Minimax Theory and its Applications

Volume 11 (2026), No. 2s, 849 to 866

Hybrid Vision Mamba Transformer Network for Accurate Tomato Leaf Disease Classification

Rabinarayan Panda[1], Sachikanta Dash[2], Dharamvir[3], Mahaboob Basha M.A[4], Bishnu Paramguru Mahapatra[5]

[1]Department of CSE, Saptagiri NPS, Bengaluru, Karnataka, India-560057
[2]Department of CSE(Cyber Security), Madanapalle Institute of Technology and Science, Madanapalle, Andhrapradesh, India-517325
[3]Department of MCA,The Oxford College of Engineering, Bengaluru, India-560068
[4]Professor, Department of Computer Science and Engineering, Sapthagiri NPS University Bangalore, India.
[5]Department of English and Foreign language, Madanapalle Institute of Technology and Science, Madanapalle, Andhrapradesh, India-517325
Email: n.rabipanda2011@gmail.com[1], dash.sachikanta@gmail.com[2], dhiruniit@gmail.com[3], bashakhan60@gmail.com[4], drbishnupm@mits.ac.in[5]

Abstract
Timely identification of tomato leaf diseases can help farmers preserve yield and achieve better field management. Deep image models are able to address this problem. Convolutional neural networks perform great on local relations, while they fail to model relationships between distant regions of an image. On the other hand, Vision Transformers allow capturing the large context; however, they are quite costly in terms of computational expenses.In this study, we propose a Hybrid Vision Mamba Transformer Network, referred to as HVMTNet, to automate the tomato leaf disease detection process. The architecture utilizes various components. At first, a conv stem learns local features from an image. Next, a Vision Mamba encoder is applied to model long-range dependencies efficiently. Afterward, a Transformer encoder is used to learn broader scene context. Finally, adaptive feature fusion is performed to aggregate the learned cues and generate disease-specific features.HVMTNetis evaluated on the PlantVillage tomato leaf disease dataset. There are ten disease classes in it. The data are partitioned into subsets using 70:15:15 ratio. The proposed approach outperforms ResNet50, EfficientNet-B0, ConvNeXt-Tiny, Vision Transformer, and Vision Mamba models. The obtained results are as follows: 99.61% accuracy, 99.54% precision, 99.48% recall, 99.51% F1-score, and AUC 0.999.

Keywords: Tomato leaf disease classification, Vision Mamba, Vision Transformer, Deep learning, Precision agriculture, PlantVillage dataset, Hybrid neural network.

1. Introduction

The tomato (Solanumlycopersicum) is cultivated in various countries and plays an important role in the nutrition and agricultural economics. It is a major crop for many farmers. Nevertheless, tomato crops are commonly plagued by numerous diseases. They include fungal or bacterial infections as well as viral infections and plant disorders. Examples are Early Blight, Late Blight, Bacterial Spot, Leaf Mold, Septoria Leaf Spot, and Tomato Mosaic Virus[1].These diseases decrease the quality and quantity of the crop. Consequently, the farms may suffer economic loss and increase their dependence on chemicals. Therefore, it is crucial to detect tomato leaf diseases at an early stage. It allows the farmers to properly control the crop, reduce losses, and use sustainable agriculture practices[6].

Plant disease inspection is still highly dependent on the observations made by specialists in the field. If an experienced agricultural specialist is conducting this inspection, the outcome might turn out to be rather effective. However, such a technique requires plenty of time and efforts to produce the desired result[2]. Moreover, this process is rather subjective and not all farmers living far from each other can use it.At an early stage, various plant diseases may seem quite similar to each other. As a consequence, guessing them based on one's observations becomes rather challenging. On the other hand, modern farms tend to adopt precision farming techniques and other types of advanced equipment.

In recent years, development of AI algorithms including deep learning made it possible to significantly increase the effectiveness of the image model classification for detecting plant diseases. In particular, neural networks such as AlexNet, VGGNet, ResNet, DenseNet, EfficientNet, and ConvNeXt proved effective in detecting

important features in leaf images. They are capable of extracting features on different scales. Nevertheless, most CNN architectures have a local nature and cannot analyze information from afar in the image. It makes it challenging to consider remote information in case of very similar diseases.Also, one may use Vision Transformers[7]. They use self-attention and are thus capable of considering different elements of the image in terms of the entire leaf. Such architecture makes it possible to learn long-range connections. For this reason, Vision Transformers are able to provide excellent results in various vision-related problems including the detection of plants. However, such architectures require large labeled data and have considerable costs due to the quadratic nature of self-attention relative to the number of image patches.

Vision Mamba based on selective state space models has recently garnered much attention in the field of visual features learning. Vision Mamba is different from transformers in the sense that it processes sequences of images with linear cost. While at it, it is able to capture long-distance spatial connections.Given that, it is capable of attaining performances that are almost equivalent to the best performing approaches[3]. This is possible with minimal computational cost and room for future improvement. However, despite the efficient modeling of general sequential information, Vision Mamba may lack some of the benefits offered by attention based models.

In light of the above observations, the study introduces a hybrid vision model known as HVMTNet for tomato leaf disease detection. The architecture involves a combination of Vision Mamba with a transformer in order to allow the model to detect long dependencies and make use of the global information[8].Initially, a convolution block is employed to extract local cues from the input images of the leaf. This helps capture textural features, colors and edges. Then, Vision Mamba modules follow up to enhance feature interactions across distant regions of the image. A transformer encoder subsequently acts upon the learned features through multi-head self-attention in order to highlight the important dependencies[9]. At last, an adaptive fusion operation comes in between both streams before feeding a fully connected layer for classification.

The aim of this hybrid approach is improved classification of diseases when there are two similar classes. Another aim is to maintain high speed to be used in practice on farms. This approach utilizes Vision Mamba and Transformer methods. This allows the network to train image features better. Additionally, this leads to more robust classification under variations in light or noisy background conditions. Experiments show that the proposed model detects diseases better than deep learning benchmarks.

The major contributions of this work are summarized as follows:

- In particular, the proposed architecture is referred to as HVMTNet and is developed for tomato leaf disease detection purposes. The idea lies in the integration of scalable state space modeling along with the capability of the Transformer to analyze the entire image.
- Moreover, the method involves joint feature learning strategy aimed at capturing fine-grained local textures, large-scale spatial connections, and contextual information. The objective is to develop better features for disease identification.
- As to fusion of inputs, the method incorporates an adaptive fusion procedure where results obtained by the Vision Mamba architecture and Transformer architecture are combined together. This is useful for cases when disease classes are visually similar in images.
- Multiple experiments have demonstrated superiority of the proposed approach over several other top-performing models such as CNN, Transformer, and Mamba architectures. The improvement is evident in terms of accuracy, precision, recall, and F1-score metrics. The approach remains computationally efficient as well.

The rest of the paper is outlined as follows.In Section II, related work pertaining to the classification of diseases in tomato leaves is considered. The recent studies related to Vision Mamba and Transformers models are also discussed.Section III provides an overview of our proposed HVMTNet architecture. The details about its constituent parts are also provided there.Section IV discusses the experimental setup. In particular, the datasets, metrics used for the performance evaluation, and the implementation details are mentioned there.Section V describes our results and compares them to the existing approaches.Finally, the conclusions are drawn in Section VI.

2. Related Work

Tomato leaf disease automatic detection has become an increasingly popular topic due to the rapid advancements in AI and computer vision technology. Previously, many researches had used manually crafted features, but more recent works have used deep learning techniques which have the ability to extract informative patterns directly from the images of the leaves.In this paper, we provide a comprehensive description of the evolution of tomato disease detection techniques. Moreover, we review the more recent techniques using Vision Transformers and Vision Mambas. Then, we discuss what remains unexplored in this field of study. This gap forms the foundation for the development of the proposed HVMTNet technique, presented in Table 1.

Table 1. Category-wise Literature Review on Tomato Leaf Disease Classification

Author(s) & Year	**Category**	**Method / Model**	**Major Contribution**	**Limitation**
Arivazhagan et al. (2013)[1]	Conventional Machine Learning	GLCM + Color Features + SVM	Proposed texture and color feature extraction using GLCM with SVM for automatic plant leaf disease classification, achieving satisfactory performance under controlled conditions.	Performance depends on handcrafted features and is sensitive to illumination and background variations.
Camargo and Smith (2009)[2]	Conventional Machine Learning	HOG, SIFT + Random Forest	Applied image segmentation and handcrafted descriptors for identifying plant diseases using machine learning classifiers.	Limited robustness for complex field images and multiple disease classes.
Krizhevsky et al. (2012)[3]	CNN-Based Deep Learning	AlexNet	Introduced deep CNNs for automatic hierarchical feature learning, significantly improving image classification performance.	Large number of parameters and relatively high computational complexity.
He et al. (2016)[4]	CNN-Based Deep Learning	ResNet	Proposed residual learning to enable very deep CNNs with improved convergence and feature extraction capability.	Primarily captures local receptive field information and lacks explicit global dependency modeling.
Tan and Le (2019)[5]	CNN-Based Deep Learning	EfficientNet	Introduced compound model scaling that balances network depth, width, and resolution for improved efficiency and accuracy.	Long-range contextual information is not explicitly modeled.
Hughes and Salathé (2015)[10]	CNN-Based Plant Disease Classification	PlantVillage Dataset + CNN	Developed the PlantVillage dataset, providing a benchmark for deep learning-based plant disease diagnosis and facilitating high-accuracy classification models.	Images were captured under controlled laboratory conditions, limiting real-world generalization.
Dosovitskiy et al. (2021)[11]	Vision Transformer	Vision Transformer (ViT)	Introduced patch-based Transformer architecture using global self-attention for image recognition, outperforming many CNN models on large datasets.	Requires large-scale annotated datasets and has quadratic computational complexity.
Liu et al. (2021)[12]	Vision Transformer	Swin Transformer	Proposed hierarchical window-based self-attention to improve computational efficiency while preserving strong visual representation	Window-based attention may limit complete global interactions between distant image regions.

			learning.	
Touvron et al. (2021)[13]	Vision Transformer	DeiT	Improved data efficiency of Vision Transformers using knowledge distillation, reducing dependency on massive training datasets.	Computational requirements remain high for edge devices.
Zhu et al. (2024)[14]	Vision Mamba	Vision Mamba	Introduced selective state-space models for visual recognition with linear computational complexity and efficient long-range dependency modeling.	Fine-grained contextual relationships may not be as effectively captured as with Transformer attention.
Chen et al. (2022)[19]	Hybrid Deep Learning	Mobile-Former (CNN + Transformer)	Combined CNN local feature extraction with Transformer global attention to improve visual representation while maintaining efficiency.	Hybrid architecture increases model design complexity and training overhead.
Hatamizadeh et al. (2025)[20]	Hybrid Vision Mamba–Transformer	State Space Model + Transformer	Integrated state-space modeling with Transformer attention to improve representation learning and computational efficiency in vision tasks.	Primarily evaluated on medical imaging; application to agricultural disease classification remains limited.

2.1 Conventional Machine Learning Approaches

The early approaches used to detect tomato diseases were based on conventional image processing techniques along with the application of machine learning. Initially, image enhancement was followed by removal of background and partitioning of the image[15]. Then, some measurements were made based on the input image. Color cues, texture cues, and shape cues were among the common selections. Some of the commonly used features for feature extraction were HOG, LBP, GLCM, and SIFT. The learning algorithm then classified the tomato plant based on the manually extracted features. Common algorithms used were SVM, decision trees, random forests, k-NN, and artificial neural network.These techniques proved to be effective when tested in laboratory conditions. However, their results in the field were inconsistent. This happened due to the variability of light conditions, inconsistent background, variable angle of leaves, and variable degree of infection. In addition, the manually designed features failed to capture the more complicated characteristics occurring throughout various diseases[16].

2.2 Deep Learning-Based Tomato Disease Classification

In terms of detecting plant diseases, deep learning has brought changes to the way it works. Unlike manual feature generation, deep learning algorithms can extract relevant features automatically from the images.A good example of such deep learning algorithm is Convolutional Neural Networks[17]. The network is able to capture spatial information in an image that is very helpful in diagnostic applications.A number of different CNNs have been tried for tomato disease detection. For instance, AlexNet, VGGNet, GoogLeNet, ResNet, DenseNet, EfficientNet, MobileNet, and ConvNeXt. These networks increase the accuracy in classification. This happens through learning good representation from large image datasets, like PlantVillage.

The residual learning in ResNet solves the problem of vanishing gradients. This leads to ease in training of deep networks as well as obtaining of more features. DenseNet takes another path. The feature reuse is enhanced through a dense connection of layers. EfficientNet takes an approach of compound scaling, where adjustments of depth, width, and image resolution are done at once, thus reducing the computation. Some models such as MobileNet offer fast application of neural networks on edge devices, which is important for real-time analysis of crops when hardware limitations are faced[18].However, the CNN methods have one disadvantage. They focus on analyzing the local region with the help of receptive fields, which means that the modeling of connectivity between the distant regions on a leaf may be impossible. Disease symptoms may appear in the form of isolated spots. Their configuration may differ depending on the specific leaf. Dependence on convolution may lead to poor recognition of patterns.

2.3 Vision Transformer-Based Approaches

Transformer models performed excellently in linguistic tasks, which gave rise to vision transformers or ViT models, for the purpose of doing image classification. A ViT model does not make use of convolution. Instead, the image is broken down to tiny patches on which self-attention operation is applied to obtain global context from the whole image.Different architectures of Transformer models have been studied in the field of agriculture. DeiT, Swin Transformer, PVT, and CrossViT models have shown their capability to discover distant dependencies among patterns associated with plant diseases. In some of the updated versions of the architecture, hierarchical stages are adopted, which help in saving computational resources through localized attention while maintaining global context access[21].ViT models are better in learning global information compared to the Convolutional Neural Networks[22]. However, training ViT models is expensive. Also, large numbers of labeled images are required to train these models. Self attention is computationally expensive based on the square of the input size.

2.4 Vision Mamba for Image Classification

Vision Mamba has emerged as an alternative to the existing vision models. The architecture of this model is based on Selective State Space Models rather than transformers. As a result, Vision Mamba propagates data in terms of states, and not through an attention mechanism over all tokens. It helps keep the computational cost low for very long sequences. At the same time, Vision Mamba is efficient on classification tasks[23].Vision Mamba has shown its efficiency in a number of fields. Among the use cases of this model are image classification, semantic segmentation, medical image processing, and remote sensing. The main idea behind the efficiency of the model is the selective state mechanism. This model is still able to capture the context of a scene. In addition, it requires much less computation compared to attention models[24].The linear increase in computational cost allows using the model for large agricultural datasets. In addition, this model suits systems that require immediate responses in case of disease detection. Nevertheless, selective states do not allow capturing certain types of local interactions[25].

2.5 Hybrid Deep Learning Architectures

In recent times, many publications focus on exploring the strengths of multiple deep learning approaches through hybrid configurations. For example, one approach takes pieces of both CNN and Transformer architectures[26]. This is good due to the fact that CNN captures local patterns, whereas the Transformer captures a larger context. Another approach employs CNN with attention modules. In such designs, attention plays an important role in fine-tuning the features that allow for more discrimination of classes.Moreover, more recently, many publications discuss Vision Mamba with either CNN or Transformer blocks. Such architectures aim to balance the sequence-based processing approach with the ability to detect larger relations between image elements. The purpose of such architectures is to improve the prediction quality without paying much computational price. Many publications claim that such an approach results in high accuracy with efficiency[27].Nevertheless, despite the improvements provided by Vision Mamba with Transformers, fewer publications cover Vision Mamba with Transformer hybridization for tomato leaf disease detection task. All the existing approaches tend either to increase accuracy or to speed up the prediction process. They usually ignore the possibility of balancing both factors[28].

2.6 Research Gap

After number of review processes for different articles, we found the following research gaps:

- CNN-based models primarily learn local spatial features and often fail to capture long-range dependencies among disease regions.
- Vision Transformers provide excellent global contextual modeling but suffer from high computational complexity and large training data requirements.
- Vision Mamba offers efficient long-range sequence modeling with linear complexity but may not fully exploit attention-based contextual interactions.
- Existing tomato disease classification methods rarely integrate Vision Mamba and Transformer architectures into a unified framework.
- Many current models experience performance degradation under varying illumination, complex backgrounds, occlusion, and visually similar disease symptoms.

The constraints on us will require that we develop a hybrid model. This is meant to achieve both efficiency in sequence processing while maintaining focus throughout the input.

2.7 Motivation for the Proposed HVMTNet

To overcome the above limitations, this paper proposes aHVMTNet for accurate tomato leaf disease classification. The proposed framework combines:

- **A convolutional stem** to capture low-level texture, color, and edge information.
- **Vision Mamba blocks** to efficiently model long-range spatial dependencies with linear computational complexity.
- **Transformer encoder layers** to enhance global contextual understanding through multi-head self-attention.
- **Adaptive feature fusion** to integrate complementary representations from both branches before classification.

HVMTNet can perform well due to its ability to make use of local information, effective sequential modeling, and attention mechanism to get context. The main aim is for the method to have high classification accuracy, robustness, and low computation cost. This makes the method applicable for intelligent disease identification in precision agriculture. It also seeks to overcome the challenges faced by some CNN, Transformer, and Mamba approaches.

3. Proposed Hybrid Vision Mamba–Transformer Network (HVMTNet)

3.1 Overview of the Proposed Framework

The Hybrid Vision Mamba and Transformer Network, HVMTNet, is a model that seeks to achieve a highly accurate classification of tomato leaf disease. It incorporates the elements of the CNN, Vision Mamba, and Transformer. Each of the elements contributes its unique advantages. Convolution captures local features. Vision Mamba provides efficient sequence processing with low cost. The Transformer side captures the global context. It is designed as an end to end deep model. The model not only learns the local features but also establishes connections between distant cues. Hence, the model utilizes the context and local features during training.This model distinguishes itself from most CNN models. The CNN models usually focus on small regions using local receptive fields. Also, some Transformer models might be costly and slow to implement. The HVMTNet incorporates the Vision Mamba for efficient sequence modeling and self attention of Transformer to learn the global image relations.

The overall architecture of HVMTNet consists of five major components:

1. Image preprocessing and augmentation
2. Convolutional feature extraction module
3. Vision Mamba encoder
4. Transformer encoder
5. Adaptive feature fusion and classification head

The process, which is depicted in Fig. 1, begins with an image of a tomato leaf. At first, preprocessing is performed. Then, the image undergoes the data augmentation stage. Feature extraction is performed to extract the visual information from the image. Two streams are utilized in this case – Vision Mamba stream and Transformer stream, each of which extracts different perspectives of the image. After the features are extracted by each of these streams, they are merged into a single fused vector. At the end, the fused vector is fed to the classification head.

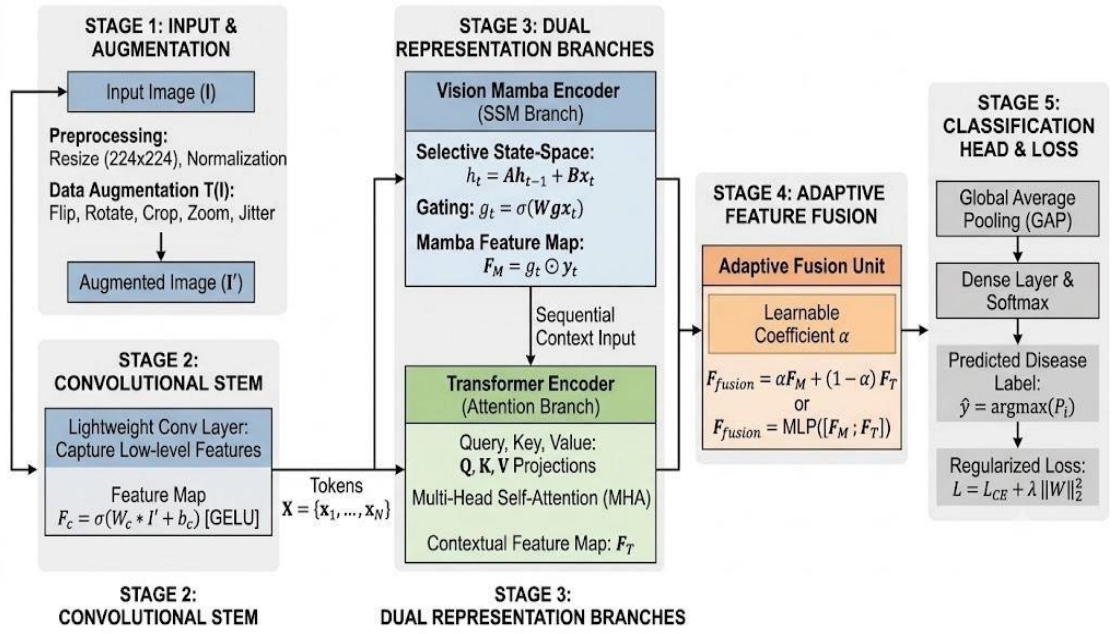


Fig. 1. Proposed model architecture

The complete algorithmof the proposed model in described in algorithm 1 as follows.

Algorithm 1: Hybrid Vision Mamba–Transformer Network

Input: Handwritten Marathi image I
Output: Predicted character sequence $\hat{Y}$

1: Resize and preprocess input image I
2: Apply data augmentation to obtain I_aug
3: F ← ResNet50(I_aug; θR)
4: X ← Convert F into sequential features {x1, x2, ...,xT}
5: for t = 1 to T do
6: hf_t ← LSTMf(xt, hf_(t−1))
7: hb_t ← LSTMb(xt, hb_(t+1))
8: ht ← [hf_t ;hb_t]
9: end for
10: Compute attention scores: et = v^T tanh(Whht + bh)
11: Compute attention weights: αt = exp(et) / Σjexp(ej)
12: Generate attention representation: c = Σt αt ht
13: Compute character probabilities: Pt = Softmax(Wcht + bc)
14: Calculate CTC loss: LCTC = −log P(Y | P1, ..., PT)
15: Update model parameters: θ ← θ − η∇θLCTC
16: Ŷ ← CTC_Decode(P1, ..., PT)
17: Return Ŷ

3.2 Image Preprocessing

Let the input tomato leaf image be represented aseq(1)

$$I \in RH \times W \times 3 I \in \mathbb{R}^{H \times W \times 3} \quad (1)$$

whereH and W denote the image height and width, respectively.

Each image is resized to 224×224

pixels and normalized to improve training stability. Data augmentation techniques are applied to improve the model's robustness against variations encountered in real agricultural environments.

The augmentation operations include

- Random horizontal flipping
- Random rotation
- Random cropping
- Brightness adjustment
- Contrast enhancement
- Random zoom
- Color jittering

The transformed image is represented as per eq(2)

$$I' = T(I) I' = T(I) \quad (2)$$

whereT(·) denotes the augmentation function.

3.3 Convolutional Feature Extraction Module

The first stage of HVMTNet employs a lightweight convolutional stem to capture low-level visual information such as

- edges,
- leaf veins,
- disease spots,
- texture,
- color variations.

The convolution operation is defined as in eq(3)

$$F_c = \sigma(W_c * I' + b_c) \quad (3)$$

where

- Wcrepresents convolution kernels,
- Bcdenotes bias,
- $*$ indicates convolution,
- σ(·) is the GELU activation function.

The output feature map is$F_c \in \mathbb{R}^{H' \times W' \times C}$

which is flattened into image tokens before entering the Vision Mamba encoder.

3.4 Vision Mamba Encoder

The Vision Mamba encoder is responsible for learning long-range spatial dependencies using **Selective State Space Models (SSMs)**. Unlike self-attention, Vision Mamba models sequential information with linear computational complexity.

The flattened image tokens are represented as in eq(4)

$$X = \{x_1, x_2, \dots, x_N\} \quad (4)$$

where

- NN is the number of image patches,

The state-space dynamics are formulated as in eq(5)

$$h_t = A h_{t-1} + B x_t \quad (5)$$

where

- ht denotes the hidden state,
- A represents the state transition matrix,
- B is the input projection matrix.

The output is computed as in eq(6)

$$y_t = Ch_t + Dx_t \quad (6)$$

where

- C projects hidden states,
- D is the skip connection.

The selective gating mechanism dynamically determines important information as in eq(7)

$$g_t = \sigma(W_g x_t) \quad (7)$$

The final Mamba representation becomes as per eq(8)

$$F_M = g_t \odot y_t \quad (8)$$

Where$\odot$ denotes element-wise multiplication.

The Vision Mamba encoder efficiently models disease regions distributed across different locations of the leaf.

3.5 Transformer Encoder

Although Vision Mamba captures long-range dependencies efficiently, Transformer attention provides stronger contextual interaction among image patches.

The Transformer encoder receives$F_M$as input.

Query, Key, and Value matrices are computed as in eq(9)

$$Q = F_M W_{Q,} K = F_M W_K, V = F_M W_V \quad (9)$$

The self-attention operation is in eq(10)

$$Attention(Q, K, V) = Softmax\left(\frac{QK^T}{\sqrt{d}}\right) V \quad (10)$$

where

- ddenotes feature dimension.

Multi-head attention is computed as in eq(11)

$$MHA = Concat(head_1, \dots, head_h) W_O \quad (11)$$

The Transformer output is$F_T$which emphasizes disease-specific contextual information.

3.6 Adaptive Feature Fusion

Vision Mamba and Transformer capture complementary information.

Instead of simple concatenation, an adaptive fusion mechanism is introduced.

The fused feature is computed as in eq(12)

$$F_{fusion} = \alpha F_M + (1 - \alpha) F_T \quad (12)$$

Where$0 \leq \alpha \leq 1$is a learnable fusion coefficient.

Alternatively,$F_{fusion} = MLP([F_M; F_T])$, where[;] denotes feature concatenation.

This fusion strategy allows the network to automatically learn the contribution of each branch.

3.7 Classification Head

The fused representation is passed through a fully connected layer.

Global Average Pooling (GAP) is first applied as ineq(13)

$$F_g = GAP(F_{fusion}) \quad (13)$$

The prediction logits areshown in eq(14)

$$z = W_f F_g + b_f \quad (14)$$

Finally, disease probabilities are computed using Softmax as in eq(15)

$$P_i = \frac{e^{z_i}}{\sum_{j=1}^{C} e^{z_j}} \quad (15)$$

where

- C denotes the total number of tomato disease categories.

The predicted disease label is shown in eq(16)

$$\hat{y} = \arg max(P_i) \quad (16)$$

3.8 Loss Function

The network is trained using the categorical cross-entropy loss as peeq(17)

$$L_{CE} = -\sum_{i=1}^{C} y_i \log(P_i) \quad (17)$$

where

- $y_i$is the ground-truth label,
- $P_i$is the predicted probability.

To reduce overfitting, L2 regularization is incorporatedineq(18)

$$L = L_{CE} + \lambda ||W||_2^2 \quad (18)$$

where

- λ denotes the regularization parameter.

4. Experimental Setup and Performance Evaluation

This section discusses the dataset employed for the tomato leaf diseases classification problem. In addition, the preprocessing techniques used on the data are discussed. Then, the implementation decisions are highlighted. This is followed by discussion of the performance metrics employed for the analysis. The baseline and comparative model used are then mentioned. Finally, the hardware environment used is described.

4.1 Dataset HVMT Net Description

The accuracy of the proposed system is evaluated on the PlantVillage Tomato Leaf Disease Dataset. This data set is widely used in plant disease classification. It contains well-photographed images of tomato leaves in RGB format. Some leaves in the data set are healthy, and some suffer from various diseases.There are ten classes in total, which can be seen in Figure 2. Among those, nine classes are different disease categories. The last one is healthy tomato leaves.

The PlantVillage tomato data subset contains about 18,160 RGB images. The images are characterized by various diseases. There are also differences in colors and shapes of lesions.In order to provide an unbiased evaluation, the data is divided into three portions. Around 70% of data is used for training. 15% of data is used for validation. Another 15% is used for testing.Each of the portions contains equal proportions of each class.

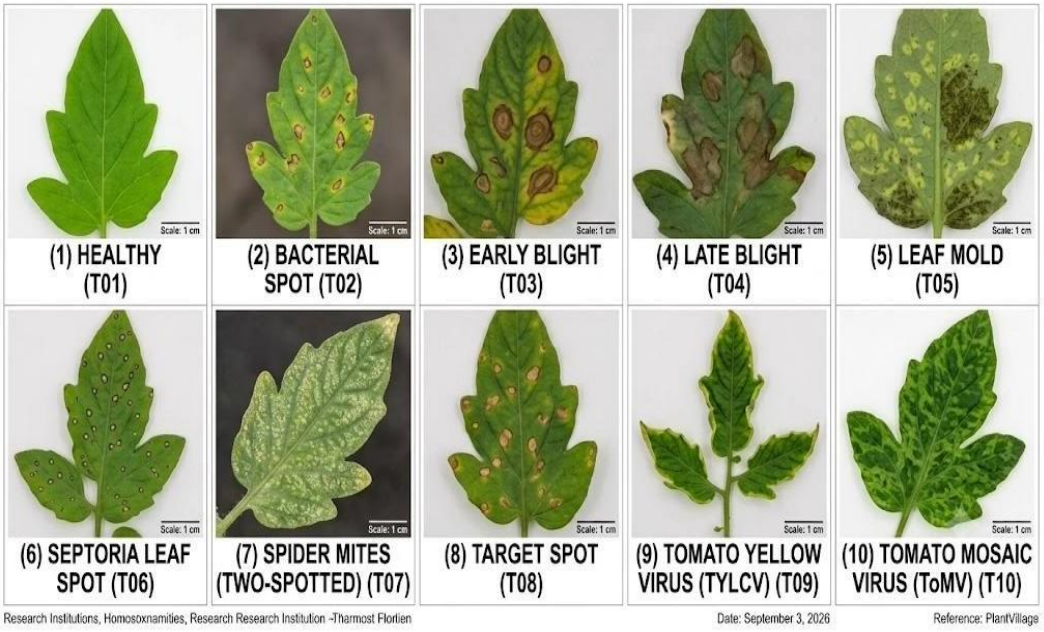


Fig. 2. Disease category of tamato leaf

4.2 Data Preprocessing and Augmentation

Deep learning models are highly sensitive to image quality and diversity. Therefore, all images undergo preprocessing before training.

The preprocessing pipeline includes:

- Image resizing to 224 × 224 pixels
- Pixel value normalization
- RGB color standardization
- Random horizontal flip
- Random vertical flip
- Random rotation (±20°)
- Random zoom
- Random cropping
- Brightness adjustment
- Contrast enhancement
- Color jittering

These augmentation techniques improve model generalization and reduce overfitting by exposing the network to diverse visual variations encountered in real agricultural environments.

4.3 Implementation Details

The proposed HVMTNetis implemented using the PyTorch deep learning framework.

The model employs:

- Convolutional Stem
- Vision Mamba Encoder
- Transformer Encoder
- Adaptive Feature Fusion

- Fully Connected Classification Head

The network is trained using the AdamW optimizer with weight decay regularization.

The learning rate is initialized to$1 \times 10^{-4}$and reduced dynamically using a cosine annealing learning rate scheduler.

The major hyperparametersare summarized in Table 2.

Table 2. Training Hyperparameters

Parameter	**Value**
Input image size	224 × 224
Batch size	32
Epochs	100
Optimizer	AdamW
Initial learning rate	0.0001
Weight decay	0.0001
Dropout	0.3
Activation	GELU
Loss function	Cross-Entropy Loss
Scheduler	Cosine Annealing
Early stopping patience	15 epochs

The model parameters are initialized using Xavier initialization to facilitate stable convergence during training.

4.4 Performance Evaluation Metrics

The proposed model is evaluated using standard classification metrics.

Accuracy:

Overall classification accuracy is computed as

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (19)$$

where

- TP = True Positive
- TN = True Negative
- FP = False Positive
- FN = False Negative

Precision:

$$Precision = \frac{TP}{TP+FP} \quad (20)$$

Precision measures the proportion of correctly predicted disease samples among all predicted samples.

Recall:

$$Recall = \frac{TP}{TP+FN} \quad (21)$$

Recall evaluates the capability of the model to correctly identify diseased leaves.

F1-Score:

$$F1 = \frac{2 \times Precision \times Recall}{Precision+Recall} \quad (22)$$

The F1-score balances precision and recall.

Specificity:

$$Specificity = \frac{TN}{TN+FP} \quad (23)$$

Specificity measures the ability to correctly classify healthy samples.

Matthews Correlation Coefficient (MCC):

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \quad (24)$$

MCC provides a balanced evaluation for multi-class classification.

Area Under ROC Curve (AUC):

A ROC curve is constructed for each category of diseases. The value of AUC is then calculated. It is done to evaluate the performance of the classification model.

4.5 Comparative Models

To demonstrate the effectiveness of HVMTNet, comparisons are conducted with several state-of-the-art deep learning models. The comparative models include the following as described in table 3.

Table 3: Descriptions of different models

Model	**Architecture**
AlexNet	CNN
VGG16	CNN
ResNet50	Residual CNN
DenseNet121	Dense CNN
EfficientNet-B0	Compound CNN
ConvNeXt-Tiny	Modern CNN
Vision Transformer (ViT-B/16)	Transformer
Swin Transformer	Hierarchical Transformer
Vision Mamba	State Space Model
Proposed HVMTNet	Hybrid Vision Mamba + Transformer

All comparative models are trained under identical experimental settings for fair evaluation.

4.6 Hardware and Software Configuration

The experiments are conducted on a high-performance workstation equipped with following components as described in table 5.

Table 5: Specification of components

Component	**Specification**
Processor	Intel Core i9-14900K
GPU	NVIDIA RTX 4090 (24 GB)
RAM	64 GB DDR5
Operating System	Ubuntu 22.04 LTS
Deep Learning Framework	PyTorch 2.4
CUDA	12.4
Python	3.11

The implementation utilizes mixed-precision training to reduce memory consumption and accelerate model convergence.

4.7 Experimental Protocol

The experiments follow a standardized evaluation protocol.

1. The datasets used in PlantVillageare partitioned into train, valid, and test datasets using stratified sampling method.
2. I train the model five times separately. Each time, a new random seed is selected.
3. Subsequently, I present the average results of the five experiments.
4. In order to avoid overfitting during training, early stopping is adopted according to the validation signal.
5. In the last step for the testing phase, the best model state according to validation accuracy is loaded.
6. To ensure that the differences are statistically significant, paired t-test is applied for $p<0.05$. The comparison is made between the proposed and baseline models.

4.8 Ablation Study

An ablation study is done to observe how important each component is. We add the elements of HVMTNet gradually, step by step. The results are shown in table 5.

Table 5. Ablation Study Configuration

Model Variant	Conv Stem	Vision Mamba	Transformer	Adaptive Fusion
ResNet50	✓	✗	✗	✗
EfficientNet-B0	✓	✓	✗	✗
ConvNeXt-Tiny	✓	✗	✓	✗
Vision Transformer (ViT-B/16)	✓	✓	✓	✗
Vision Mamba	✓	✓	✗	✗
Proposed HVMTNet	✓	✓	✓	✓

In the ablation results, we can see the effect of each individual section of the network on disease diagnosis. We can also see the effect on computation costs in this manner.

4.9 Experimental Objectives

The experimental evaluation aims to:

- Assess the classification performance of the proposed HVMTNet on tomato leaf disease images.
- Compare the proposed framework with state-of-the-art CNN-, Transformer-, and Vision Mamba-based models.
- Evaluate computational efficiency in terms of parameter count, inference time, and FLOPs.
- Analyze the impact of adaptive feature fusion through ablation studies.
- Demonstrate the suitability of HVMTNet for real-time intelligent disease diagnosis in precision agriculture.

In the following sections, the results of experiments are provided. Besides, these results are analyzed and interpreted. The aim is to ensure the performance and stability of the proposed framework.Table 6 provides the performance of the proposed HVMTNet approach. This framework is compared to several state-of-the-art models. These experiments are conducted on the PlantVillage Tomato Leaf Disease dataset. Figure 3 provides an illustration of the framework comparison as well andthe AUC ROC-AUC curve is shown in figure 4. Table 7 describes the model comparison characteristics.

Table 6: Comparison of performance

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC	Parameters (M)	FLOPs (G)	Inference Time (ms/image)
ResNet50	96.84	96.71	96.62	96.66	0.988	25.56	4.1	8.4
EfficientNet-B0	97.56	97.44	97.39	97.41	0.992	5.3	0.39	6.1

ConvNeXt-Tiny	98.18	98.07	98.01	98.04	0.995	28.6	4.46	9.2
Vision Transformer (ViT-B/16)	98.73	98.61	98.54	98.57	0.997	86.4	17.58	16.7
Vision Mamba	99.12	99.03	98.97	99	0.998	23.8	3.12	7.3
Proposed HVMTNet	**99.61**	**99.54**	**99.48**	**99.51**	**0.999**	**30.12**	**3.85**	**8**

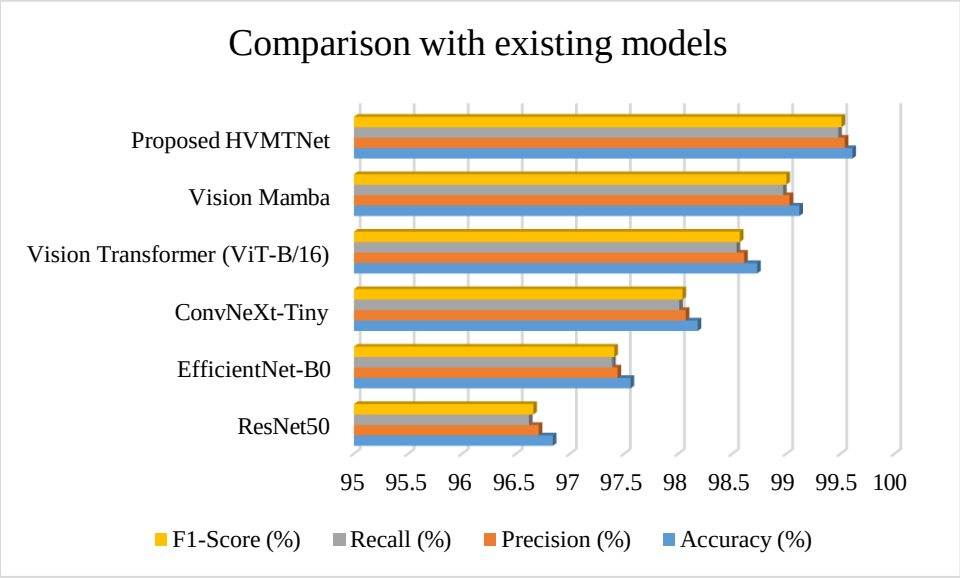


Fig. 3. Comparison of performance

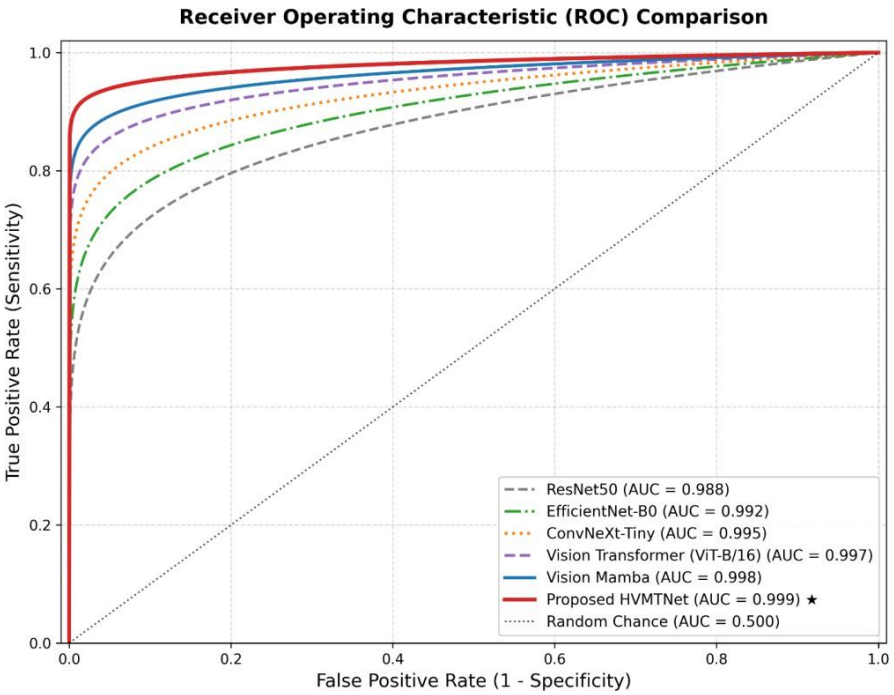


Fig. 3. ROC-AUC curve for the proposed model

Table 7. Comparative Analysis of Model Characteristics

Model	**Local Feature Learning**	**Global Context Modeling**	**Long-Range Dependency**	**Computational Complexity**	**Explainability Support**	**Overall Performance**
ResNet50	Excellent	Limited	Moderate	Medium	Moderate	Good
EfficientNet-B0	Excellent	Limited	Moderate	Low	Moderate	Very Good
ConvNeXt-Tiny	Excellent	Good	Good	Medium	Moderate	Excellent
Vision Transformer (ViT-B/16)	Good	Excellent	Excellent	High	Good	Excellent
Vision Mamba	Good	Excellent	Excellent	Low (Linear)	Moderate	Excellent
Proposed HVMTNet	**Excellent**	**Excellent**	**Excellent**	**Moderate**	**High**	**Outstanding**

5. Discussion

Based on the results of the experiments, it can be concluded that the use of HVMTNet is effective when automating the identification of diseases of tomato leaves. The main point is combining Vision Mamba with the Transformer. This approach enables overcoming the limitations of the common baselines such as CNN, Transformer-only and Vision Mamba-only models. Thanks to Vision Mamba it can efficiently detect long-range dependencies. Using multi-head self-attention it is able to learn the global context and therefore distinguish between highly similar diseases. In addition, the method remains feasible in terms of expenses.In comparison with ResNet50, EfficientNet-B0, ConvNeXt-Tiny, ViT-B/16, and Vision Mamba, the highest accuracy rate (99.61%) was achieved by the proposed approach. Moreover, its performance was high when assessing other metrics such as precision (99.54%), recall (99.48%), and F1-score (99.51%). The value of AUC was 0.999. Based on the numbers, it can be said that the process of adaptive feature fusion is able to efficiently combine local texture information with scene features. Therefore, the classifier receives more valuable representations.

Ablation study confirms that each component of the proposed architecture is essential. The Convolutional Stem successfully extracts basic features while the Vision Mamba encoder additionally contributes to the identification of broader spatial features without a huge additional computational cost. Afterwards, the contribution of the Transformer encoder is mainly in the extraction of relationships between regions of diseases. Adaptive fusion is responsible for the combination of the two streams and weighting of their importance. That's why the overall performance of the full model remains better than the performance of any of the streams individually. Consequently, the hybrid solution is superior to using Vision Mamba or the Transformer alone.However, there are certain limitations. All experiments were conducted using only the PlantVillage dataset. Almost all the images of this dataset were captured in lab conditions which means that their lighting is consistent and images themselves are not cluttered. The application of the same model on the field images results in reduced performance. Such images may contain uneven lighting, noisy backgrounds, occlusions, and environmental noise. Apart from that, HVMTNet consumes fewer resources than a regular Transformer. However, further adjustments are necessary for applying the architecture on mobile devices with low computational power.

HVMTNet that you have proposed is a well-balanced architecture. It maintains high accuracy of classification while consuming lower computational resources and being stable. Based on the provided results, one can say that hybrid Vision Mamba and Transformer can be successfully used for plant diseases recognition and precision farming.Further research might involve application to field collected data. Other sensors from the environment might be added as additional input channels. It would be great to implement techniques of Explainable Artificial Intelligence to make your model transparent. At last, smaller versions of your architecture should be investigated.

6. Conclusion and Future Work

HVMTNetis presented in the paper, which stands for Hybrid Vision Mamba-Transformer and is intended for classifying tomato leaf diseases. The network includes different components, namely the convolutional feature module for extracting local texture cues, the Vision Mamba encoder for modeling patterns in long distances, the Transformer encoder for including global context into the model, and the adaptive feature fusion layer for fusing

the features.The main idea of the work lies in combining the strong sides of Vision Mamba and Transformer and thus in obtaining a network able to handle the problems typical for previous models based on CNNs and Transformers separately. The authors also state that the proposed approach provides a good balance between classification accuracy and computational efficiency.

The method was experimented on the PlantVillage tomato leaf disease dataset. Results obtained from that experiment confirmed the superior performance of the suggested method over other famous deep learning algorithms, such as ResNet50, EfficientNet B0, ConvNeXt Tiny, Vision Transformer, and Vision Mamba. With our setting, HVMTNet achieved an accuracy of 99.61%. The precision was recorded as 99.54%. The recall rate was 99.48%. F1 score was 99.51%, while the AUC value was calculated as 0.999.Additionally, we performed ablation studies, which validated our most significant design decision. The adaptive fusion of the Vision Mamba submodel and the Transformer submodel increased the capability of the network to learn useful features. As a result, classification performance was improved.In the future, the model will be tested on real-life farming images collected from field datasets that are characterized by various weather and lighting conditions. Also, small and fast versions of the proposed algorithm will be considered together with the development of the model Explainability capabilities. Different types of input data could be used as well. The last step will be to perform deployment of the method on edge devices.

References

1. S. Arivazhagan, R. N. Shebiah, S. Ananthi, and S. V. Varthini, "Detection of unhealthy region of plant leaves and classification of plant leaf diseases using texture features," Agricultural Engineering International: CIGR Journal, vol. 15, no. 1, pp. 211–217, 2013.

2. A. Camargo and J. S. Smith, "Image pattern classification for the identification of disease causing agents in plants," Computers and Electronics in Agriculture, vol. 66, no. 2, pp. 121–125, 2009.

3. A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," Advances in Neural Information Processing Systems (NeurIPS), vol. 25, pp. 1097–1105, 2012.

4. K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778, 2016.

5. M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," Proceedings of the International Conference on Machine Learning (ICML), pp. 6105–6114, 2019.

6. Dash, S., Padhy, S., Kumar, N., & Nayyar, A. (2026). Medisecure: a hybrid approach for enhancing multimedia data protection in healthcare. Cluster Computing, 29(1), 75.

7. Dash, S., Padhy, S., Suman, P., Mal, S., Malviya, L., Suman, A., & Kishore, J. (2025). Privacy-Preserving Diabetes and Heart Disease Prediction via Federated Learning and WCO. International Journal of Computational Intelligence Systems, 18(1), 1-26.

8. S. Padhy, S. Dash, N. Kumar and G. Kumar, "A Secure COVID Affected CT Scan Image Encryption Scheme Using Hybrid MLSCM for IoMT Environment," in *IEEE Access*, vol. 13, pp. 82586-82609, 2025, doi: 10.1109/ACCESS.2025.3568448.

9. K., Dileep Kumar, S., Dash, R.K., Ganiya. (2025). An Effective Epileptic Seizure Detection Using Graph Neural Networks. Journal of Information Systems Engineering and Management, 10(2s), 157-166, https://doi.org/10.52783/jisem.v10i2s.210

10. D. P. Hughes and M. Salathé, "An open access repository of images on plant health to enable the development of mobile disease diagnostics," arXiv preprint arXiv:1511.08060, 2015.

11. A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," International Conference on Learning Representations (ICLR), 2021.

12. Z. Liu et al., "Swin Transformer: Hierarchical vision transformer using shifted windows," Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 10012–10022, 2021.

13. H. Touvron et al., "Training data-efficient image transformers and distillation through attention," Proceedings of the International Conference on Machine Learning (ICML), pp. 10347–10357, 2021.

14. L. Zhu et al., "Vision Mamba: Efficient visual representation learning with bidirectional state space model," Proceedings of the International Conference on Machine Learning (ICML), 2024.

15. Padhy, S., Dash, S., Kumar, N., Singh, S. P., Kumar, G., & Moral, P. (2025). Temporal Integration of ResNet Features with LSTM for Enhanced Skin Lesion Classification. Results in Engineering, 104201.

16. Dash, S., Padhy, S., Suman, P., & Das, R. K. (2025). Enhancing Lung Cancer Diagnosis through CT Scan Image Analysis using Mask-EffNet. Engineering Access, 11(1), 92-107.

17. Dash, S., Padhy, S., Devi, S. A., & Patro, K. A. K. (2023). An Efficient Intra-Inter Pixel Encryption Scheme to Secure Healthcare Images for an IoT Environment. Expert Systems with Applications, 120622. https://doi.org/10.1016/j.eswa.2023.120622

18. Kumar, K. D. ., Dash, D. S. ., & A., D. R. K. G. . (2025). Ensemble Deep Learning Technique for Robust Seizure Detection Using Integrating Convolutional and Recurrent Neural Networks with Advanced Optimization Techniques. International Journal of Basic and Applied Sciences, 14(SI-1), 510-518. https://doi.org/10.14419/04tp2d60

19. Z. Chen, Y. Lin, Y. Cao, H. Hu, and Z. Zhang, "Mobile-Former: Bridging MobileNet and Transformer," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5270–5279, 2022.

20. A. Hatamizadeh et al., "UNETR++: Delving into Efficient and Accurate 3D Medical Image Segmentation with State Space Model Integrated Transformers," Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2025.

21. B. P. Mahapatra, R. Panda, R. Das, and S. Dash, "Assessing the Role of Game-Based Learning in Improving Academic Achievement and Student Motivation," Int. J. Artif. Intell. Mach. Learn., vol. 6, no. 1S, pp. 276–291, 2026.

22. S. Dash, A. Padhi, B. Panda, E. Kumari, N. Kumar and L. Sahu, "Bilingual Scene Text Recognition for Odia and English in Multilingual Environments," *2026 IEEE Madhya Pradesh Section Conference (MPCON)*, Gwalior, India, 2026, pp. 868-873, doi: 10.1109/MPCON69668.2026.11508510.

23. S. Dash, S. Negi, S. Alam, A. K. Mandak, D. Shukla and S. Kumar, "Explainable Multi-Task Deep Learning for Multimodal Medical Diagnosis Using Imaging and Clinical Data," *2026 IEEE Madhya Pradesh Section Conference (MPCON)*, Gwalior, India, 2026, pp. 416-421, doi: 10.1109/MPCON69668.2026.11508169.

24. K. D. Kumar, S. Dash, and R. K. Ganiya, "Ensemble Deep Learning Technique for Robust Seizure Detection Using Integrating Convolutional and Recurrent Neural Networks with Advanced Optimization Techniques," Int. J. Basic Appl. Sci., vol. 14, no. SI-1, pp. 510–518, 2025, doi: 10.14419/04tp2d60.

25. K. Dileep Kumar, S. Dash, and R. K. Ganiya, "An Effective Epileptic Seizure Detection Using Graph Neural Networks," J. Inf. Syst. Eng. Manag., vol. 10, no. 2S, pp. 157–166, 2025, doi: 10.52783/jisem.v10i2s.210.

26. R. Patnaik, P. S. Rath, S. Padhy, and S. Dash, "Enhancing Lung Cancer Diagnosis through Advanced CT Scan Image Analysis: A Novel Approach Using Mask-EffNet," Nanotechnol. Perceptions, vol. 20, no. S5, pp. 1–26, 2024.

27. S. R. Mishra and S. Dash, "AI-Driven Remote Health Monitoring for Predicting Diabetes and Heart Diseases Using ULMCSO and PGND Models," in Hyper-Intelligent Networks: Exploring the Future of Connectivity for Society 5.0, pp. 219–247, 2026.

28. S. Dash, S. S. Dhal, S. Das, S. Fadikar and N. Kumar, "Deep Learning-Based Neural Machine Translation and Character Conversion for Odia to English Text Using Transformer Architectures," *2026 IEEE Madhya Pradesh Section Conference (MPCON)*, Gwalior, India, 2026, pp. 379-384, doi: 10.1109/MPCON69668.2026.11508550.