IEEE Open Journal of the Computer Society

Received 10 April 2024; revised 6 May 2024; accepted 23 June 2024. Date of publication 25 June 2024;
date of current version 18 July 2024. The review of this article was arranged by Associate Editor Jianzong Wang.

Digital Object Identifier 10.1109/OJCS.2024.3419027

The Power of Vision Transformers and Acoustic Sensors for Cotton Pest Detection

REMYA S [1], ANJALI T [1], ABHISHEK S [1], SOMULA RAMASUBBAREDDY [2], AND YONGYUN CHO [3]

[1]Department of CSE, Amrita School of Computing, Kollam 690525, India
[2]Department of Information Technology, Vallurupalli Nageswara Rao Vignana Jyothi Institute of Engineering and Technology, Hyderabad 500090, India
[3]Department of Information and Communication Engineering, Sunchon National University, Suncheon 57922, South Korea

CORRESPONDING AUTHOR: YONGYUN CHO (e-mail: yycho@scnu.ac.kr).

This work was supported in part by the Innovative Human Resource Development for Local Intellectualization program through the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea Government, MSIT, under Grant IITP-2024-2020-0-01489, 50%, and in part by the MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program supervised by the IITP (Institute for Information & Communications Technology Planning & Evaluation) under Grant RS-2024-00259703, 50%.
This article has supplementary downloadable material available at https://doi.org/10.1109/OJCS.2024.3419027, provided by the authors.

ABSTRACT Whitefly infestations have posed a severe threats to cotton crops in recent years, affecting farmers globally. These little insects consume food on cotton plants, causing leaf damage and lower crop yields. In response to this agricultural dilemma, we developed a novel method for detecting whitefly infestations in cotton fields. To improve pest detection accuracy, we use the combined efficiency of visual transformers and low-cost acoustic sensors. We train the vision transformer with a large dataset of cotton fields with and without whitefly infestations. Our studies yielded encouraging results, with the vision transformer obtaining an amazing 99% accuracy. Surprisingly, this high degree of accuracy is reached after only 10-20 training epochs, outperforming benchmark approaches, which normally give accuracies ranging from 80% to 90%. These outcomes underline the cost-effective potential of the vision transformer in detecting whitefly attacks on cotton crops. Moreover, the successful integration of acoustic sensors and vision transformers opens doors for further research and advancements in the domain of cotton pest detection, promising more robust and efficient solutions for farmers facing the challenges of whitefly infestations.

INDEX TERMS Acoustic sensors, vision transformer, neural network, VGG16, cotton white fly attack.

I. INTRODUCTION

Every nation's economy depends heavily on agriculture, and India is considered a country with a robust agricultural basis. The production of healthy, disease-free crops is one of the fundamental goals of agriculture. In India, the income-producing yield that is most important is cotton. In terms of global agricultural production, India comes in second. Farmers here in India grow an unlimited range of crops. About 41.49% of the Indian population is employed there, contributing 18% of the country's total GDP. Around the world, plant diseases are thought to be responsible for $60 billion worth of crop losses each year [1]. Rapid agricultural expansion is essential for earning lucrative foreign exchange and achieving self-sufficiency. The early loss of leaves or the development of diseases impacts cotton yields. As a result of the increased need for food, agriculture is severely strained. The agricultural information sector desperately needs a quick, effective, less expensive, and trustworthy method for identifying cotton infestations since severe disease cases can lead to no grain harvest. Accuracy, cost, and time might improve with a computerized practice for spotting plant diseases.

Plant diseases are expensive for the agriculture sector. For a consistent supply of high-quality food, crop disease control is crucial. Early diagnosis of plant illnesses is essential for reducing disease transmission and enhancing effective treatment strategies. Effective crop management depends on assessing the health of the plant disease, which involves predicting methods and models for treatment application. The agronomist benefits from using computer methods for identifying and diagnosing plant diseases. An arbitrary process of determining disease detection is possible using older methods.

However, modern technology offers a way that may be used to diagnose plant illnesses accurately. The traditional approach for identifying plant diseases relies solely on visual inspection, which requires additional human effort, laboratories with specialized equipment, etc., which starts with hiring a domain expert who visits the location and examines the crop using the visual examination. This approach takes a lot of time and effort. Crop monitoring must be done continuously as well. On the other hand, for centuries, farmers and planting specialists had to deal with various issues and persistent agricultural challenges, such as many cotton diseases. Another significant challenge is that farmers need access to professionals in some regions. The need to gather data on cotton's healthy growth and development in real-time is stressed in intensive agriculture and hydroponics.

Crop diseases cannot always be found with optical tracking. It will not be possible to accurately assess crop disease using conventional diagnostic techniques. Plant diseases cause losses that are both qualitative and quantitative. Automatically detecting disease signs as soon as they develop on plant leaves is crucial in agricultural research because it allows for the monitoring of vast fields of crops. Identifying numerous plant diseases may be quite challenging for a farmer. Researchers have increasingly concentrated on the agricultural sector to find crop illnesses using automated approaches. Cotton is one of the world's most significant crops, with a global production area of over 31 million hectares. However, the crop is vulnerable to a variety of pests and illnesses, one of which is whitefly infestation. The whitefly is a small insect that feeds on the sap of the cotton plant, causing leaf damage and lowering crop production.

The deep learning-based strategy suggested in this study provides a more efficient and accurate solution for detecting whitefly infestations in cotton fields, in comparison to previous manual methods. By utilising the vision transformer architecture, the model is capable of identifying whitefly attacks, even during the first stages when the harm is limited. This allows for the prompt adoption of efficient management strategies. Incorporating this cutting-edge technology for pest identification has the potential to greatly enhance agricultural sustainability and output. The main contributions of the article are presented below:

1) Development of a comprehensive dataset featuring 32,000 images showcasing whitefly infestations on cotton crops, accompanied by multi-class ground truth annotations.
2) Introduction of an innovative approach that combines vision transformers with cost-effective acoustic sensors to improve pest detection accuracy in cotton fields.
3) Comparative evaluation against state-of-the-art models, including MobileNet, ResNet152v2, and VGG-16, demonstrating the superior performance of our proposed vision transformer method.
4) Acknowledgment of the potential of integrating vision transformers and acoustic sensors for more effective cotton pest detection in agricultural settings.

The previous studies on recognizing whitefly attacks on cotton growing was conducted using various techniques including Support Vector Machines, which is the common machine learning algorithm of the traditional methods. Convolutional Neural Networks [2] and Hybrid Artificial Neural Networks [3], which are the new and more developed methods. On the other hand, the SVM-based approach [4] displayed a perfect level of accuracy in categorizing the degrees of the whitefly attacks though mostly there are such impediments as the complexity of computation, optimization of the hyperparameters [5], [6] and interpretation of the results. Although neural network-based systems have made significant progress in disease detection and staging, the bottleneck of overfitting combined with the hardware resource demands mainly act as barriers to the enthroning of such systems in the environment [7], [8], and [9]. In addition to this, other hybrid models are also developed to tackle this pest attack. However, there exist many challenges regarding hyperparameter tuning and generalization of the model. (The summary of the literature review is shown in supplemental material). Our proposed approach has proven to be a promising method for reducing the negative effects of cotton whiteflies on crop output with an excellent accuracy rate of 99%, which guarantees reliability, saves time, and improves efficiency in detecting and managing pests.

This article is organized as follows: Section II describes the materials and methods which includes the dataset used in this work, data preprocessing, and the role of acoustic sensors in cotton pest detection. The methodology and experimental setup of the proposed approach are presented in Sections III and IV respectively. Section V concludes the article and Section VI highlights the directions for future work.

II. MATERIALS & METHODS

A. DATASET

In this study, we first developed a dataset on whitefly attacked and healthy leaves containing 20000 images with the help of Kerala & Tamil Nadu agriculture department, which are divided into different classes such as healthy, unhealthy, mild, severe, and nutrition deficiency. The data for this study was gathered by conducting thorough field surveys in cotton fields that were impacted by infestations of whiteflies. Comprehensive coverage of the agricultural environment was achieved by capturing high-resolution photographs utilizing mobile phones and digital SLR cameras. The AgriPK is the other dataset that is used here for the experimental analysis and study. In addition, cost-effective sensors were strategically positioned in the fields to capture insect sounds and environmental cues, resulting in acoustic recordings. Subsequently, we proposed a vision transformer-based approach to classify the image dataset.

To train the vision transformer, ResNet152v2, VGG16, and MobileNet, cotton fields with and without whitefly infestations were photographed [10]. The total samples, which had over 32,000 images, were compiled from various sources,

including fieldwork and other repositories. Infested cotton fields comprised half of the images, while uninfected cotton fields made up the other half. The collected data set is fair, clear, and has minimal interference. However, images taken in real-life situations are prone to interference from factors such as overlapping leaves, wind resistance, and intense sunlight. To overcome these challenges, a controlled environment was created to reduce the impact of environmental factors [5] and [11].

B. DATA PRE-PROCESSING

The efficacy of our proposed system for detecting cotton pests using ViTs and transfer learning techniques is extremely dependent on data preprocessing. Early detection of cotton pest infestations is crucial for efficient pest control. For this research, we collected image data and acoustic data from different sources. Then we included pertinent environmental metadata and resized the images to a predetermined resolution to provide consistent data input for the proposed ViT models. The detailed preprocessing methods are discussed in the following sub-sections. We employed data normalization, standardization, and sampling to minimize class imbalances and divided the data into training, testing validation sets. This preprocessed multimodal data is fed into the ViT model through transfer learning.

1) DATA AUGMENTATION

Using this technique, the sensor-based and visual data are randomly modified by flipping, rotating, zooming, and adjusting brightness. To make the model more resilient and scenario-adaptable, a variety of data augmentation techniques are used in addition to this, including normalization, standardization, contrast enhancement, and noise introduction. By using these tactics, we can expand our training set of data and expose the model to a wider range of events. Gaussian noise addition, which involves randomly choosing noise from a Gaussian distribution and merging it with the original data, was used in this work to add noise to the data. By training the model with additional permutations, this strategy improves its performance and noise tolerance in a variety of circumstances. The images were subjected to both contrast enhancement and noise injection. This brings up subtle patterns and characteristics, which help to identify infestations of whiteflies. One common technique for enhancing contrast is histogram equalization. The uniqueness is maximized by redistributing pixel intensities. This improves the ability of the model to identify pest traits.

Then the data is standardized to have a zero mean and one standard deviation. This eliminates scale disparities that could impair model learning. On the other hand, normalization adjusts pixel intensities or amplitudes to a range, usually from 0 to 1. This keeps numerical stability and prevents activation function saturation in deep neural networks. We efficiently prepared the dataset for ViT model training utilizing these preprocessing methods. This makes the model significant whitefly infection patterns in cotton fields. Noise injection, contrast enhancement, standardization, and normalization increased model performance and generalization.

Maintaining data integrity protects cotton pest detection parameters including pest borders and environmental characteristics. By using data augmentation methodologies, we reduce overfitting and increase the model's ability to process cotton pest detection sensors and visual data with different patterns and characteristics.

C. DATA ACQUISITION

In the data acquisition process for cotton pest detection, we used different cameras and acoustic sensors. A structured approach is employed to guarantee thorough coverage and precision of data. At first, thorough planning is done for field surveys, which involves identifying specific regions that are prone to pest infestations. SLR cameras are strategically positioned around the cotton fields to collect high-resolution images. We can manually activate the cameras to ensure optimal timing and coverage. Assigning accurate geographical information to each image by geotagging enables exact spatial referencing. In addition, strategically placed acoustic sensors capture insect sounds and environmental cues, enhancing the visual data. Data integrity is maintained through the implementation of quality assurance techniques, which include real-time monitoring and regular calibration of cameras and sensors. Post-acquisition validation is a process that confirms the sharpness of images, the accuracy of colors, and the alignment with geolocation data. It also verifies the clarity and consistency of sound recordings.

D. METHODS

Whitefly attack detection in cotton fields was accomplished using neural network architecture of vision transformer and compared with ResNet152v2, VGG16, and MobileNet architectures. Several computer vision applications have used the well-known convolutional neural network designs ResNet152v2 and VGG16. The vision transformer architecture, which was created primarily to address this issue, comprises several blocks of feedforward and self-attention layers.

The Vision Transformer is an extension of the Transformer architecture, initially introduced for NLP tasks, to computer vision tasks such as image classification. The key idea is to treat the image as a sequence of patches or feature maps, which are then processed by the transformer encoder to generate a global representation of the image. The encoded data is subsequently processed into a classifier, such as a linear layer, to provide the eventual prediction. The VIT model distinguishes itself from conventional convolutional neural networks (CNNs) by employing self-attention methods to explicitly represent the connections among various regions in the image, instead of relying on convolutional filters to extract local characteristics [12].

1) ROLE OF ACOUSTIC SENSORS IN COTTON PEST DETECTION

Incorporating acoustic sensors is a significant improvement in our method of detecting cotton pests. These sensors are essential for quickly identifying cotton pests because they pick up on the quiet sounds made by the pests as they move or eat. This feature allows farmers to spot pests when they are still young, making it easier to take action and protect the crops from damage. One of the main advantages of these sensors is that they provide a non-intrusive way of monitoring cotton fields, ensuring that the harvest remains undisturbed. They also offer a more cost-effective and environmentally friendly option compared to existing methods, such as visually inspecting the plants or using expensive chemicals.

In addition to this, the continuous observation of insect activity within the cotton fields, and acoustic sensors provide timely surveillance. The provision of live data to farmers equips them with practical knowledge, enabling them to promptly address the identification of pests. By leveraging data-driven analysis, not only are insect control initiatives enhanced, but also methods for safeguarding crops are optimized. Acoustic sensors help to reduce the requirement for wide-ranging chemical treatments in agriculture, thereby supporting sustainable farming techniques.

Piezoelectric sensors are utilized for this indication, in this research work. The operational mechanism of these sensors involves the conversion of electrical signals from mechanical vibrations, such as those produced by insects or feeding activities in cotton fields. These vibrations are detected and converted into electrical impulses. To improve the signal quality and discern significant attributes that signify insect activity, the unprocessed sensor data undergoes preprocessing techniques, including filtering and signal conditioning. To identify and classify instances of pest infestation, the machine learning algorithms analyze data comprising temporal dynamics, frequency components, and amplitude patterns. Insights into instances of pest activity can be generated by machine learning models, which can aid in the implementation of suitable pest management strategies and reduce crop loss in cotton fields. As a result of their ability to detect minute mechanical vibrations, piezoelectric sensors are highly beneficial for agricultural pest detection and insect monitoring. These sensors can transform cotton pest control, guaranteeing the durability and success of cotton crops while preserving resources and minimizing environmental harm.

2) TRANSFER LEARNING

Each image was downsized to 224 × 224 pixels to provide a uniform size, and the pixel values were set to range from 0 to 1. Subsequently, we applied data augmentation techniques to enlarge the training dataset and prevent overfitting. We applied a pre trained EfficientNet-B0 model as a feature extractor. Except for the final layer, whose weights were frozen, a new fully connected layer with different output units, one for each type, was added [13].

3) VISION TRANSFORMER

In the proposed method, we used a ViT architecture to classify and detect the pest types. A feedforward neural network was placed after several multi-head self-attention layers in the ViT architecture. The Adam optimizer was used to train our model across 20 epochs with a learning rate of 0.0001 and a batch size of 32.

4) NETWORK ARCHITECTURES

ViT has recently gained prominence as a potent tool for diverse applications, particularly in the field of computer vision. The significant potential for transforming the identification of cotton pests is notably appealing owing to its exceptional efficacy in visual data processing. ViT is exclusively employed to analyze image data to accurately identify cotton pests through this innovative approach.

The proposed methodology makes efficient use of the capabilities of the Vision Transformer architecture to analyse image data. The implementation of the ViT model enables the framework to accurately detect and classify cotton pests in images, thereby contributing to the improvement of cotton crop management and pest detection. In general, ViTis composed of the following layers: input, encoder, fully connected, and output layer. The transformation in Fig. 1 depicts the various components of the proposed model and their respective connections. This contributes to an enhanced comprehension of the model.

5) DIFFERENT PHASES OF VIT MODEL

1) *Patch Embedding:* The patch embedding technique is employed to generate a sequence of non-overlapping patches from the input image. Each individual patch is encoded as a vector consisting of pixel values. Denoted as X, the input image possesses the following dimensions: H × W × C. The image is partitioned into P × P non-overlapping patches, and a sequence of patch vectors $X = [x_1, x_2, \ldots, x_n]$ is generated, with the dimension of x_i being P × P × C. Every patch vector is transformed into a solitary vector with the length $P2timesC$, represented as x_i'.
2) *Linear Embedding:* In this stage, a learnable linear projection is used to project each patch vector into a higher-dimensional space. The model then acquires a more expressive representation of each patch. Let W be a learnable weight matrix, with dimensions $D \times (P^2 \times C)$, where D is the output dimension of the linear projection, resulting in a new sequence of vectors $H = [h_1, h_2, \ldots, h_n]$, where $h_i = W \times x_{i'}$ where $x_{i'}$ is the patch vector.
3) *Positional Encoding:* To capture the spatial structure of the image, the model also adds a learnable position embedding to each patch vector. This provided a model with information on the location of each patch within the image. Let E be a learnable position embedding matrix, with dimensions $D \times N$, where N is the total

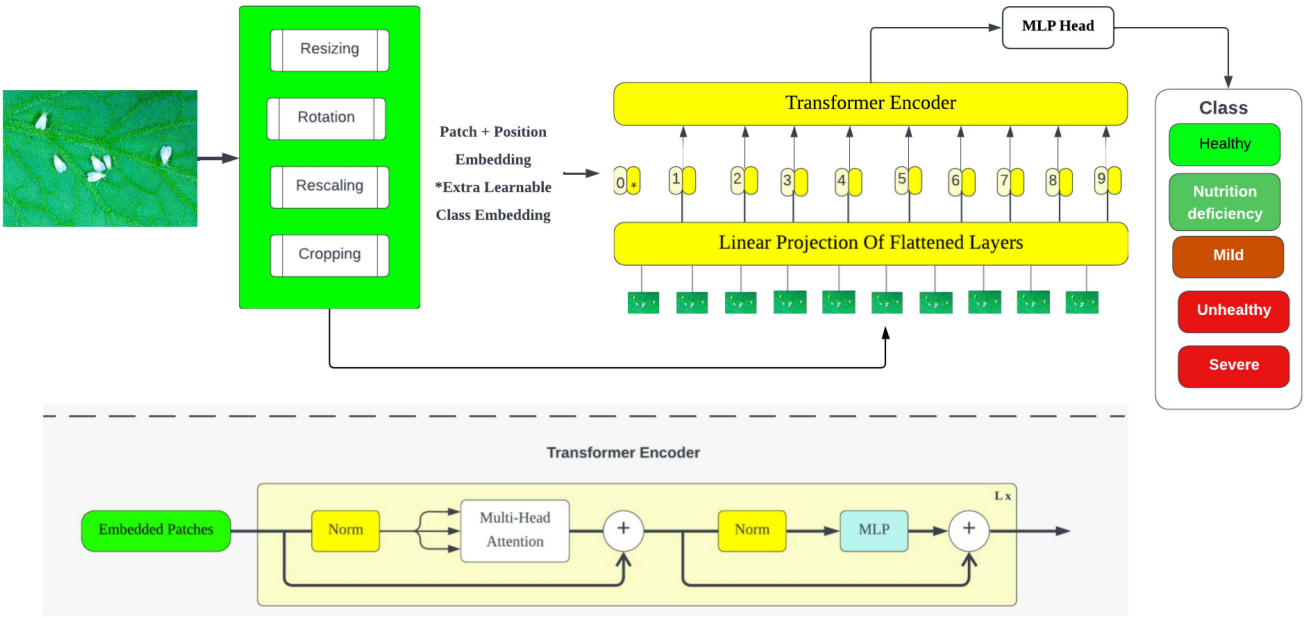

FIGURE 1. The proposed model architecture.

number of patches in the image. We concatenate the position embedding to each patch embedding, resulting in a new sequence of vectors $Z = [z_1, z_2, \ldots, z_n]$, where $z_i = h_i + e_i$.

The position embedding can be computed as follows:

$$e_{i,j} = \begin{cases} sin\left(\frac{j}{(10000(\frac{2i}{D}))}\right) & \text{for } i = 0, 2.., D-1 \\ cos\left(\frac{j}{(10000(\frac{2i}{D}))}\right) & \text{for } i = 1, 3, \ldots, D-1 \end{cases} \quad (1)$$

Here in both cases j = 0, 1, 2,.., N-1

4) *Transformer Layers:* The L times repeated self-attention and feedforward layers with ReLU activation function make up the Transformer layers. Each layer receives a set of patch embeddings with positional encodings X_l as input and produces a new set of embeddings X_{l+1} which represents the transformer layer X_l. Each transformer layer employs a self-attention mechanism that calculates the weighted sum of each embedding in X_l, with the weights determined by how similar the embeddings are to one another. Like the residual connections in other neural network architectures, the resulting weighted sum was subsequently added to the input embeddings. The skip connection is then used to pass the output embeddings through [14].
5) *Classification Head:* The output embeddings of the final transformer layer are averaged to generate a single vector representation of the input image using the ViT model. The classification head, which comprises a linear layer and softmax activation function, is then fed into this vector. The output of the linear layer is normalized by the softmax activation function such that the values add up to 1. Consequently, the model can forecast the likelihood of each class. Let Q, K, and V be linear projections of Z of dimensions $D \times N$, and let Z be the vector sequence that is input to the Transformer encoder layer.

We compute the self-attention matrix A as follows:

$$A = \text{softmax}\left((Q^T * K)/\sqrt{(D)}\right) * V \quad (2)$$

where softmax is the softmax function applied element-wise to each row of the matrix, and $\sqrt{(D)}$ is a scaling factor to reduce the variance of the dot product. The connections between each pair of input vectors are represented by self-attention matrix A. Standard gradient descent and backpropagation methods are used to train the ViT model. For improved outcomes, the cross-entropy loss function was also used. By switching out the classification head and retraining the model on a labelled dataset, the model can then be fine-tuned for a particular downstream task, such as image classification. Each vector in the series was then subjected to FF neural network with two linear transformations. Let W_1 and W_2 be D' × D and D × D', respectively, and be the learnable weight matrices for the FFN. We created a new sequence of vectors $U = [u_1, u_2, \ldots, u_n]$ by applying an FFN to each vector z_i in the sequence. Therefore, the vision transformer architecture takes advantage of both transformers and convolutional neural networks (CNNs) to enhance image analysis. However, to handle different modalities, the original ViT design probably needs to be modified. Additionally, to boost performance, additional neural network topologies might be combined.

In summary, the vision Transformer architecture is designed to leverage the strengths of both convolutional neural networks (CNNs) and Transformers to improve image analysis tasks.

III. PROPOSED METHODOLOGY

A. SYSTEM MODEL AND ARCHITECTURE

In the domain of cotton pest detection, the adoption of cutting-edge neural network architecture, known as the vision transformer, offers substantial advancements. Our approach in this research leverages the power of pre-trained vision transformer models for feature extraction, capitalizing on their remarkable performance in various computer vision applications. More precisely, we employ a pre-trained vision transformer model as the basis, augmenting it by incorporating fully connected layers to effectively carry out the essential duty of pest categorization. The layers are tailored to correspond with the distinct kinds of pests identified in our databases, guaranteeing precise classification of pest species. Furthermore, our model uses a multitask learning strategy that allows binary and multiclass classification tasks to be completed simultaneously. This flexible approach allows us to identify a wide range of pests, which is essential for accurate pest identification.

ViTs' ability to recognize complex correlations between images makes them a very effective tool for pest identification in agricultural settings. Unlike traditional models that focus on specific regions of an image, ViTs can scan the entire image, allowing them to detect minute patterns associated with insect infestations. These patterns could include colour variances, distinct patterns of development, or the presence of pests themselves.

1) ViTs analyse the images globally, which enables them to gather contextual data and deep correlations for the entirety of the image. CNNs deviate from this comprehensive method since they frequently rely on particular regions of the image to extract information.
2) ViTs focus on important regions of the image and use a self-attention mechanism that allows them to consider the connections between different patches.
3) ViTs' excel data handling capabilities make them ideal for difficult computer vision tasks. This is particularly crucial for agricultural pest identification because vast datasets are needed to create accurate models.
4) Even though ViTs are smaller than other CNNs, they require less training epochs to achieve acceptable accuracy. The accuracy rate of the ViT model in this investigation reached 99% after a maximum of 20 training epochs.

We improve our cotton pest monitoring system by using sound sensors, which provide an inexpensive and non-intrusive way to identify insects early on. When insects are active, these sensors listen for and analyse the noises they make. The benefits of acoustic sensors and vision transformers are combined in this architecture to produce a complete insect detection system. It provides accurate categorization results and fast analysis. This novel strategy can be used to greatly improve cotton pest control and is also an economical and environmentally responsible choice. Putting this strategy into practice will improve crop vitality and long-term viability while also lowering costs to the environment and the economy [15].

During the model training phase, a 20-epoch training method was used on the dataset, to train ViT. The ViT architecture, comprised of FF and self-attention layers, was developed to overcome the drawbacks of these approaches. The ViT architecture forms the basis of a visual transformer originally developed for NLP tasks [16]. The transformer's capability to interpret coincident data sequences simultaneously is one of its prime - advantages, coveted for CV applications [17]. The self-attention mechanism of a vision transformer is one of its primary characteristics. By doing so, the network can discover fine-grained information necessary for object detection and semantic segmentation tasks, by analyzing smaller sections of an image instead of a complete image. Besides, ViT is adept at managing enormous amounts of data, which makes it suitable for sophisticated CV applications. The detailed Network architecture and workflow of the vision transformer is shown in Fig. 2. In this study, ViTs were utilized to handle large datasets, enabling the learning of diverse features and patterns. The ViT model was trained using a dataset obtained from the previous phase, with a 20-epoch training schedule and a batch size of 32 on the Adam optimizer with a learning rate of 0.001. The dataset was partitioned into training, testing, and validation sets in the proportions of 60:20:20; evaluation metrics such as accuracy, log loss, ROC, and AUC curves were employed to assess the neural networks' efficacy.

The proposed strategy was evaluated using a separate test set, and performance metrics such as accuracy, precision, recall, and F1-score were calculated. Comparisons were made with benchmark strategies, including ResNet152v2, VGG16, and MobileNet. The model was trained using the Stochastic Gradient Descent (SGD) optimizer with a learning rate scheduler to prevent overfitting and improve generalization. The training process was monitored using accuracy, loss, and validation metrics, and early termination was applied to avoid overfitting. The parameters were fine-tuned using the validation set to optimize the hyperparameters of the model. Overall, the proposed approach showed promising results and areas for further improvement were identified during evaluation and the detailed analysis is explained in the results and discussion section.

B. PSEUDO-CODE FOR VISION TRANSFORMER

ViT is a DL architecture for CV applications that manipulates the transformer architecture, initially created for NLP. The pseudocode for the ViT is described in Algorithm 1.

An input image is broken down into non-overlapping patches, which are then linearly projected into flattened feature vectors. These vectors are fed into a typical transformer

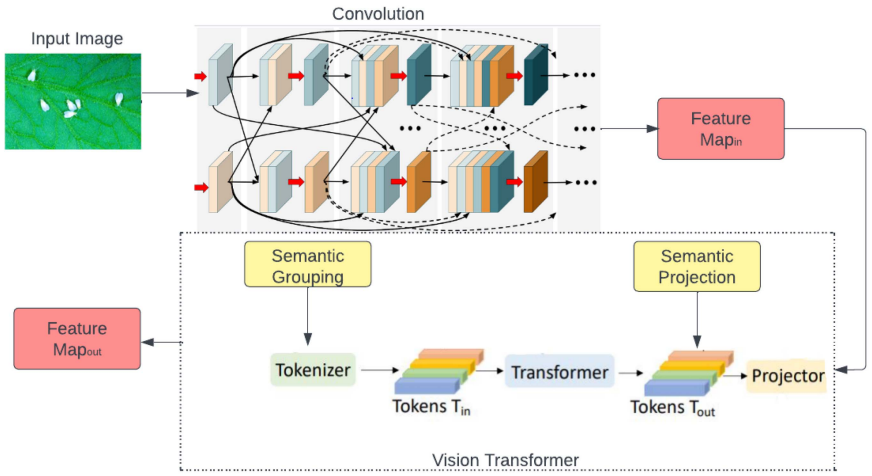

FIGURE 2. Network architecture of vision transformer.

encoder, where self-attention mechanisms capture global relationships among the patches, allowing for effective feature extraction. Relative positional embeddings are added to patch embeddings to incorporate positional information.

The transformer encoder runs these embeddings through successive layers, improving hierarchical representations, and the final output of the last layer is used for classification tasks through a conventional fully connected layer. During training, the model is optimized, minimizing cross-entropy loss, and the model's parameters are modified through backpropagation. The modular and self-attention-based architecture of the ViT empowers it to achieve competitive performance on many image recognition benchmarks while demonstrating flexible generalization capabilities.

IV. EXPERIMENT RESULTS

A. EXPERIMENTAL SETUP

We curated a diverse dataset for our project on the power of Vision Transformers in cotton pest detection, drawing data from multiple sources, including visual imagery and sensor readings, to ensure a comprehensive representation of various cotton pest indicators and data collection methods. We scaled all of the images to a standard resolution and normalized the sensor data before training the Vision Transformer models. This ensured that the neural network would perceive the data consistently. We used data preparation techniques including rotation, flipping, and random cropping to enhance the model's capacity for generalization. The model's ability to handle the numerous cotton pest data types and their sources was greatly enhanced by these techniques.

The visualization technique assists in identifying patterns, relationships, and potentially anomalous data points by presenting numerous elements about each other. These understandings are necessary to verify the authenticity of information, evaluate the significance of specific features, and comprehend potential relationships between various parts. Pair plots provide important information to enhance the architecture of our system and its ability to accurately identify

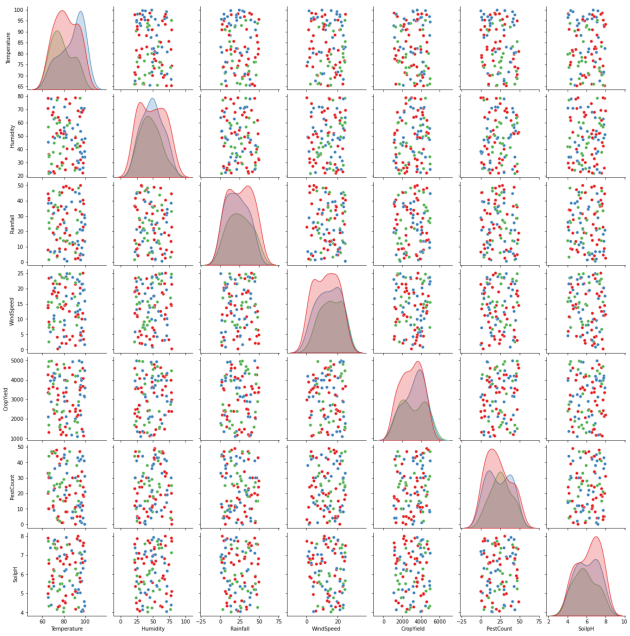

FIGURE 3. Pairwise relationships based on the dataset.

and classify cotton pest problems [18]. Comprehending the data and making critical decisions, including feature selection and preprocessing, are crucial for enhancing the classification model's efficacy and accuracy. Fig. 3 shows pair plots representing various classes and traits.

The experimental results of the proposed approach demonstrate that the vision transformer achieved a remarkable accuracy of 99% even after undergoing twenty training epochs. This technology represents a substantial advancement compared to conventional manual techniques for detecting whitefly infestations, which are often time-consuming and susceptible to human error. Moreover, to handle substantial volumes of data enables it to acquire a diverse set of characteristics and patterns from the data, rendering it resilient and capable of managing real-world situations [19], [20], and [21].

Algorithm 1: Pseudo Code for Vision Transformer.

Input: Pre-trained model, dataset
Output: Trained ensemble model with transfer learning and Vision Transformer
DataPreprocessing:
Image tensor of shape $(batchsize, numofchannels, height, width)$
Predicted class probabilities
Load the dataset, including images and corresponding labels.
Resizeall images to a consistent resolution.
Normalize pixel values to the range [0, 1]
Perform data augmentation techniques (random rotation, flipping) to enhance dataset diversity.
Transfer Learning: Load the pre-trained model
Replace the final classification layer(s) for the new task
Freeze pre-trained layers. If FineTuningLayers is not None:
Model.FreezeLayersExcept(FineTuningLayers)
Train the model on the new data.
Evaluate the fine-tuned model
Training Procedure:
Settraining parameters, including the number of epochs, batch size, and early stopping criteria.
For each training epoch
Initialize training metrics
For each batch in the fine-tuning dataset
For each batch in the fine-tuning dataset:
Load a batch of lesion images and their corresponding labels.
Pass the images through the model to extract features.
Concatenate the feature maps
Pass the concatenated features through fully connected layers.
Compute the loss between predicted and true labels.
Backpropagate the loss to update the weights of the ensemble model.
Update training metrics.
Vision Transformer:
$embeddings \leftarrow convolutionallayer(input)$
$embeddings \leftarrow reshape(embeddings, (batchsize, numofpatches, embeddingsize))$
For $i = 1$ to $numoflayers$ **do**
$attentionoutput \leftarrow multiheadattention(embeddings)$
$attentionoutput \leftarrow layernorm(embeddings + attentionoutput)$
$feedforwardoutput \leftarrow feedforward(attentionoutput)$
$embeddings \leftarrow layernorm(attentionoutput + feedforwardoutput)$
End For
$classtokens \leftarrow learnableparametervector(shape = (batchsize, 1, embeddingsize))$ Add learnable class tokens to embeddings
$embeddings \leftarrow concatenate([classtokens, embeddings], axis = 1)$
$output \leftarrow globalaveragepooling(embeddings)$
$output \leftarrow fullyconnected(output, numofclasses)$
return $softmax(output)$

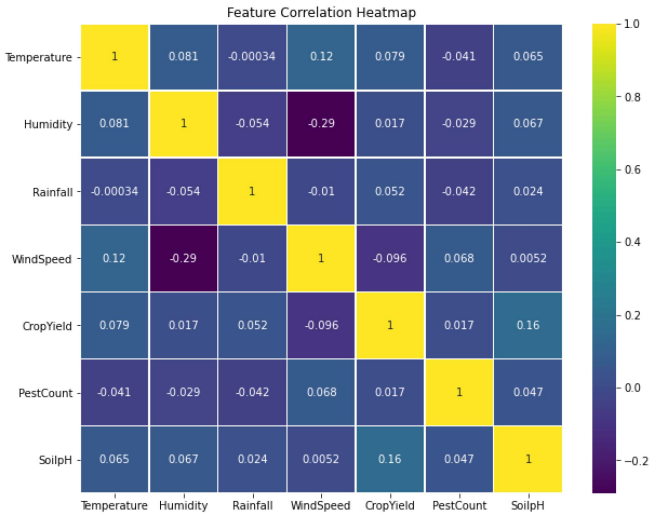

FIGURE 4. Correlation Heatmap of different features and classes.

The heatmap is a valuable tool for analyzing similarities and variances among classes. It improves the precision of the classification model, as seen in Fig. 4. Strong positive correlations occur when there are shared image attributes that result in challenges with categorization, whereas negative correlations imply contrasting visual traits. Heatmap data can be used to impact feature selection and model design, resulting in improved classification performance and data augmentation operations. Furthermore, the heatmap exposes possible disparities in class distribution, highlighting the necessity of employing data balancing methodologies [22], [23], [24], and [25]

The model's performance can be improved by tweaking this data. The model may be adjusted, for instance, to more accurately depict the difference between the images of crops with whitefly infestations and crops in good condition, according to the distribution center [26] and [27].

Ensuring the convergence of the model is a primary objective for accuracy. As the model becomes more acquainted with the data during the training process, its accuracy on the training set is expected to increase. Practitioners can determine if the model is converging as expected by examining the train accuracy graph, which provides a clear view of the learning process. If the graph shows a significant increase in accuracy during the early stages, it suggests that the model is learning quickly and effectively generalizing to the training data.

If the accuracy graph shows slow or irregular growth, the model may be either overfitting or underfitting to the training data. Another possibility is that the training set is either too small or too similar to the test set. In addition to monitoring convergence, the training accuracy can also identify any problems with the model or the training process. If the graph shows a sudden fall in accuracy, it could indicate that the model is overfitting to the training data and needs to be regularised. Conversely, if the accuracy graph levels out, it may indicate

that the model has reached its maximum potential and further training is unlikely to enhance its performance [29].

For instance, a specific model may be more appropriate for the task if it exhibits a better trade-off between accuracy, precision, and recall. Likewise, a model might be considered more reliable if it generates a curve with a greater area under the curve (AUC). This suggests an improved balance between the ability to retrieve relevant results (recall) and the accuracy of those results (precision) over a broader spectrum of thresholds.

Evaluating the trade-off between the accuracy of correctly identified cotton fly attacks and the rate of incorrectly identified attacks is essential when analyzing the ROC curve of cotton fly attack detection using the MobileNet architecture. Unlike the false positive rate, which measures the number of negative samples that are mistakenly identified as positive, the true positive rate quantifies the number of positive samples that are accurately discovered.

Our project's goal is to evaluate Vision Transformers' ability to identify pests in cotton. To accomplish this, we performed a performance assessment on three well-known models: ResNet152v2, VGG16, and MobileNet. The goal was to ascertain whether neural network designs were suitable for this task [30]. Based on their demonstrated performance in a variety of computer vision tasks, including the representation of complex images, the choice to employ any of these models was made. By employing several strategies, we were able to assess their competence in the difficult task of recognizing cotton pests. We carefully separated the data into sets for testing, validation, and training, making sure to preserve the class distribution for an objective assessment of the model. To improve and fine-tune the model parameters and avoid overfitting, the training data set has to be used.

The pre-trained weights of these neural network designs were leveraged using transfer learning, which resulted in a reduction of training time and data requirements. Our main goal was to fine-tune the network's last layers such that they were especially useful for recognizing pests that affect cotton. We performed ablation tests to assess how different types of data affected the multimodal system's overall performance. This complex experimental setup was designed to clarify which neural network topologies function best and whether they are suitable for using numerous detection techniques to find cotton pests.

We used the concept of segmentation in our proposed cotton pest identification approach. This comprises segmenting specific interest regions such as the cotton fields and potential pest indicators inside the dataset. We carefully marked off these sections, examined the characteristics of the cotton plants, and looked for evidence of insect infestations. We combined bespoke rules that are specifically created based on our knowledge in the sector with traditional image processing approaches to achieve accurate segmentation results. Our multimodal analysis, which employed Vision Transformer in a transfer learning approach, was predicated on the segmented regions. This method produced positive results in accurately identifying and classifying cotton pests as shown in Fig. 5.

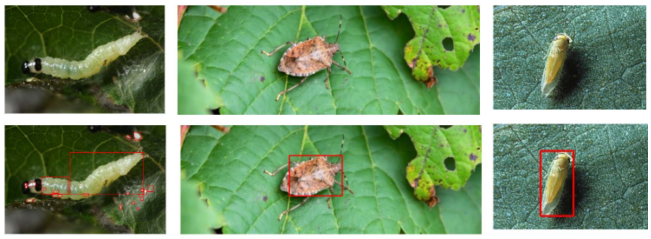

FIGURE 5. A set of input images and its detection results.

B. ANALYSIS BASED ON PERFORMANCE METRICS

In the suggested approach, assessing the deep learning model comes last. A test set that wasn't used during training was used to evaluate the model's performance. Several measures, such as accuracy, precision, recall, and F1-score, were used to determine the model's performance. Using the same criteria, the model was contrasted with various benchmark techniques, including ResNet152v2, VGG16, and MobileNet. The performance of the suggested strategy was evaluated using the assessment data, and any potential improvement areas were noted.

ResNet152v2, an enhanced version of ResNet152, addresses the vanishing gradient issue in deep neural networks during training. It's suitable for recognizing multimodal images as it consists of residual blocks with convolutional and batch normalization layers, allowing the network to learn complex features. VGG16, originally designed for image categorization, can identify infestations in photos by preprocessing images with scaling, grayscale conversion, and pixel value normalization. Its architecture, with max-pooling and convolutional layers, aids feature extraction from images. MobileNet, designed for mobile and embedded devices, uses depth-wise separable convolutions to speed up processing. Depth-wise convolution filters input channels individually, while point-wise convolution combines filtered outputs. This design optimizes computation for less powerful devices.

In our experimental design, applying transfer learning with pre-trained weights was a successful approach. The pre-trained models, which were initialized with information from sizable image datasets, permitted faster convergence during training, and showed that the learned features could be applied to detection problems. The performance of the model was enhanced by fine-tuning the final few layers to conform to the special features images.

The strong generalization of the models to new data was supported by their strong performance on the test set. The consistency of our multimodal categorization approach in terms of performance metrics across many evaluation measures was reassuring the efficiency. Compared with traditional manual procedures, the proposed methodology is designed to provide a more efficient and accurate means of detecting infestations.

TABLE 1. Analysis of Performance Metrics for Different Models

Method	ROC-AUC	Accuracy	P	R	F1 Score
SVM	75	85.2	0.72	0.88	0.79
KNN	70	81.2	0.69	0.65	0.66
MobileNet	85	83	0.86	0.88	0.87
ResNet152V2	75	50	0.43	0.46	0.45
VGG16	59	79	0.75	0.74	0.74
ViT	98	99	0.98	0.97	0.98

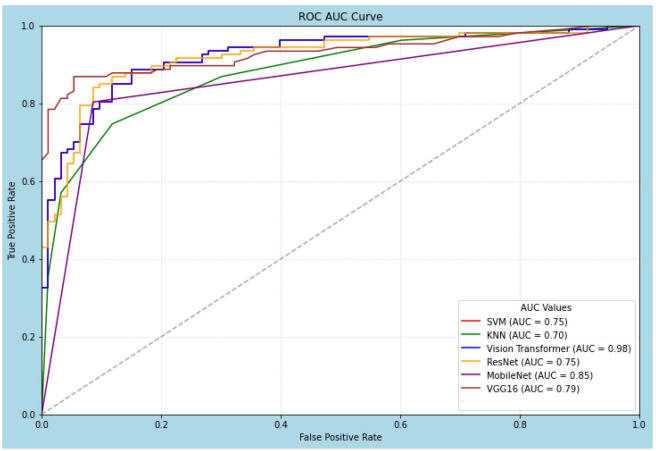

FIGURE 6. Comparative ROC-AUC analysis across benchmark methods.

Using a vision transformer as a deep learning model can effectively detect multimodal images because it can achieve high accuracy even with a constrained number of training epochs.

ViTs substantially improve over conventional manual approaches as they are less prone to human error besides faster processing. Large datasets can be processed by ViTs, allowing them to learn a variety of characteristics and patterns, boosting their robustness in dealing with real-world scenarios. Incorporation of a self-attention mechanism into the network is a crucial advantage, as it enables the network to focus on specific areas of an image instead of the entire image. This capability helps the network to identify fine-grained details in an image, which is essential for various tasks. The proposed strategy successfully achieved a classification accuracy of 99.3%. This shows that DL-based multimodal classification has the potential to increase classification accuracy. The performance of the proposed method was evaluated on different machine learning models, and a comparison between the proposed model and the various benchmark models is presented in Table 1.

The ROC-AUC compares the TPR to the FPR,and a curve above the diagonal line indicates a model with greater performance, on the other hand, indicating a model with random performance. We can assess the model's performance at various potential thresholds by analyzing the ROC curve. We compared the different benchmark methods against ROC-AUC which is shown in Fig. 6.

TABLE 2. Computational Resource Comparison

Model	Computational Resources Used
ViT	Low
SVM	Moderate
KNN	Low
MobileNet	Moderate
ResNet-152	High
VGG-16	High

TABLE 3. Hyperparameter Tuning Comparison

Model	Hyper parameter Tuning Required	Optimal Parame-ters Found	Methodology for Tuning
ViT	No	N/A	N/A
SVM	Yes	Yes	Grid Search
KNN	Yes	Yes	Cross-validation
MobileNet	Yes	Yes	Random Search
ResNet-152	Yes	Yes	Bayesian Optimization
VGG-16	Yes	Yes	Cross-validation

TABLE 4. Model Complexity Comparison

Model	Number of Layers	Computational Complexity
ViT	12	High
SVM	N/A	Low
KNN	N/A	Low
MobileNet	28	Moderate
ResNet-152	152	High
VGG-16	16	High

The combination of diverse imaging modalities and transfer learning techniques resulted in accurate and reliable - classification outcomes. Another comparison has been done based on computational resource, complexity of the model and hyper parameter tuning. Table 2 defines the level of required computational power by each model: ResNet-152 and VGG-16 as worst-case scenarios requires roughly three times more power than ViT. Table 3 reports that except for the proposed ViT model, all other models promptly adjust other hyperparameters, but the ViT model does this automatically. Methods such as grid search, cross-validation, etc. were applied for the determination of optimal model parameters of other models as well. Table 4 gives us information about the number of layers and the computational complexity that is associated with each model. On the other hand, although ViT complexity is high in the number of parameters, they are smaller than that of the SVM and KNN models with low complexity.

V. CONCLUSION

The strategic implementation of a vision transformer, coupled with acoustic sensors, for the detection of pests in cotton fields, presents a groundbreaking approach that promises to

revolutionize the precision and efficiency of pest identification. Utilizing the unique properties of a vision transformer allows the system to effectively handle large datasets, facilitating the extraction of a wide range of features and patterns. The system's adaptability improves its perseverance, ensuring its ability to effectively tackle the complexities of real-world pest detection scenarios. The vision transformer enhances the network's capacity to recognize intricate and unique features by incorporating a self-attention mechanism that enables the network to selectively focus on specific regions of an image. This is a significant step in the right direction towards better pest identification, and it highlights the need for additional research and development in this field. This technology is the finest choice for cotton pest detection and other applications because of its cost and user-friendliness.

FUTURE ENHANCEMENT

The recommended approach can smoothly integrate into a computerized system designed specifically for detecting whiteflies. This transformation presents a hopeful opportunity to simplify the labor-intensive procedure of manual inspection, greatly improving efficiency and precision. In addition, the automated ability to continuously monitor enables early detection of attacks, effectively reducing possible harm. Although the performance of the model appears favorable when evaluated using an image dataset, it is crucial to conduct real-world testing of the suggested strategy to completely determine its usefulness. Conducting real-world testing will yield vital insights into the model's performance under actual field situations and uncover any difficulties that may need to be addressed. An exciting future research direction is conducting a comparative analysis, where the proposed strategy is compared to traditional manual whitefly detection methods to determine the practical benefits of this unique methodology.

REFERENCES

[1] S. Rafi, R. Balasani, F. Qadir, M. Tahir, F. Ahmed, and W. S. Ahmad, "Role of artificial intelligence in agriculture sustainability, an example from India," in *Towards Sustainable Natural Resources: Monitoring and Managing Ecosystem Biodiversity*. Cham, Switzerland: Springer, 2022, pp. 275–293.

[2] A. I. Jajja et al., "Compact convolutional transformer (CCT)-based approach for whitefly attack detection in cotton crops," *Agriculture*, vol. 12, no. 10, Sep. 2022, Art. no. 1529.

[3] V. Singh et al., "Morpho-physiological and biochemical responses of cotton (gossypium hirsutum l.) genotypes upon sucking insect-pest infestations," *Physiol. Mol. Biol. Plants*, vol. 5, pp. 1–7, Dec. 2022.

[4] K. Ahmed, T. R. Shahidi, S. M. Alam, and S. Momen, "Rice leaf disease detection using machine learning techniques," in *Proc. IEEE Int. Conf. Sustain. Technol. Ind. 4.0*, 2019, pp. 1–5.

[5] H. Kukadiya and D. Meva, "Automatic cotton leaf disease classification and detection by convolutional neural network," in *Proc. Advance. Smart Comput. Inf. Secur.: First Int. Conf.*, 2022, pp. 247–266..

[6] A. Parikh, M. S. Raval, C. Parmar, and S. Chaudhary, "Disease detection and severity estimation in cotton plant from unconstrained images," in *Proc. IEEE Int. Conf. Data Sci. Adv. Anal.*, 2016, pp. 594–601.

[7] A. M. Soomro et al., "Forecasting cotton whitefly population using deep learning," *J. Comput. Biomed. Informat.*, vol. 4, no. 1, pp. 64–76, Dec. 2022.

[8] H. Du, H. X. Xu, F. Wang, L. X. Qian, S. S. Liu, and X. W. Wang, "Armet from whitefly saliva acts as an effector to suppress plant defences by targeting tobacco cystatin," *New Phytologist*, vol. 234, no. 5, pp. 1848–62, Jan. 2022.

[9] X. Sun et al., "Vitellogenin facilitates associations between the whitefly and a bacteriocyte symbiont," *Mbio*, vol. 14, Jan. 2023, Art. no. e02990–22.

[10] P. Deepika and B. Arthi, "Prediction of plant pest detection using improved mask FRCNN in cloud environment," *Meas.: Sensors*, vol. 24, Dec. 2022, Art. no. 100549.

[11] T. Pallapothu, M. Singh, R. Sinha, H. Nangia, and P. Udawant, "Cotton leaf disease detection using mask RCNN," in *Proc. AIP Conf. Proc.*, 2022, vol. 2393, no. 1, Art. no. 020114.

[12] C. U. Parab et al., "Comparison of single-shot and two-shot deep neural network models for whitefly detection in IoT web application," *AgriEngineering*, vol. 4, no. 2, pp. 507–22, Jun. 2022.

[13] J. Shin, M. Mahmud, T. U. Rehman, P. Ravichandran, B. Heung, and Y. K. Chang, "Trends and prospect of machine vision technology for stresses and diseases detection in precision agriculture," *AgriEngineering*, vol. 5, no. 1, pp. 20–39, Mar. 2023.

[14] P. S. Thakur, P. Khanna, T. Sheorey, and A. Ojha, "Explainable vision transformer enabled convolutional neural network for plant disease identification: PlantXViT," 2022, *arXiv:2207.07919*.

[15] H. T. Thai, K. H. Le, and N. L. Nguyen, "FormerLeaf: An Efficient Vision Transformer for Cassava Leaf Disease Detection," *Comput. Electron. Agriculture*, vol. 204, Jan. 2023, Art. no. 107518.

[16] S. Yu, L. Xie, and Q. Huang, "Inception convolutional vision transformers for plant disease identification," *Internet Things*, vol. 21, Apr. 2023, Art. no. 100650.

[17] H. Alshammari, K. Gasmi, I. Ben Ltaifa, M. Krichen, L. Ben Ammar, and M. A. Mahmood, "Olive disease classification based on vision transformer and CNN models," *Comput. Intell. Neurosci.*, vol. 2022, Jul. 2022, Art. no. 3998193.

[18] X. Li, X. Li, S. Zhang, G. Zhang, M. Zhang, and H. Shang, "SLViT: Shuffle-convolution-based lightweight Vision transformer for effective diagnosis of sugarcane leaf diseases," *J. King Saud Univ.-Comput. Inf. Sci.*, vol. 35, no. 6, 2023, Art. no. 101401, doi: 10.1016/j.jksuci.2022.09.013.

[19] F. Wang et al., "Practical cucumber leaf disease recognition using improved Swin Transformer and small sample size," *Comput. Electron. Agriculture*, vol. 199, Aug. 2022, Art. no. 107163.

[20] E. Li, L. Wang, Q. Xie, R. Gao, Z. Su, and Y. Li, "A novel deep learning method for maize disease identification based on small sample-size and complex background datasets," *Ecological Inform.*, vol. 75, 2023, Art. no. 102011, doi: 10.1016/j.ecoinf.2023.102011.

[21] A. Mittal and H. Gupta, "An experimental evaluation in plant disease identification based on activation-reconstruction generative adversarial network," in *Proc. IEEE 2nd Int. Conf. Adv. Comput. Innov. Technol. Eng.*, 2022 pp. 361–366.

[22] M. A. Khan et al., "Cucumber leaf diseases recognition using multi level deep entropy-ELM feature selection," *Appl. Sci.*, vol. 12, no. 2, Jan. 2022, Art. no. 593.

[23] K. Khairnar and N. Goje, "Image processing based approach for diseases detection and diagnosis on cotton plant leaf," in *Proc. 2nd Int. Conf. Adv. Technol. Societal Appl.-Volume 1*, 2018, pp. 55–65.

[24] G. P. Bhattarai, R. B. Schmid, and B. P. McCornack, "Remote sensing data to detect hessian fly infestation in commercial wheat fields," *Sci. Rep.*, vol. 9, no. 1, pp. 1–8, Apr. 2019.

[25] H. Abid, N. Nida, and A. Irtaza, "PestinaNet-A real-time crop pest detection system," in *Proc. IEEE 2nd Int. Conf. Comput. Mach. Intell.*, 2022, pp. 1–4.

[26] H. Yu, J. Liu, C. Chen, A. A. Heidari, Q. Zhang, and H. Chen, "Optimized deep residual network system for diagnosing tomato pests," *Comput. Electron. Agriculture*, vol. 195, Apr. 2022, Art. no. 106805.

[27] R. Manavalan, "Towards an intelligent approaches for cotton diseases detection: A review," *Comput. Electron. Agriculture*, vol. 200, 2022, Art. no. 107255.

[28] K. S. Kumar and A. Abdul Rahman, "Early detection of locust swarms using deep learning," in *Proc. Adv. Mach. Learn. Comput. Intell.: ICMLCI*, 2019, pp. 303–310.

[29] S. Zhang, R. Jing, and X. Shi, "Crop pest recognition based on a modified capsule network," *Syst. Sci. Control Eng.*, vol. 10, no. 1, pp. 552–61, Dec. 2022.

[30] H. T. Thai, N. Y. Tran-Van, and K. H. Le, "Artificial cognition for early leaf disease detection using vision transformers," in *Proc. IEEE Int. Conf. Adv. Technol. Commun.*, 2021, pp. 33–38.

REMYA S received the Ph.D. degree in computer science and engineering from the Vellore Institute of Technology, Vellore, India. She is currently an Assistant Professor with the Department of Computer Science and Engineering, School of Computing, Amrita Vishwa Vidyapeetham, Kollam, India. Her research interests include deep learning, data science, computer vision, and smart environments.

ANJALI T received the Master of Technology degree in computer science and engineering in 2013 from the Amrita Vishwa Vidyapeetham, Kollam, India, where she is currently working toward the Ph.D. degree in computer science and engineering. She is an esteemed Assistant Professor with the Department of Computer Science and Engineering, Amrita Vishwa Vidyapeetham, a highly regarded educational institution based in India. Her research interests include image processing, computer vision, and machine learning.

ABHISHEK S is currently working toward the undergraduate degree in computer science and engineering with the Amrita School of Computing, Amrita Vishwa Vidyapeetham, Kollam, India. He has gained valuable experience in the field of cybersecurity as a former Digital Forensics Analyst with teamBi0s, India's leading Cyber Security Research CTF team. He is also a Data Engineer with Rolls-Royce, demonstrating his remarkable versatility in the field. His research interests include data engineering, the Internet of Things (IoT), machine learning, deep learning, and cybersecurity.

SOMULA RAMASUBBAREDDY received the Ph.D. degree in computer science and engineering from the Vellore Institute of Technology University, Vellore, India, in 2022. He is currently an Assistant Professor with the Department of Information Technology, Vallurupalli Nageswara Rao Vignana Jyothi Institute of Engineering and Technology, Hyderabad, India. He completed Postdoctoral research work with the Department of Information and Communication, Sunchon National University, Suncheon, South Korea, in 2024. He has authored or coauthored more than 40 publications in reputed journals and conferences. His research interests include mobile cloud computing, IoT, machine learning, and edge computing.

YONGYUN CHO received the Ph.D. degree in computer engineering from Soongsil University, Seoul, South Korea. He is currently a Professor with the Department of Information and Communication Engineering, Sunchon National University, Suncheon, South Korea. His main research interests include system software, embedded software, and ubiquitous computing.