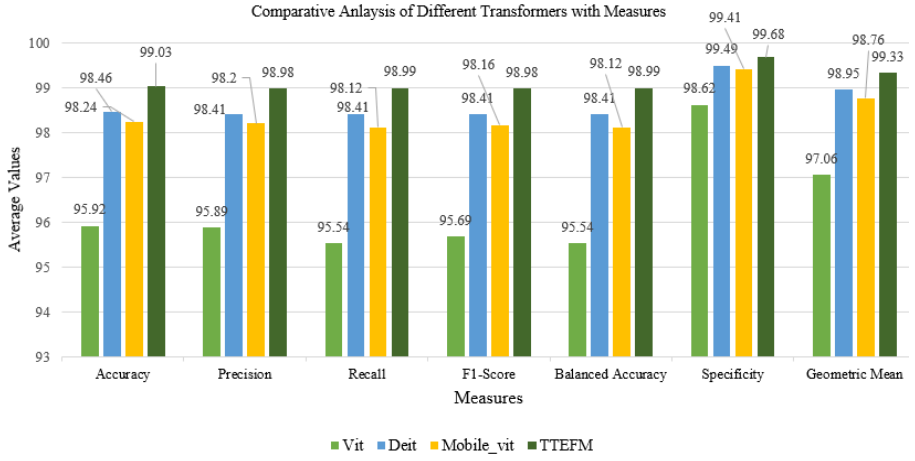

Figure 7. Comparative analysis of different transformer models with measures.

When compared to the other benchmark methods, the ensemble method, known as the TTEFM method, yields superior results. The metric values for accuracy, precision, recall, F1-score, balanced accuracy, specificity, and geometric mean are 99.03%, 98.98%, 98.99%, 98.98%, 98.98%, 99.68%, and 99.33%, respectively. Figure 7 also illustrates the result, providing a clear pictorial presentation of the comparative performance.

4.2.3. Discussion and Comparison

The proposed method, TTEFM, is an ensemble model that combines the strengths of ViT, DEiT, and Mobile_ViT transformer methods to achieve superior class-wise classification performance. The core objectives of each method are outlined below:

- ViT: ViT treats images as sequences of patches, replacing convolutional layers with a self-attention

mechanism to capture global context features and the relationships between patches. It is particularly effective for datasets with large dependencies.

- DEiT: DEiT focuses on enabling efficient training of ViTs on smaller datasets. Techniques like knowledge distillation and optimized training strategies make it competitive with CNNs in terms of accuracy and resource efficiency, even with limited data.
- Mobile_ViT: designed for resource-constrained environments such as mobile and edge devices, Mobile_ViT combines the strengths of transformers and convolutions to deliver high efficiency and accuracy on small datasets.

By combining these three methods, TTEFM exploits the individual strengths of these methods to improve generalization across classes. The ensemble method employs hard voting strategies that ensure equal importance for every class, contributing to robust performance on diverse datasets. Though ViT attains high accuracy by capturing global features, the precision, recall, and F1-score values vary across classes, as shown in Figure 8. This indicates that class-wise performance is not consistent. Compared to ViT, DEiT generalises better on new images, especially when working with smaller datasets, as shown in Figure 9. This arises from its primary focus on training efficiency.

Classification Report:

	precision	recall	f1-score	support
Chickenpox	0.96	0.90	0.93	508
Measles	0.96	0.98	0.97	591
Monkeypox	0.96	0.97	0.96	779
Normal	0.96	0.97	0.97	792
accuracy			0.96	2670
macro avg	0.96	0.96	0.96	2670
weighted avg	0.96	0.96	0.96	2670

Figure 8. Classification report of the ViT method.

Classification Report:

	precision	recall	f1-score	support
Chickenpox	0.97	0.98	0.98	508
Measles	0.99	0.98	0.99	591
Monkeypox	0.98	0.98	0.98	779
Normal	0.99	0.99	0.99	792
accuracy			0.98	2670
macro avg	0.98	0.98	0.98	2670
weighted avg	0.98	0.98	0.98	2670

Figure 9. Classification report of DEiT method.

Mobile_ViT shows excellent performance in resource-constrained environments, achieving good accuracy while maintaining efficiency. The classification report in Figure 10 highlights its effectiveness. The TTEFM model proposed here surpasses these individual methods by showing consistently high accuracy for all classes. It places equal importance on each class and provides comprehensive generalisation and reduced variability in the performance metrics. The classification report of TTEFM, as shown in Figure 11, is well-balanced and superior. Combining the goals of ViT, DEiT, and Mobile_ViT in a TTEFM ensemble model enables better general performance for a class-wise classification task.

Classification Report:

	precision	recall	f1-score	support
Chickenpox	0.97	0.96	0.97	508
Measles	0.99	0.99	0.99	591
Monkeypox	0.97	0.98	0.98	779
Normal	0.99	0.99	0.99	792
accuracy			0.98	2670
macro avg	0.98	0.98	0.98	2670
weighted avg	0.98	0.98	0.98	2670

Figure 10. Classification report of Mobile_ViT method.

Classification Report:

	precision	recall	f1-score	support
0	0.98	0.99	0.99	508
1	0.99	0.99	0.99	591
2	0.99	0.99	0.99	779
3	0.99	0.99	0.99	792
accuracy			0.99	2670
macro avg	0.99	0.99	0.99	2670
weighted avg	0.99	0.99	0.99	2670

Figure 11. Classification report of TTEFM method.

The batch size determines the quantity of samples handled during one forward and backwards pass. A reasonable batch size strikes a compromise between optimal memory use and minimising noise in photos. So, the normal batch size for the model is taken as 32. An epoch refers to the total count of complete iterations across the training dataset. The number of complete passes through the training datasets is called the epochs. If the training period is too long, the model will move to overfit, and the generalization of the image will become difficult. Based on the epoch and the learning rate, the model will face the difficulties. The learning rate will control the speed of training. The step size for updating the model weights during optimization. The parameters used in the transformation models are discussed in Table 7.

Table 7. Hyperparameters for transformation models.

Parameter name	Values
Epoch	10
K-fold	5-fold
Learning rate	0.0001
Optimizer	Adam
Loss	Cross Entropy loss
Batch size	32

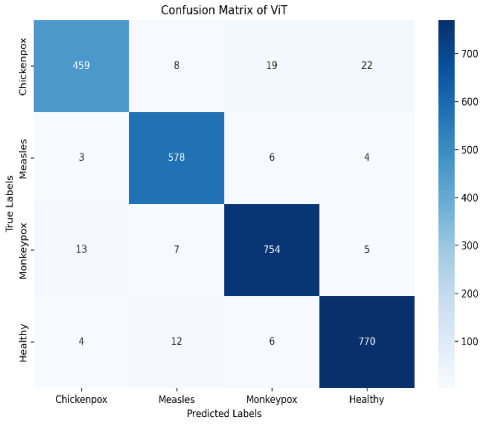

Figure 12. Confusion matrix of the ViT method.

The confusion matrix says how the transformation model works in each class. A combination of true positive, true negative, false positive, and false negative values is obtained. Each parameter has its characteristics. Figure 12 describes the ViT method of the confusion matrix. The total true values are 2561, and the false values are 109, respectively. Still, the models contain some amount of misclassification data. The DEit methods have the total true values as 2629 and the total false values as 41.

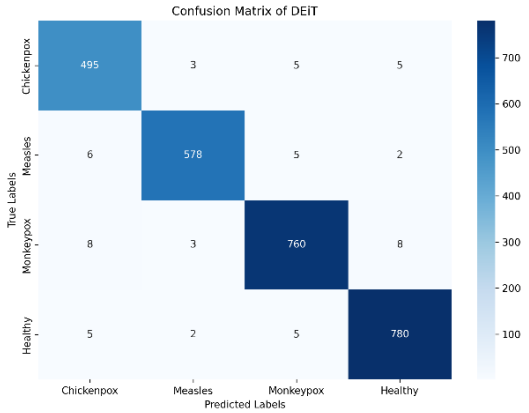

Figure 13. Confusion matrix of DEiT method.

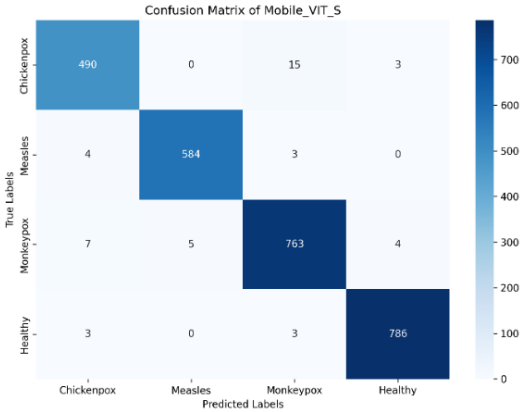

Figure 14. Confusion matrix of Mobile_ViT method.

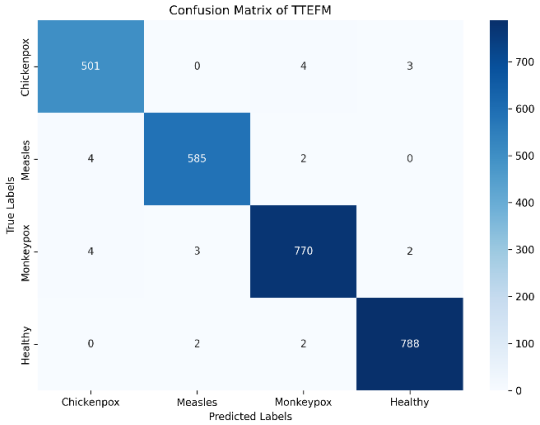

Figure 15. Confusion matrix of TTEFM method.

Figure 13 explains how the DEiT method, in comparison to the ViT model, generates a lower false rate and more true values. The Mobile_ViT has the true values as 2623 and the total false values as 47, which is explained in Figure 14. The TTEFM method has a total value of 2691, and the false values are 26. The proposed works outperform all state-of-the-art methods, ensuring that each class receives equal importance. Figure 15 expresses the TTEFM confusion matrix.

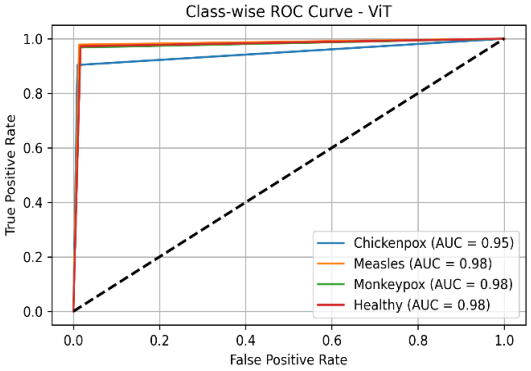

Figure 16. Class-wise ROC curve for ViT model.

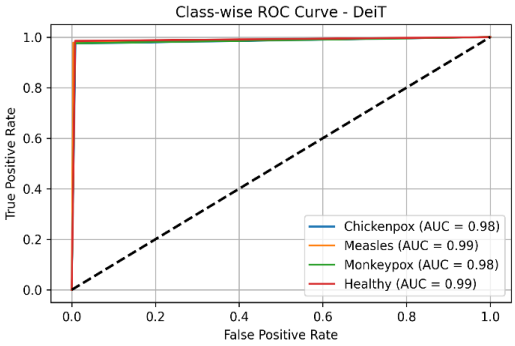

Figure 17. Class-wise ROC curve for DEiT model

The ROC curve can be derived from the confusion matrix, which is essential for determining the FPR and assessing model performance. The distribution of the ROC curve by class is presented herein. Figure 16 illustrates the ViT model, highlighting the true predictions of each model and their interactions with one another. The predicted probabilities for chickenpox, measles, monkeypox, and healthy individuals were 0.95, 0.98, 0.98, and 0.98, respectively. The model demonstrated limitations in predicting the chickenpox virus. The DEiT transformer method determines all the classes to be equal. Figure 17 illustrates that the AUC values for chickenpox, measles, monkeypox, and normal were 0.99.

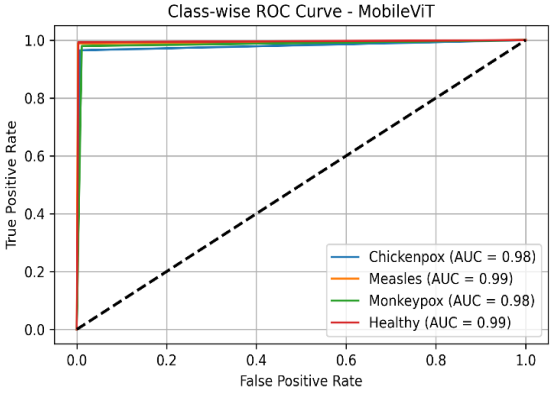

Figure 18. Class-wise ROC curve for Mobil_ViT model.

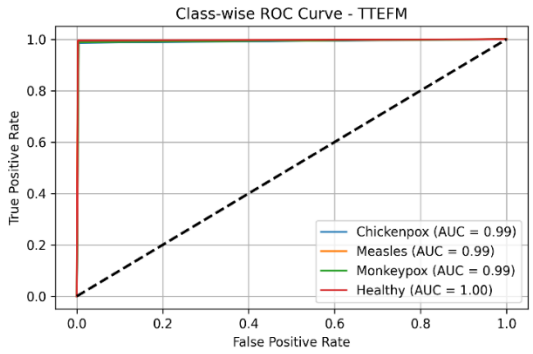

Figure 19. Class-wise ROC curve for TTEFM model.

Figures 18 and 19 illustrate the Mobile_ViT model and the proposed TTEFM models, which use a class-wise approach. The Mobile_ViT models achieve an accuracy of 0.98 for both chickenpox and monkeypox. For Healthy and Measles, it achieves a true positive value of 0.99. The TTEFM method achieves a score of 0.99 for chickenpox, measles, and monkeypox. The true value of 1.00 is achieved as healthy.

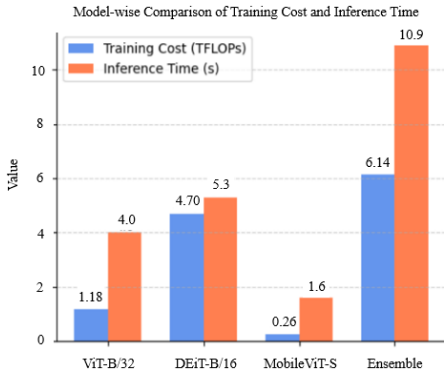

Figure 20. Model-wise comparison of training cost and inference time.

Figure 20 explains the training cost and the inference time of individual models along with the proposed method of TTEFM. Since the model has a high inference time but still the accuracy and the F1 score of the model were high when compared with individual methods. Based on the accuracy and the F1-score, the additional computational cost, especially in diagnostics of the viral infection, is necessary.

4.2.4. Ablation Analysis

In the beginning, the model overfitted and underfitted at epoch 5. The cross-validation progressively increased the epoch count to correct those issues.

DEiT and ViT with Mobile_ViT_S employed the stringent voting ensemble approach, which resulted in lower accuracy, although the individual model accuracy can exceed 98%. The deterioration is mostly caused by comparable structures and overlapping prediction properties in paired systems. The ensemble cannot utilize model diversity, which is essential for effective hard voting.

Although the ViT+DEiT combination had 97% accuracy, the ensemble could not guarantee performance increases in all class forecasts. The classification reports revealed class-wise balancing difficulties, with good predictions for measles and chickenpox but often incorrect alarms for monkeypox and vice versa. A consistent pattern in all examined ensembles suggests that models tend to complement each other in deficiencies due to structural and predictive similarities.

To prove that ensemble methods differ from other benchmark methods, the one-way ANOVA statistical method was used. Before testing the model, a hypothesis has been created.

$$H_o: \mu_{model} = \mu_{TTEFM} \quad (6)$$

$$H_1: \mu_{model} \neq \mu_{TTEFM} \quad (7)$$

In Equation (6), Ho indicates that the mean accuracy of all three models is equal to TTEFM; then the model has no significant difference when compared with the (α=0.05) value. In Equation (7), H_1 indicates that the mean of the model accuracy is not equal to TTEFM; then it significantly differs from the model. The p-value was computed for three models vs. TTEFM. The ViT vs. TTEFM model attains a p-value of 0.0164. The DEiT vs. TTEFM attains 0.0245, and the Mobile_ViT_S vs. TTEFM attains 0.0188. All three values of P indicate that the models are significantly different from one another, and this assertion was statistically confirmed.

5. Conclusions and Future Direction

In this paper, the introduction of TTEFM, which is an ensemble model combining the strengths of three variations of the Transformer: ViT, DEiT, and Mobile_ViT. ViT extracts global context features, DEiT performs well on smaller datasets, and Mobile_ViT is designed for lightweight systems. By ensembling these models, TTEFM provides more balanced and robust predictions across all classes, as verified in our class-wise performance. This balance enhances stability and reduces bias in imbalanced or multi-class datasets.

However, promising its performance, TTEFM has some drawbacks. The ensemble framework boosts computational complexity, which translates to increased inference time and memory usage relative to single models. Such trade-offs can restrict the model's usability in real-time or low-resource clinical settings, such as mobile health systems or rural clinics.

Future research will aim to minimize the computational burden of TTEFM by compressing, pruning, or distilling the model to facilitate deployment on edge devices. Additionally, integrating eXplainable AI (XAI) methods can increase clinical trust by increasing transparency in the decision process. Investigating adaptive inference methods or light ensembling may also ensure that TTEFM is made suitable for real-time diagnosis and point-of-care applications. Additionally, the cross-dataset validation

will be done in future. Based on this model, real-time handheld devices can be developed for identifying viral infections in low-resource clinical settings.

Reference

[1] ALkahla L., Saeed J., and Hussein M., "Empowering Ovarian Cancer Subtype Classification with Parallel Swin Transformers and WSI Imaging," *The International Arab Journal of Information Technology*, vol. 21, no. 6, pp. 1006-1014, 2024. https://doi.org/10.34028/iajit/21/6/5

[2] Bala D., Hossain M., Hossain M., Abdullah M., and et al., "MonkeyNet: A Robust Deep Convolutional Neural Network for Monkeypox Disease Detection and Classification," *Neural Networks*, vol. 161, pp. 757-775, 2023. https://doi.org/10.1016/j.neunet.2023.02.022

[3] Bala D., Kaggle, Monkeypox Skin Images Dataset (MSID), https://doi.org/10.34740/KAGGLE/DSV/3971903, Last Visited, 2025.

[4] Ballabio D., Grisoni F., and Todeschini R., "Multivariate Comparison of Classification Performance Measures," *Chemometrics and Intelligent Laboratory Systems*, vol. 174, pp. 33-44, 2018. https://doi.org/10.1016/j.chemolab.2017.12.004

[5] Bunge E., Hoet B., Chen L., Lienert F., and et al., "The Changing Epidemiology of Human Monkeypox a Potential Threat? A Systematic Review," *PLoS Neglected Tropical Diseases*, vol. 16, no. 2, pp. 1-20, 2022. https://doi.org/10.1371/journal.pntd.0010141

[6] Cao K., Tao H., Wang Z., and Jin X., "MSM-ViT: A Multi-Scale MobileViT for Pulmonary Nodule Classification Using CT Images," *Journal of X-Ray Science and Technology*, vol. 31, no. 4, pp. 731-744, 2023. https://pubmed.ncbi.nlm.nih.gov/37125604/

[7] Demsar J., "Statistical Comparisons of Classifiers over Multiple Data Sets," *Journal of Machine Learning Research*, vol. 7, pp. 1-30, 2006. https://www.jmlr.org/papers/volume7/demsar06a/demsar06a.pdf

[8] Din N., Zhang L., Nawaz M., and Yang Y., "Multi-Model Feature Aggregation for Classification of Laser Welding Images with Vision Transformer," *Journal of King Saud University-Computer and Information Sciences*, vol. 36, no. 5, pp. 102049, 2024. https://doi.org/10.1016/j.jksuci.2024.102049

[9] Ding Y. and Yang W., "Classification of Apple Leaf Diseases Based on MobileViT Transfer Learning," *in Proceedings of the International Conference on Image Processing and Artificial Intelligence*, Suzhou, pp. 1-7, 2024. https://doi.org/10.1117/12.3035225

[10] Dosovitskiy A., Beyer L., Kolesnikov A., Weissenborn D., and et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," *arXiv Preprint*, vol. arXiv:2010.11929v2, pp. 1-22, 2021. https://arxiv.org/abs/2010.11929

[11] Falendysz E., Lopera J., Rocke T., and Osorio J., "Monkeypox Virus in Animals: Current Knowledge of Viral Transmission and Pathogenesis in Wild Animal Reservoirs and Captive Animal Models," *Viruses*, vol. 15, no. 4, pp. 1-17, 2023. https://doi.org/10.3390/v15040905

[12] Haripriya K. and Inbarani H., "Hybrid FuzzyPCA-VGG16 Framework for Classifying Pox Virus Images," *The International Arab Journal of Information Technology*, vol. 22, no. 3, pp. 614-626, 2025. https://doi.org/10.34028/iajit/22/3/14

[13] Haripriya K. and Inbarani H., "Performance Analysis of Machine Learning Classification Approaches for Monkeypox Disease Prediction," *in Proceedings of the 6th International Conference on Electronics, Communication and Aerospace Technology*, Coimbatore, pp. 1045-1050, 2022. https://ieeexplore.ieee.org/document/10009407

[14] Hussain S., Songhua X., Aslam M., Waqas M., and Hussain F., "Hypergraph Convolutional Neural Networks for Clinical Diagnosis of Monkeypox Infections Using Skin Virological Images," *Applied Soft Computing*, vol. 170, pp. 112673, 2025. https://doi.org/10.1016/j.asoc.2024.112673

[15] Jerbi F., Aboudi N., and Khlifa N., "Automatic Classification of Ultrasound Thyroids Images Using Vision Transformers and Generative Adversarial Networks," *Scientific African*, vol. 20, pp. 1-14, 2023. https://doi.org/10.1016/j.sciaf.2023.e01679

[16] Khan S., Naseer M., Hayat M., Zamir S., and et al., "Transformers in Vision: A Survey," *ACM Computing Surveys*, vol. 54, no. 10s, pp. 1-41, 2022. https://doi.org/10.1145/3505244

[17] Luong H., Nguyen H., Le N., Le H., and et al., "A Proposed Approach for Monkeypox Classification," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 8, pp. 643-651, 2023. DOI: 10.14569/IJACSA.2023.0140871

[18] Malek A., Islam N., and Hoque A., "Investigations of Transmission Dynamics of Nipah Virus in Bangladesh," *Informatics in Medicine Unlocked*, vol. 44, pp. 101417, 2023. https://doi.org/10.1016/j.imu.2023.101417

[19] Maqsood S., Damasevicius R., Shahid S., and Forkert N., "MOX-NET: Multi-Stage Deep Hybrid Feature Fusion and Selection Framework for Monkeypox Classification," *Expert Systems*

with Applications, vol. 255, pp. 124584, 2024. https://doi.org/10.1016/j.eswa.2024.124584

[20] Mehta S. and Rastegari M., "MobileViT: Light-Weight, General-Purpose, and Mobile-Friendly Vision Transformer," *arXiv Preprint*, vol. arXiv:2110.02178v2, pp. 1-26, 2022. https://arxiv.org/abs/2110.02178

[21] Muthulakshmi A., Prasad C., Balachandran G., and Ranjith S., "Optimised Global Aware Siamese Network Based Monkeypox Disease Classification Using Skin Images," *Biomedical Signal Processing and Control*, vol. 101, pp. 107125, 2024. https://doi.org/10.1016/j.bspc.2024.107125

[22] Nayak T., Chadaga K., Sampathila N., Mayrose H., and et al., "Deep Learning Based Detection of Monkeypox Virus Using Skin Lesion Images," *Medicine in Novel Technology and Devices*, vol. 18, pp. 100243, 2023. https://doi.org/10.1016/j.medntd.2023.100243

[23] Nivetha S. and Inbarani H., "Neighborhood Rough Neural Network Approach for COVID-19 Image Classification," *Neural Processing Letters*, vol. 54, no. 3, pp. 1919-1941, 2022. https://link.springer.com/article/10.1007/s11063-021-10712-6

[24] Pereira R., Plastino A., Zadrozny B., and Merschmann L., "Correlation Analysis of Performance Measures for Multi-Label Classification," *Information Processing and Management*, vol. 54, no. 3, pp. 359-369, 2018. https://doi.org/10.1016/j.ipm.2018.01.002

[25] Sahin V., Oztel I., and Oztel G., "Human Monkeypox Classification from Skin Lesion Images with Deep Pre-Trained Network Using the Mobile Application," *Journal of Medical Systems*, vol. 46, no. 11, pp. 79, 2022. https://doi.org/10.1007/s10916-022-01863-7

[26] Salauddin M., Zheng Q., Murtuza M., Zheng C., and Hossain M., "Monkeypox Immunity: A Landscape of Host-Virus Interactions, Vaccination Strategies, and Future Research Horizons," *Animals and Zoonoses*, vol.1, no. 1, pp. 104-111, 2025. https://doi.org/10.1016/j.azn.2025.02.002

[27] Saputra V., Devi M., Diana., and Kurniawan A., "Comparative Analysis of Convolutional Neural Networks and Vision Transformers for Dermatological Image Classification," *Procedia Computer Science*, vol. 245, pp. 879-888, 2024. https://doi.org/10.1016/j.procs.2024.10.315

[28] Shankar A., Perumal P., Subramanian M., Ramu N., and et al., "An Intelligent Recommendation System in E-Commerce Using Ensemble Learning," *Multimedia Tools and Applications*, vol. 83, pp. 48521-48537, 2024. https://doi.org/10.1007/s11042-023-17415-1

[29] Sharma D. and Selwal A., "Cascading Adaptive Binary Image Feature Maps with Vision Transformer for Iris Spoof Detection," *Applied Soft Computing*, vol. 170, pp. 112713, 2025. https://doi.org/10.1016/j.asoc.2025.112713

[30] Sitaula C. and Shahi T., "Monkeypox Virus Detection Using Pre-Trained Deep Learning-based Approaches," *Journal of Medical Systems*, vol. 46, pp. 78, 2022. https://link.springer.com/article/10.1007/s10916-022-01868-2

[31] Sokolova M. and Lapalme G., "A Systematic Analysis of Performance Measures for Classification Tasks," *Information Processing and Management*, vol. 45, no. 4, pp. 427-437, 2009. https://doi.org/10.1016/j.ipm.2009.03.002

[32] Tchakoucht T., Elkari B., Chaibi Y., and Kousksou T., "Random Forest with Feature Selection and K-Fold Cross Validation for Predicting the Electrical and Thermal Efficiencies of Air Based Photovoltaic-Thermal Systems," *Energy Reports*, vol. 12, pp. 988-999, 2024. https://doi.org/10.1016/j.egyr.2024.07.002

[33] Tharwat A., "Classification Assessment Methods," *Applied Computing and Informatics*, vol. 17, no. 1, pp. 168-192, 2021. https://doi.org/10.1016/j.aci.2018.08.003

[34] Touvron H., Cord M., Douze M., Massa F., and et al., "Training Data-Efficient Image Transformers and Distillation through Attention," *arXiv Preprint*, vol. arXiv:2012.12877v2, pp. 1-22, 2021. https://arxiv.org/abs/2012.12877

[35] Ulukaya S. and Deari S., "A Robust Vision Transformer-based Approach for Classification of Labelled Rices in the Wild," *Computers and Electronics in Agriculture*, vol. 231, pp. 109950, 2025. https://doi.org/10.1016/j.compag.2025.109950

[36] Unalan S., Gunay O., Akkurt I., Gunoglu K., and Tekin H., "A Comparative Study on Breast Cancer Classification with Stratified Shuffle Split and K-Fold Cross Validation via Ensembled Machine Learning," *Journal of Radiation Research and Applied Sciences*, vol. 17, no. 4, pp. 101080, 2024. https://doi.org/10.1016/j.jrras.2024.101080

[37] Yadav S. and Qidwai T., "Machine Learning-based Monkeypox Virus Image Prognosis with Feature Selection and Advanced Statistical Loss Function," *Medicine in Microecology*, vol. 19, pp. 100098, 2024. https://doi.org/10.1016/j.medmic.2024.100098

[38] Zhu C., Wang W., Lu K., Zhang J., and et al., "Corneal Ulcer Automatic Classification Network Based on Improved Mobile ViT," *in Proceedings of the 19th International Conference on Advanced Intelligent Computing Technology and Applications*, Zhengzhou, pp. 614-625, 2023. https://doi.org/10.1007/978-981-99-4742-3_51

K.P. Haripriya received her MCA from Kumaraguru College of Technology, Anna University, Coimbatore, India, in 2018. She is currently pursuing a Ph.D. as a Research Scholar in the Department of Computer Science at Periyar University, Salem, Tamil Nadu, India. She qualified for the UGC NET-JRF in June 2023. Her research interests include Image Processing, Machine Learning, and Deep Learning.

H. Hannah Inbarani received her M.Sc. from Bharathidasan University, Trichy, India, in 1991, her M.Phil. from M.S. University, Tirunelveli, India, in 2003, and her MTech. degree from AAI University, Allahabad, India, in 2006, and her Ph.D. from Periyar University, Salem, India, in 2012. She is a Professor of Computer Science at Periyar University, Salem, Tamil Nadu, India. She completed a UGC major research project in 2016 and mentored a DST-NASI project, which was completed in 2015. She has received five best research awards at various regional, national, and international conferences. She has authored and co-authored more than 110 papers in international journals. Her research interests include Machine Learning, Deep Learning, Image Processing, Signal Processing, and Bioinformatics. She is the corresponding author for this paper.