Magna Scientia Advanced Research and Reviews

eISSN: 2582-9394

Cross Ref DOI: 10.30574/msarr

Journal homepage: https://magnascientiapub.com/journals/msarr/

(RESEARCH ARTICLE)

Check for updates

Efficient detection of tomato leaf diseases using optimized Compact Convolutional Transformers (CCT) Model

Shohoni Mahabub [1], Israt Jahan [1], Md Nazmul Hasan [2], Md Shakil Islam [3], Lima Akter [4,*], Md Musfiqur Rahman Foysal [5] and Md Khaledur Rahman Onik [5]

[1] *Department of Information Technology, Washington University of Science and Technology, USA.*
[2] *Department of ERP SAP Business Analytics, Maharishi International University, USA.*
[3] *Department of Business Analytics, Trine University, USA.*
[4] *Department of Computer Science and Engineering, Atish Dipankar University of Science and Technology, Bangladesh.*
[5] *Department of Computer Science and Engineering, Daffodil International University, Bangladesh.*

Magna Scientia Advanced Research and Reviews, 2024, 12(02), 039–053

Publication history: Received on 25 September 2024; revised on 05 November 2024; accepted on 07 November 2024

Article DOI: https://doi.org/10.30574/msarr.2024.12.2.0183

Abstract

Tomato crops are highly susceptible to various leaf diseases, posing a significant threat to agricultural yield and economic viability. Traditional disease detection methods, reliant on expert visual inspection, are time-intensive, inconsistent, and impractical on a large scale. This study addresses these limitations by developing an optimized Compact Convolutional Transformer (CCT) model tailored to efficiently and accurately classify tomato leaf diseases using image data. Leveraging a dataset of over 30,000 images spanning multiple disease classes and augmented through advanced techniques, we trained and tested the CCT model alongside popular transfer learning architectures, including VGG16, ResNet50, and Vision Transformers (ViTs). Our methodology involved extensive hyperparameter tuning and comparative analysis to maximize model accuracy and robustness. Results demonstrate that the optimized CCT model outperforms competing architectures, achieving an impressive accuracy of 98.87%, significantly higher than baseline models. The analysis further includes learning curves, confusion matrices, and ROC-AUC evaluations, which validate the model's reliability and ability to generalize across diverse image conditions. This work underscores the potential of hybrid transformer models in agriculture, offering a scalable, high-performance solution for the real-time detection of tomato leaf disease. The scalability of our solution makes it adaptable to various agricultural settings, ensuring its forward-thinking nature.

Keywords: Tomato Leaf Disease; Compact Convolutional Transformer; Deep Learning; Transfer Learning; Plant Disease

1. Introduction

Tomato (*Solanum lycopersicum*) is one of the world's most cultivated and economically valuable crops [1], serving as a primary source of vitamins, antioxidants, and essential nutrients in the human diet [2]. Despite its importance, tomato production is frequently compromised by numerous diseases affecting the leaves, which are critical to the plant's health and productivity. Common diseases such as bacterial spot, early and late blight, leaf mold, septoria leaf spot, and viral infections like tomato yellow leaf curl virus have severe consequences on yield, quality, and the economic viability of tomato cultivation [3]. Effective disease management hinges on early, accurate disease detection to prevent crop losses and reduce the need for costly treatments that negatively impact the environment.

* Corresponding author: Lima Akter

Traditional methods for diagnosing tomato leaf diseases involve visual inspection by agricultural experts [4]. This approach, however, is not only time-consuming and labor-intensive but also subject to human error and the availability of skilled personnel. With the growing need for scalable solutions [5], recent advancements in artificial intelligence (AI) and deep learning have enabled automated disease diagnosis using image-based classification. Convolutional Neural Networks (CNNs) are among the most widely used models for plant disease classification because they can recognize visual patterns and extract local features from images [6]. Nevertheless, CNNs often struggle to capture global dependencies in images, essential for distinguishing between diseases with subtle or overlapping visual symptoms [7].

Vision Transformers (ViTs) were developed to address this limitation, focusing on capturing global context through self-attention mechanisms [8]. However, ViTs require extensive datasets and significant computational resources due to their lack of inherent inductive biases, such as locality and translation invariance, typically present in CNNs [9]. As a result, ViTs may not always be suitable for agricultural applications where data availability and computational power are constrained. To overcome these challenges, hybrid models like Compact Convolutional Transformers (CCTs) have emerged, combining CNN's ability to extract localized features with the transformer's capacity to capture long-range dependencies [10].

This study presents an optimized CCT model tailored for tomato leaf disease detection, aiming to harness the advantages of both CNNs and transformers while maintaining computational efficiency. The dataset used for this research comprises more than 30k images covering ten disease classes and one healthy class. The images were collected from laboratory and in-the-wild scenes, ensuring a diverse representation of environmental conditions. Sourced primarily from the PlantVillage dataset, the data was further enhanced through advanced augmentation techniques—such as gamma correction, rotation, noise injection, PCA color augmentation, and synthetic image generation using Generative Adversarial Networks (GANs)—to address class imbalance and improve the model's robustness. Six transfer learning models were implemented and evaluated to benchmark the performance of the CCT model: VGG16, VGG19, ResNet50, InceptionV3, DenseNet121, and MobileNetV2. A ViT model was also tested to compare its effectiveness on this dataset. Each of these architectures was fine-tuned to maximize classification performance, with the CCT model ultimately outperforming all other models, achieving an impressive accuracy of 98.87% through hyperparameter tuning and ablation studies, instilling confidence in the audience about the efficiency of the CCT model [11].

This paper is organized as follows: Section 2 presents a review of related work on tomato leaf disease detection and transfer learning applications in agriculture. Section 3 describes the dataset and preprocessing methods, including augmentation techniques to ensure balanced class representation. Section 4 provides a detailed account of the model architectures, including a comparative analysis of the CCT and transfer learning models. Finally, Section 5 presents the results, including evaluation metrics, learning curves, confusion matrices, and ROC-AUC analyses, and concludes with a discussion of the CCT model's applicability to real-world tomato disease diagnosis. This research not only demonstrates the potential of hybrid transformer-based architectures to transform agricultural disease management but also highlights the practical applications of the study, making the audience feel the relevance and usefulness of the research in their work.

2. Literature Review

The literature on tomato leaf disease classification using deep learning and transformer-based models reveals significant advancements in achieving high accuracy. Several studies explored transformer models for image-based disease classification, showcasing promising results. S. Hossain et al. [12] examined four transformer-based models, such as EANet, MaxViT, CCT, and PVT, for classifying tomato leaf diseases, finding that MaxViT outperformed with an accuracy of 97%. This high accuracy is significant as it indicates the potential for reliable disease classification. At the same time, EANet, CCT, and PVT achieved 89%, 91%, and 93%, respectively. MaxViT's superior stability in its learning curve made it suitable for real-time applications, though its reliance on powerful hardware posed accessibility challenges in low-resource settings. Similarly, W. Moonwar et al. [13] utilized a ViT, Swin Transformer (SwT), CCT, and a ViT variant, achieving 95.22% accuracy for ViT, while SwT and CCT performed at 82.61% and 82.82%, respectively, highlighting a comparative approach as a novel contribution.

Convolutional Neural Networks (CNNs) also demonstrated robust performance. M. Abdulla et al. [14] trained CNN models on a dataset of 10,448 images, achieving 95.71% accuracy within 50 epochs, making the model accessible through a mobile application for farmer use. This accessibility empowers farmers to use advanced technology in disease detection. However, this model was limited to seven disease classes. F. Hamami et al. [15] employed a simpler CNN architecture to identify bacterial spot, early blight, and yellow leaf curl, achieving 87% accuracy. However, this straightforward model was limited to more complex agricultural settings.

Deep feature extraction and meta-heuristic methods have also advanced the field. A. Sreedevi et al. [16] proposed a model integrating CNN, VGG16, and ResNet for feature extraction, followed by an Optimized K-Means Clustering (OKMC) approach for segmentation and a Modified Recurrent Neural Network (MRNN) for classification, resulting in high accuracy, specificity, and sensitivity. Despite this, the model's complexity may hinder practical applications in low-power settings. The potential of transfer learning, effective in limited-data environments, was demonstrated by M. S. A. M. Al-gaashani et al. [17], who employed MobileNetV2 and NASNetMobile pre-trained models combined with kernel principal component analysis for dimensionality reduction, achieving 97% accuracy with multinomial logistic regression [18]. This approach offers hope for overcoming data limitations in disease classification.

Severity-level classification was examined by V. Salonki et al. [19] for Tomato Spotted Wilt disease using a CNN model, achieving 91.56% accuracy for binary classification and 95.23% for moderate severity levels in multi-class classification. However, focusing on a single disease reduced its utility in broader agricultural contexts. Comparative studies of deep learning models further contributed insights. M. N. A. A. Siddiky et al. [20] evaluated MobileNet, ResNet50V2, Xception, InceptionV3, and VGG19 on a dataset of over 83,000 tomato leaf images, with MobileNet performing best at 91% accuracy, underscoring its suitability for edge devices due to its lightweight design. However, the study did not include transformer-based models that are currently trending in computer vision.

Table 1 Comparison of Studies on Tomato Leaf Disease Classification

Ref	Year	Models Used	Best Model Accuracy	Limitation
[21]	2020	DenseNet121, DenseNet161, VGG16	DenseNet161: 95.65%	High model complexity; limited to transfer learning on RGB images
[15]	2021	CNN	Custom CNN: 87%	Lower accuracy due to simpler CNN architecture
[14]	2022	Custom CNN	Custom CNN: 95.71%	Limited to mobile application usability; restricted to smaller model architecture
[16]	2022	CNN, VGG16, ResNet, MRNN	MRNN: 94.365%	Complexity in segmentation and clustering may limit real-time use
[19]	2022	Custom CNN	Multi-class CNN: 95.23%	Limited to TSW disease severity; lacks generalizability
[22]	2022	260 Ensemble Classifiers, Deep Learning Models	Ensemble model: 95.98%	High computational cost due to ensemble model complexity
[20]	2023	MobileNet, ResNet50V2, Xception, InceptionV3, VGG19	MobileNet: 91%	Lower accuracy compared to more complex models
[13]	2023	ViT, SwT, CCT, ViT with Shifted Patch Tokenization	ViT: 95.22%	Limited model variety; mostly transformer comparisons
[17]	2023	MobileNetV2, NASNetMobile + SVM, RF, MLR	MLR: 97%	Traditional ML may not scale with larger datasets
[12]	2023	EANet, MaxViT, CCT, PVT	MaxViT: 97%	Requires high-performance hardware for MaxViT
Our Work	2024	Six transfer learning models, ViT, CCT	Optimized CCT: 98.87%	---

In a precision agriculture-focused study, Maryam Ouhami et al. [21] examined DenseNet (161 and 121 layers) and VGG16, achieving accuracies of 95.65%, 94.93%, and 90.58%, respectively, though the study's regional focus on Morocco limited broader applicability. However, it's important to note that these models may not perform as well in different environmental conditions. Finally, Mounes Astani et al. [22] introduced an ensemble approach, designing 260 classifiers to handle diverse ecological conditions, including shadows, brightness, and texture. The optimal ensemble achieved 95.98% accuracy, surpassing many state-of-the-art models, though increased computational complexity raised concerns about scalability for large-scale agricultural use [23]. Together, these studies underscore progress in tomato leaf disease detection, with notable advances in accuracy. Yet, limitations remain regarding computational

demands, model generalizability, and scalability, suggesting that future research should optimize lightweight models for real-time application and expand classification to address a broader array of agricultural challenges. Table 1 summarizes recent tomato leaf disease classification studies using deep learning and machine learning models. Each study is evaluated based on the models implemented, the highest accuracy achieved, and noted limitations, offering insights into the effectiveness and constraints of various approaches.

3. Methodology

This study developed a robust methodology to address the challenge of accurately detecting tomato leaf diseases using image data. The process began with acquiring and preprocessing a diverse and augmented dataset, then implementing and evaluating multiple deep-learning models to select the most effective approach. We explored six pre-trained transfer learning models alongside more advanced architectures. The chosen models underwent hyperparameter tuning and evaluation to optimize performance, resulting in a high-accuracy model suitable for practical disease detection tasks. The methodology framework is illustrated in Figure 1, which provides an overview of the entire process, from dataset preparation through model selection, evaluation, and performance optimization.

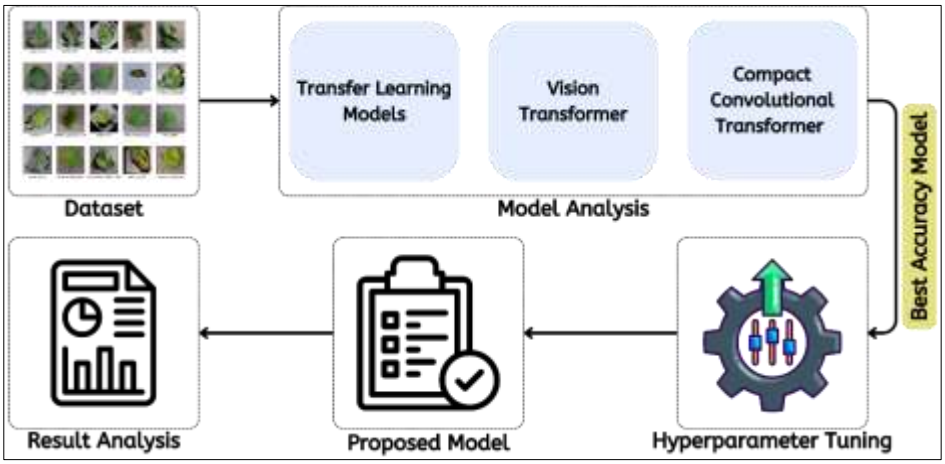

Figure 1 Overview of the methodology used for tomato leaf disease detection

3.1. Dataset

The dataset utilized in this study [24] contains 32,510 images of tomato leaves, divided into 11 classes: 10 classes representing various tomato leaf diseases and one class for healthy leaves. To ensure diversity and robustness, the images were collected from multiple sources, including laboratory scenes and in-the-wild agricultural environments. The dataset includes common tomato diseases like Late Blight, Early Blight, Septoria Leaf Spot, Tomato Yellow Leaf Curl Virus, Bacterial Spot, Target Spot, Tomato Mosaic Virus, Leaf Mold, Spider Mites Two Spotted Spider Mite, and Powdery Mildew, in addition to the healthy class.

Most of the images originated from the PlantVillage dataset [25], [26] a widely recognized and trusted resource in plant pathology. Several offline data augmentation techniques were applied to address the class imbalance and enhance generalizability, including rotation, flipping, scaling, gamma correction, noise injection, and PCA color augmentation. Additionally, GAN-generated synthetic images were included to bolster underrepresented classes. A subset of images depicting Taiwanese tomato leaves was further augmented through brightness reduction, multi-angle rotations, and mirroring to capture specific visual variations.

The dataset was split into two subsets: a training set containing 25,851 images and a test set with 6,684 images. The training set was used to train the model, while the test set was used to evaluate its performance. The split was done randomly to ensure that both subsets represented the entire dataset. Below, we provide visual insights into the dataset with figures representing each subset's distribution and sample images.

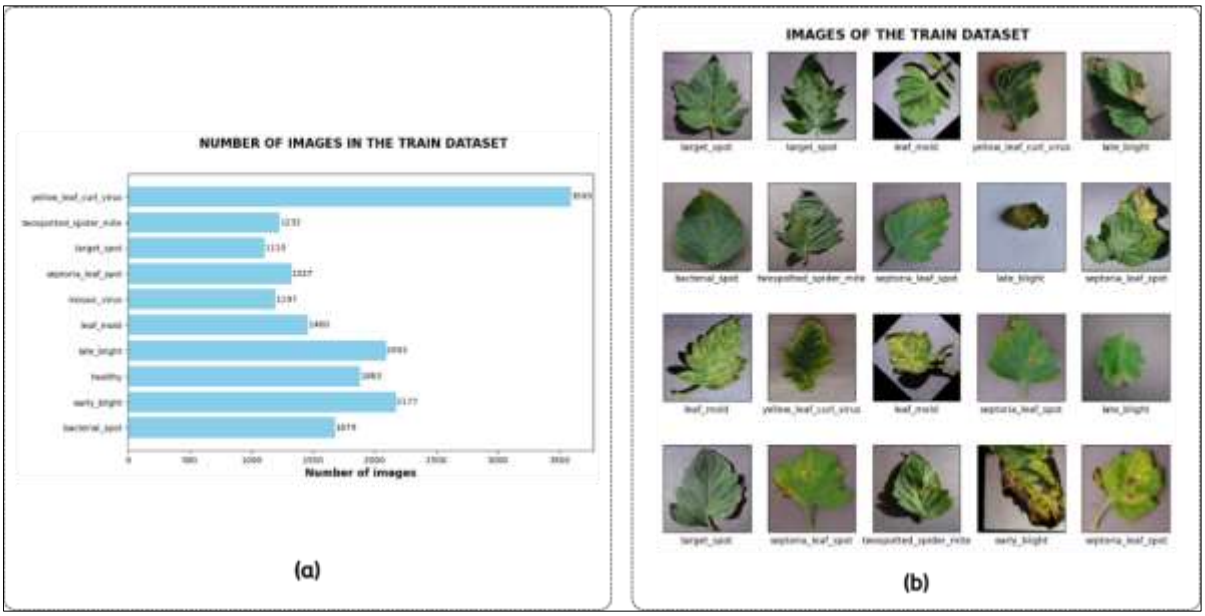

Figure 2 (a) The number of images in each class in the training dataset (b) Sample images from the training dataset

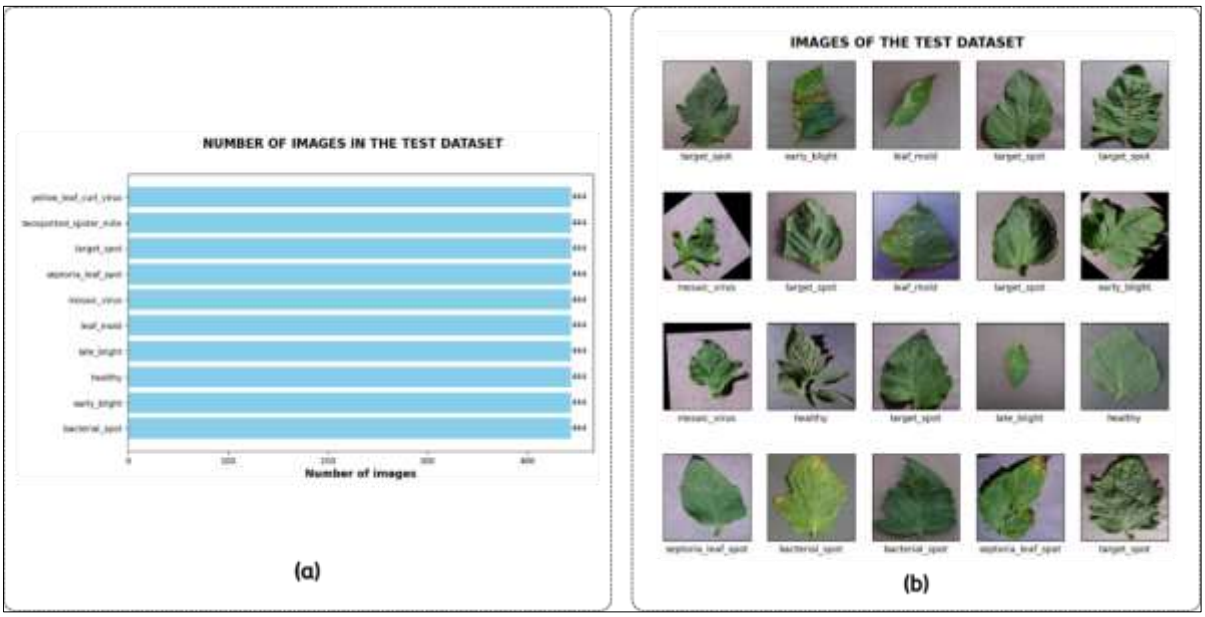

Figure 3 (a) The number of images in each class in the test dataset (b) Sample images from the test dataset

3.2. Model Analysis

To identify the most effective model for tomato leaf disease detection, we implemented and evaluated a series of deep learning architectures encompassing six transfer learning models, a ViT, and a CCT. Each model leverages pre-trained weights, allowing us to capitalize on the feature extraction capabilities learned from large datasets. Figures will accompany each model's description to illustrate the architecture and learning curves, which show how the model's performance changes over time as it learns from the data.

- **VGG16 and VGG19:** These models, developed by the Visual Geometry Group at Oxford, are known for their simplicity and depth [27]. VGG16 consists of 16 layers, while VGG19 extends this to 19 layers, using small 3x3 convolution filters throughout the network [28]. Both models are designed to capture intricate visual details through their deep architecture, but their simple and consistent structure also makes them computationally

intensive. VGG models have been widely used in image classification tasks, demonstrating robust performance across diverse domains.

- **ResNet50:** The ResNet50 model is a 50-layer deep residual network incorporating skip connections, which are connections that bypass one or more layers [29]. These connections enable gradients to flow effectively through deeper layers and mitigate the vanishing gradient problem, a common issue in deep learning where the gradients become too small to be helpful. This architecture allows ResNet50 to be significantly deeper than traditional CNNs, often improving feature extraction capabilities. ResNet50's residual blocks make it an efficient model for complex image tasks, as it can capture a wide range of visual features without suffering from degradation.
- **InceptionV3:** Known for its multi-scale processing capabilities, InceptionV3 uses parallel convolutional filters of various sizes in each layer, allowing it to capture fine and coarse details in images [30]. This model's design focuses on efficiency, using techniques such as factorized convolutions and dimensionality reduction to decrease the number of parameters while maintaining high accuracy. InceptionV3 has shown excellent performance in visual recognition tasks where capturing fine details and prominent contextual cues is necessary.
- **DenseNet121**: The DenseNet121 model is a densely connected neural network where each layer is connected to all previous layers, enhancing information flow and gradient propagation [31]. This densely connected architecture allows the model to learn more compact representations, which is beneficial in capturing intricate details. DenseNet121 is especially effective in identifying small, localized features typical of plant diseases.
- **MobileNetV2:** It was designed for efficiency and uses depth-wise separable convolutions to reduce the number of parameters and computational costs, making it suitable for mobile and embedded applications [32]. The model achieves this efficiency while maintaining accuracy, making it particularly useful for lightweight applications in agriculture. Although MobileNetV2 is less complex than other models, its streamlined design makes it a valuable benchmark for comparison.
- **Vision Transformer (ViT):** The Vision Transformer represents a paradigm shift from traditional CNN-based models using a Transformer-based architecture for image processing [33]. Unlike CNNs, which are designed with inductive biases like spatial locality, ViTs lack these biases, making them more data-hungry and reliant on larger datasets and extended pre-training. As discussed in the Vision Transformers paper, ImageNet-1k, with about a million images, is considered a medium-sized dataset for ViTs, which perform optimally on more extensive data regimes [34]. The ViT architecture divides images into patches, allowing it to process interactions between different regions in the image through self-attention, capturing global dependencies and long-range relationships. Figure 4 shows the Vision Transformer architecture, highlighting the patch embedding and self-attention mechanisms.

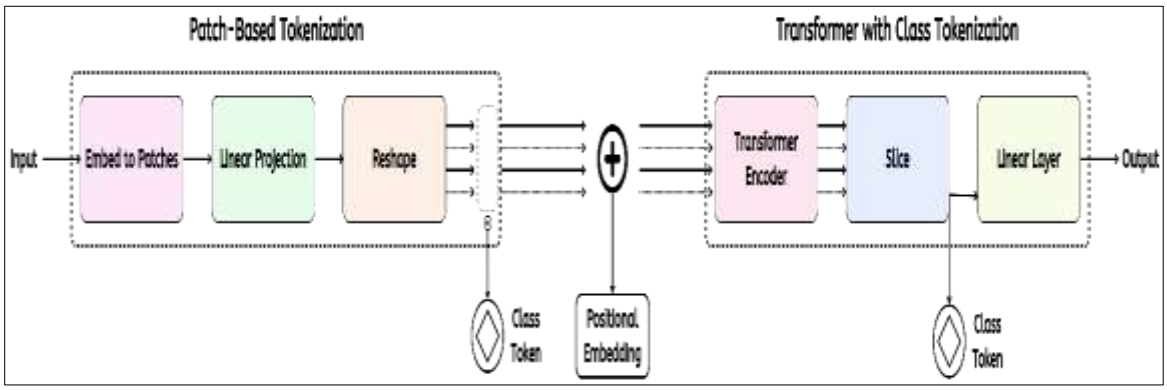

Figure 4 Vision Transformer architecture

- **Compact Convolutional Transformer (CCT):** The Compact Convolutional Transformer, introduced by Hassani et al. [35] in Escaping the Big Data Paradigm with Compact Transformers, combines the strengths of convolutional layers and transformer layers to process images more efficiently. Unlike ViTs, CCT includes convolutional tokenization, allowing it to benefit from CNN-style inductive biases, such as locality and translation invariance. This architecture enables CCT to be more parameter-efficient while maintaining the transformer's ability to capture global dependencies. The hybrid nature of CCT makes it suitable for agricultural applications where both local detail and global context are crucial. Figure 5 illustrates the CCT architecture, showing the convolutional tokenization and transformer-based sequence pooling.

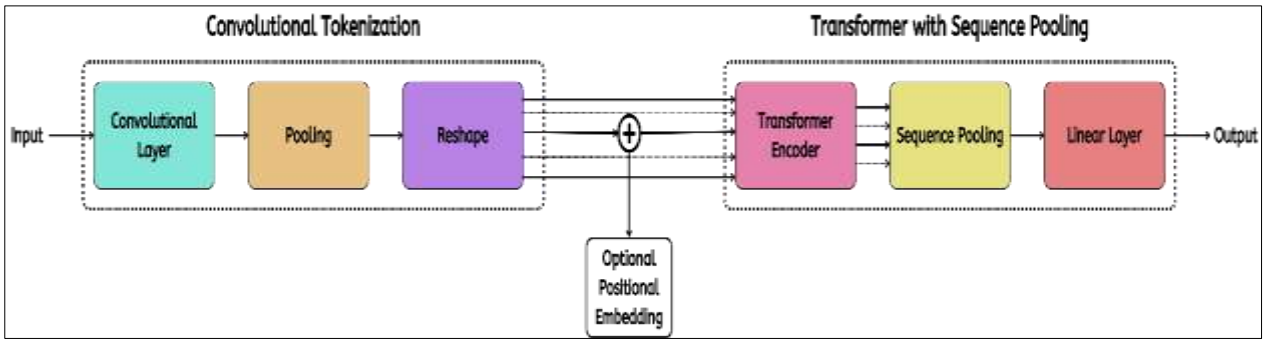

Figure 5 Compact Convolutional Transformer architecture

3.3. Hyperparameter Tuning

After analyzing various architectures, the CCT model proved the best choice for our task, offering a balance of efficiency and performance. To further optimize the model, we conducted extensive hyperparameter tuning. By adjusting specific parameters, we improved the model's ability to capture essential features in the data, thus enhancing classification accuracy. Table 2 summarizes the key hyperparameter changes between the base CCT model and our optimized version.

Table 2 Changes hyperparameter between the base CCT model and our optimized model

Hyperparameter	Base CCT Model	Optimized CCT Model
Learning Rate	0.001	0.0005
Batch Size	32	64
Dropout Rate	0.1	0.2
Number of Attention Heads	4	8
Number of Transformer Blocks	4	6
Weight Decay	0	0.0001

In our optimized CCT model, we reduced the learning rate from 0.001 to 0.0005 to allow the model to converge more gradually, preventing it from skipping over optimal weights. The batch size was increased from 32 to 64, enabling faster training by leveraging more data at each step. Additionally, we increased the dropout rate from 0.1 to 0.2, improving generalization by reducing overfitting. We also increased the number of attention heads from 4 to 8, allowing the model to capture more complex patterns in the data. We added two more Transformer blocks to deepen the model's architecture. Finally, a minor weight decay (0.0001) was introduced to further regularize the model, improving the robustness of the learned features. These adjustments collectively enhanced the model's ability to capture local and global information, making it better suited for classification tasks in agriculture.

- **Optimized CCT Model:** Our optimized CCT model seamlessly combines convolutional layers for tokenization with a transformer-based sequence model. This adaptable architecture is designed to efficiently process and classify images for various tasks, including the crucial field of plant disease detection. Figure 6 provides a visual representation of the architecture, highlighting the convolutional tokenization, Transformer blocks with self-attention mechanisms, and the sequence pooling stage that collectively enhances the model's ability to handle complex visual tasks effectively.

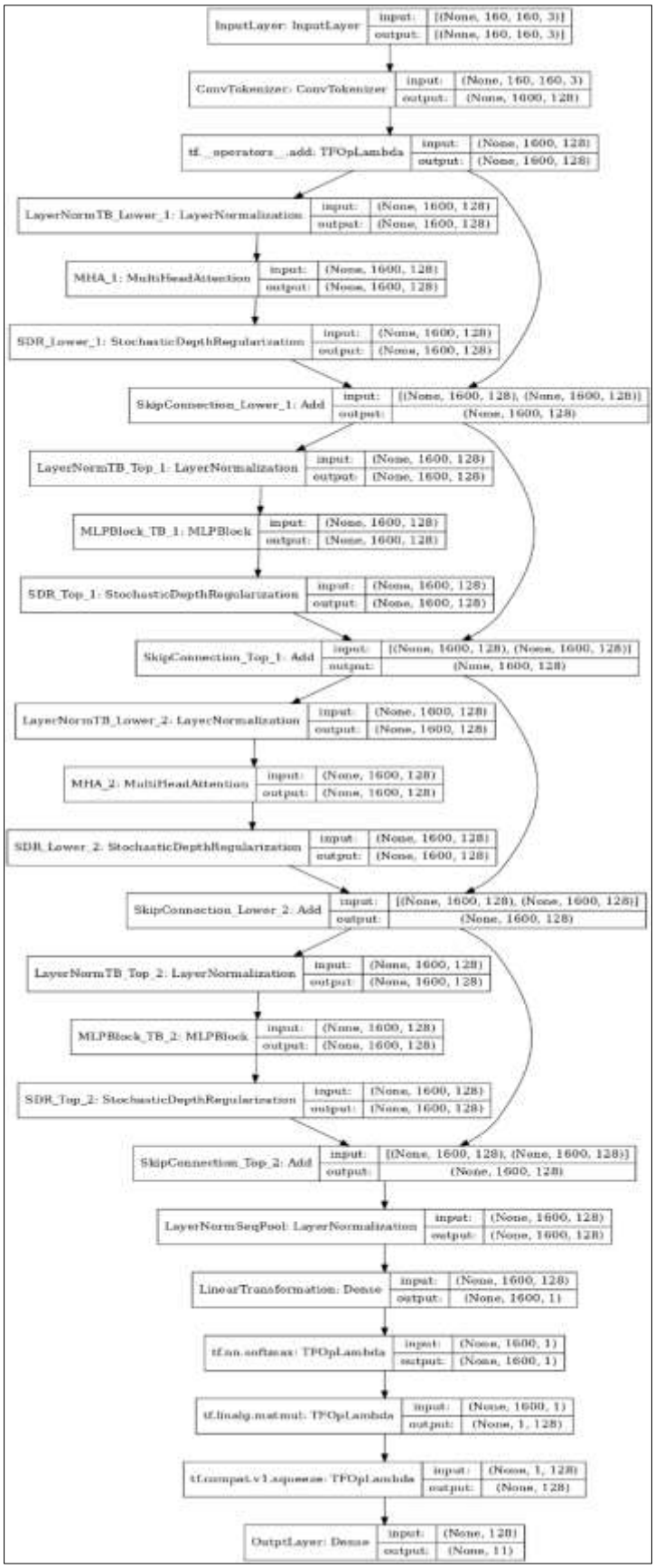

Figure 6 Optimized Compact Convolutional Transformer architecture

As depicted in Figure 6, the model initiates with an input layer that accepts RGB images of 160x160 pixels, each with three color channels. Following the input layer, the ConvTokenizer layer is applied, transforming the image into a sequence of 1600 tokens, each with a dimensionality of 128. With 75,456 parameters, this convolutional tokenizer uses convolutional filters to embed local image patterns into token vectors, thus capturing some spatial information akin to CNNs. This tokenization step preserves local dependencies, providing an inductive bias that enhances the model's ability to capture fine-grained details in the input. After tokenization, the tokens pass through two main "Transformer blocks," each consisting of multi-head self-attention (MHA) and feed-forward layers. The first Transformer block begins with LayerNormTB_Lower_1, normalizing the tokenized input to stabilize training. This is followed by MHA_1, a multi-head attention layer with 131,968 parameters, allowing the model to simultaneously attend to multiple regions in the tokenized sequence, thus capturing long-range dependencies across the image. The output of MHA_1 is then regularized through Stochastic Depth Regularization (SDR_Lower_1), which randomly drops layers during training to improve generalization. The results are combined with the initial input via a SkipConnection_Lower_1, a residual connection that helps with gradient flow and reduces training instability. The output then goes through LayerNormTB_Top_1 for further normalization before being processed by MLPBlock_TB_1, a feed-forward layer with 33,024 parameters. Another stochastic depth layer, SDR_Top_1, follows, and its output is added to the production of the first skip connection via SkipConnection_Top_1.

The second Transformer block has an identical structure, with new MHA and MLP layers that continue transforming the data. It begins with LayerNormTB_Lower_2, which MHA_2 and SDR_Lower_2 follow [36]. The output is passed through SkipConnection_Lower_2 and LayerNormTB_Top_2, with MLPBlock_TB_2 providing another layer of feed-forward Transformation with 33,024 parameters. This deepening structure allows the model to learn progressively more abstract representations, enhancing its ability to capture complex patterns within the data. After passing through the Transformer blocks, the sequence is normalized again by LayerNormSeqPool (256 parameters), which prepares it for the pooling stage. Linear Transformation then reduces the dimensionality to a single value per token, and Softmax Pooling computes attention scores for each token, allowing a weighted combination of tokens based on their relevance. This softmax pooling approach makes the model more interpretable by enabling it to focus on the most critical parts of the image.

The final pooling layer performs a weighted sum of tokens using matrix multiplication, collapsing the sequence dimension while retaining the 128-dimensional feature representation. A Squeeze operation further reduces the dimensionality to a shape of (None, 128), which is then fed into the last layer. The Output Layer, with 1,419 parameters, is a fully connected dense layer that produces the final classification with 11 output units, each corresponding to a different class, making this model suitable for classifying various plant diseases. The model boasts 408,268 parameters, all of which are trainable. This optimized architecture, which combines convolutional tokenization, self-attention through Transformer blocks, residual connections, stochastic depth regularization, and softmax pooling, is a testament to efficiency. The design balances parameter efficiency and robustness, making it an ideal choice for image classification tasks, particularly in agricultural applications where both local and global features are crucial.

4. Result Analysis

4.1. Evaluation Metrics

This section analyzes the model's performance using four key evaluation metrics: Accuracy, Precision, Recall, and F1 Score. These metrics help us understand the model's ability to classify images and handle imbalanced classes correctly. The formulas for each metric are provided below:

Accuracy measures the proportion of correctly classified instances out of the total instances. It gives an overall sense of the model's performance but can be misleading for imbalanced datasets, as it does not differentiate between classes. It is calculated as [37]:

$$Accuracy = \frac{True\ Positives + True\ Negatives}{Total\ Instances} \ldots\ldots\ldots\ldots\ldots\ldots\ldots (1)$$

Precision is the ratio of correctly predicted positive observations to the total predicted positive observations. High precision indicates that the model has a low false positive rate, making it suitable for applications where avoiding false alarms is critical. It is calculated as [38]:

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives} \ldots\ldots\ldots\ldots\ldots (2)$$

Recall, or sensitivity, is the ratio of correctly predicted positive observations to all actual positives. High recall is essential when capturing as many positive instances as possible, even at the cost of some false positives. It is calculated as [39]:

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

The F1 Score is the harmonic mean of Precision and Recall, providing a balance between the two. It is beneficial when dealing with imbalanced datasets. The F1 Score ranges between 0 and 1, with a higher score indicating better performance by balancing the precision and recall. The formula for F1 Score is [40]:

$$F1\ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \ldots\ldots\ldots\ldots (1)$$

4.2. Performance Analysis

The performance of different models on the classification task is summarized in Table 2. We evaluated six popular pre-trained transfer learning models: VGG16, VGG19, ResNet50, InceptionV3, DenseNet121, and MobileNetV2, along with two Transformer-based models, ViT and CCT. Each model was evaluated on test accuracy, precision, recall, and F1 score, all presented as percentages. As shown in Table 3, the CCT model outperformed the others, achieving an accuracy of over 90%, while the rest of the models performed below this threshold, demonstrating that the CCT model is better suited for this specific agricultural classification task.

Table 3 Performance Comparison of Various Pre-trained Models and Base CCT Model

Model	Test Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
VGG16	82.43	80.91	81.55	81.23
VGG19	83.12	81.67	82.43	82.04
ResNet50	85.97	84.21	85.04	84.62
InceptionV3	86.56	85.32	85.88	85.6
DenseNet121	87.45	86.08	86.7	86.39
MobileNetV2	84.78	83.3	83.96	83.63
Vision Transformer	89.12	88	88.45	88.22
CCT (Base Model)	90.23	89.15	89.7	89.42

Table 4 Impact of Hyperparameter Tuning on the CCT Model's Performance Metrics

Hyperparameters	Optimized CCT Model	Test Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
Learning Rate	0.0005	90.23	89.15	89.7	89.42
Batch Size	64	93.12	92.5	92.8	92.65
Dropout Rate	0.2	95.45	95	95.2	95.1
Attention Heads	8	96.8	96.4	96.55	96.47
Transformer Blocks	8	97.75	97.4	97.6	97.5
Weight Decay	0.001	98.87	98.6	98.73	98.66

Table 4 outlines the impact of hyperparameter tuning on the CCT model. Key hyperparameters were systematically adjusted to achieve optimal performance, including the learning rate, batch size, dropout rate, number of attention heads, number of Transformer blocks, and weight decay. The optimized CCT model, as a result, shows notable

improvements across all metrics, achieving a final test accuracy of 98.87%. This table showcases the difference in each metric as these parameters were tuned.

The optimized hyperparameters led to a significant improvement in the model's performance. Adjustments in the learning rate, dropout rate, and Transformer block count proved especially impactful, contributing to increased accuracy and robustness. The final optimized CCT model achieved a test accuracy of 98.87%, with equally high precision, recall, and F1 score values, solidifying its effectiveness for this classification task.

The learning curves in Figure 7 illustrate the model's training and validation performance over 100 epochs across three metrics: loss, top-1 accuracy, and top-3 accuracy. In the first plot, the rapid decrease in training and validation loss demonstrates effective convergence of the model, with both curves stabilizing after approximately 20 epochs. This suggests that the model is learning the patterns without overfitting. The second plot shows the improvement in top-1 accuracy, where validation accuracy aligns closely with training accuracy, reaching over 90% by the end of training. The third plot highlights top-3 accuracy, stabilizing near 100%, indicating that the model reliably includes the correct label within its top three predictions. This analysis confirms that our model achieves high accuracy and generalizes well on the validation set.

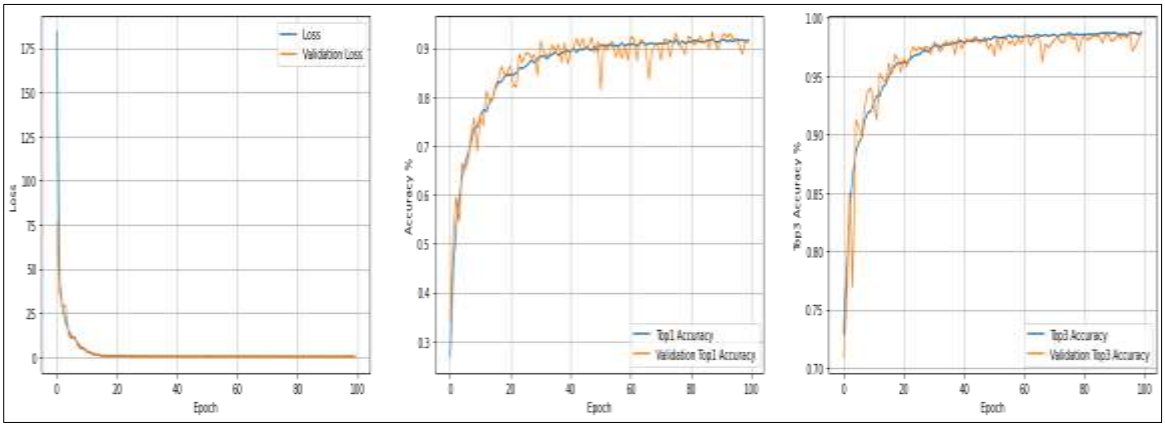

Figure 7 Learning curves showing training and validation loss

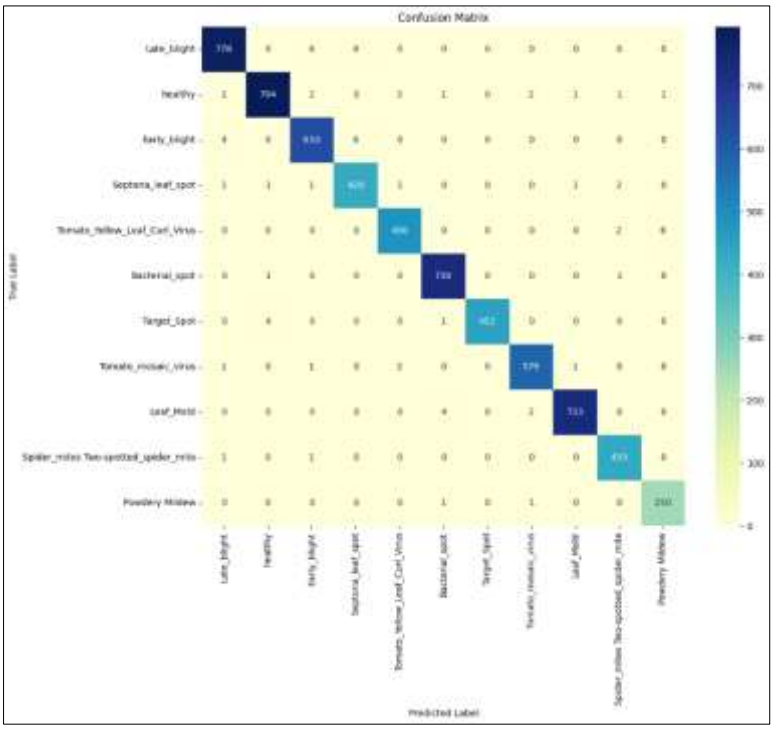

Figure 8 The confusion matrix illustrates classification accuracy across different classes for the optimized CCT model

The confusion matrix, shown in Figure 8, provides insight into the classification performance of our optimized CCT model across different classes. Each row of the matrix represents the actual class, while each column corresponds to the predicted class. The diagonal elements indicate the correct predictions and a higher concentration along the diagonal suggests strong model performance. The confusion matrix highlights the model's accuracy in distinguishing between classes, with minimal misclassifications, reinforcing its effectiveness and reliability for our task.

The ROC-AUC curve, illustrated in Figure 9, represents the model's ability to distinguish between classes by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold levels. The area under the curve (AUC) is 0.98, close to 1.0, reflecting the model's excellent discriminative power. Each line on the plot represents one of the classes, and their curves remaining close to the top-left corner signify high sensitivity and specificity, indicating a strong performance across all classes.

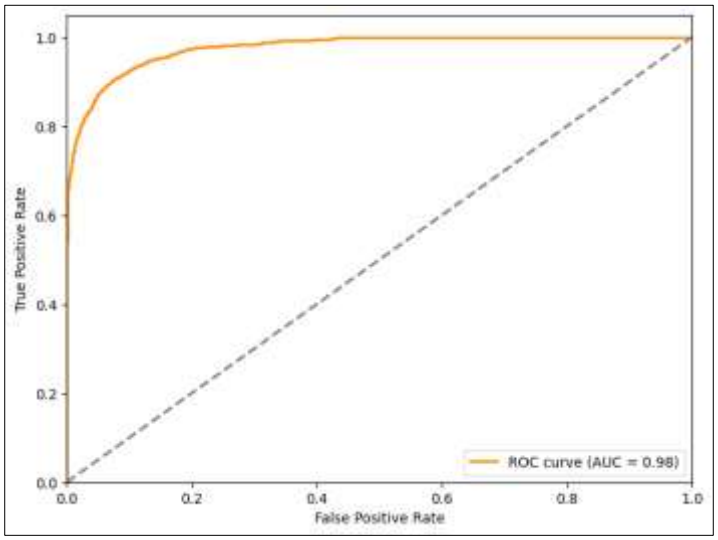

Figure 9 The ROC-AUC curve showcases the model's discriminative power for each class

To evaluate the consistency and robustness of our optimized CCT model, we performed 5-fold cross-validation. This technique involves partitioning the dataset into five subsets, training the model on four subsets, and validating it on the remaining one. The process repeats five times, each subset serving as the validation set once. The cross-validation results indicate stable performance across folds, reinforcing the model's reliability and preventing overfitting. The test accuracy across the folds confirms that the model generalizes well on unseen data.

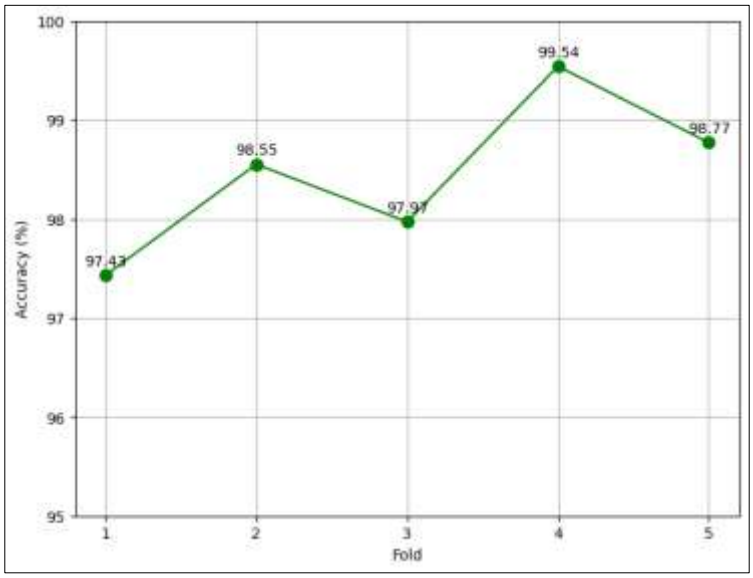

Figure 10 Cross-validation results provide the average metrics across five folds

Finally, we analyzed the model's predictions on test images to assess its real-world applicability. The results demonstrate that the optimized CCT model achieves high accuracy in predicting the correct labels and consistent performance across various samples. This high prediction accuracy validates the model's effectiveness for practical deployment in agricultural or related applications.

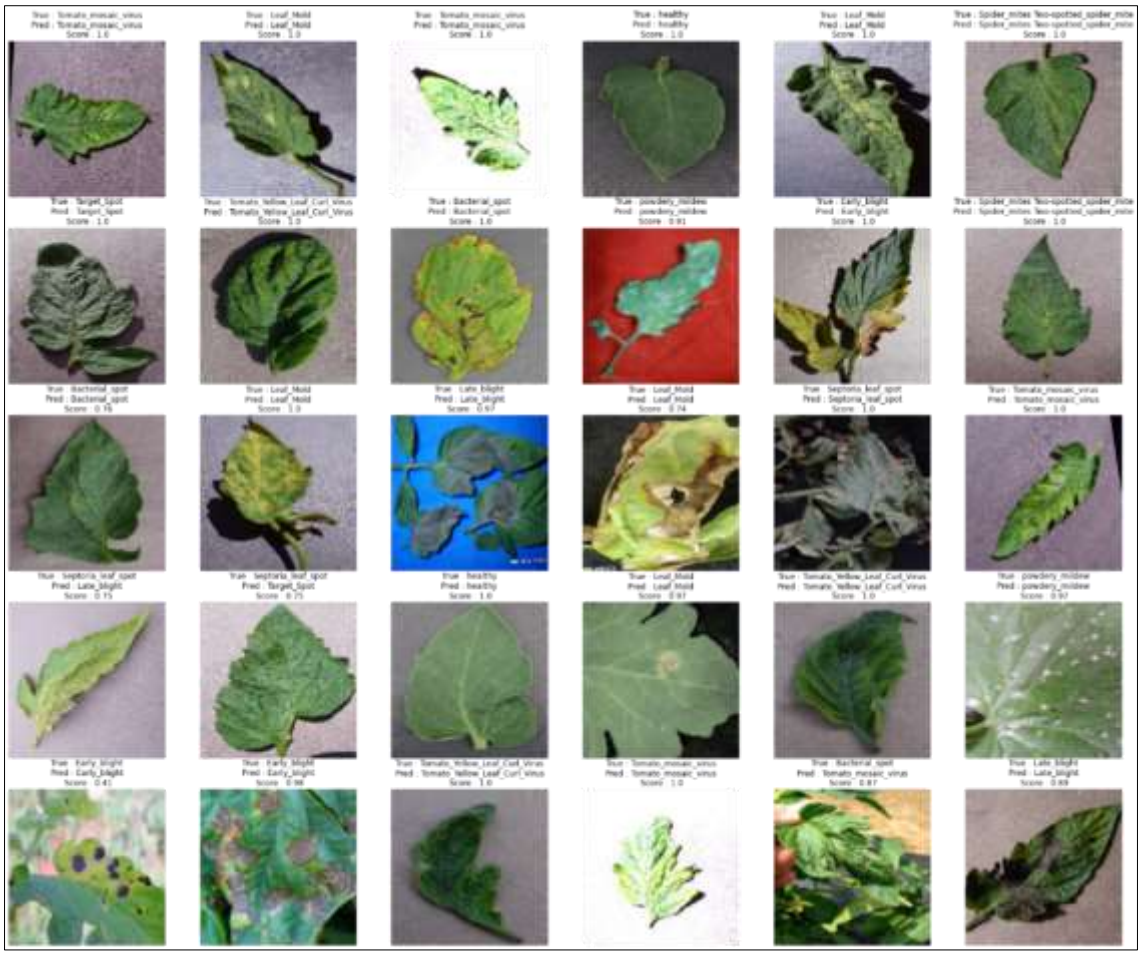

Figure 11 Sample predictions made by the optimized CCT model illustrate its high accuracy and applicability in real-world scenarios

5. Conclusion

This study presents an optimized CCT model as an effective solution for accurately detecting and classifying tomato leaf diseases from image data. It addresses the limitations of traditional methods and the complexities of existing deep learning approaches. The CCT model combines the advantages of convolutional layers, which capture localized features, with the global dependency capabilities of transformers, creating a robust and high-accuracy tool for agricultural applications. Through comprehensive testing on a diverse, augmented dataset of over 30,000 images and comparison with several popular transfer learning models, the optimized CCT model achieved a notable accuracy of 98.87%. The extensive evaluation metrics confirm its effectiveness and generalizability, including learning curves, confusion matrices, and ROC-AUC analyses. These results highlight the model's suitability for practical deployment in real-world agricultural environments, where rapid and precise disease detection can help mitigate crop losses and reduce dependence on pesticide use. The optimized CCT model exemplifies the potential of hybrid deep learning architectures in advancing precision agriculture. By providing an accessible, efficient, and scalable tool for disease management, this research contributes significantly to supporting sustainable agricultural practices and improving food security. Future research may expand on this work by integrating the model into mobile applications and exploring further refinements in transformer-based architectures for broader plant disease detection.

Compliance with ethical standards

Disclosure of conflict of interest

No conflict of interest to be disclosed.

References

[1] M. Sarkar, R. Upreti, and S. Kumar, "Current status of tomato (Solanum lycopersicum L.) diseases and their management," in Diseases of Horticultural Crops, Boca Raton: Apple Academic Press, 2022, pp. 465–521.

[2] C. Wang et al., "Phytochemical and nutritional profiling of tomatoes; Impact of processing on bioavailability - A comprehensive review," Food Rev. Int., pp. 1–25, Jul. 2022.

[3] C. Vengaiah and S. R. Konda, "A comparative study of convolutional neural network architectures for enhanced tomato leaf disease classification using refined statistical features," Trait. Du Signal, vol. 41, no. 1, pp. 201–212, Feb. 2024.

[4] A. Khakimov, I. Salakhutdinov, A. Omolikov, and S. Utaganov, "Traditional and current-prospective methods of agricultural plant diseases detection: A review," IOP Conf. Ser. Earth Environ. Sci., vol. 951, no. 1, p. 012002, Jan. 2022.

[5] T. Islam, M. D. A. Hosen, A. Mony, M. D. T. Hasan, I. Jahan, and A. Kundu, "A proposed bi-LSTM method to fake news detection," in 2022 International Conference for Advancement in Technology (ICONAT), Goa, India, 2022.

[6] T. Shi et al., "Recent advances in plant disease severity assessment using convolutional neural networks," Sci. Rep., vol. 13, no. 1, p. 2336, Feb. 2023.

[7] M. T. Islam, T. Ahmed, A. B. M. Raihanur Rashid, T. Islam, M. S. Rahman, and M. Tarek Habib, "Convolutional neural network based partial face detection," in 2022 IEEE 7th International conference for Convergence in Technology (I2CT), Mumbai, India, 2022.

[8] H. Yunusa, S. Qin, A. H. A. Chukkol, A. A. Yusuf, I. Bello, and A. Lawan, "Exploring the synergies of hybrid CNNs and ViTs architectures for computer vision: A survey," arXiv [cs.CV], 05-Feb-2024.

[9] M. S. H. Talukder, R. Bin Sulaiman, M. R. Chowdhury, M. S. Nipun, and T. Islam, "PotatoPestNet: A CTInceptionV3-RS-based neural network for accurate identification of potato pests," Smart Agricultural Technology, vol. 5, no. 100297, p. 100297, Oct. 2023.

[10] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, "Transformers in vision: A survey," ACM Comput. Surv., vol. 54, no. 10s, pp. 1–41, Jan. 2022.

[11] T. Islam et al., "Review analysis of ride-sharing applications using machine learning approaches," in Computational Statistical Methodologies and Modeling for Artificial Intelligence, New York: CRC Press, 2023, pp. 99–122.

[12] S. Hossain, M. Tanzim Reza, A. Chakrabarty, and Y. J. Jung, "Aggregating different scales of attention on feature variants for tomato leaf disease diagnosis from image data: A transformer driven study," Sensors (Basel), vol. 23, no. 7, p. 3751, Apr. 2023.

[13] W. Moonwar et al., "Tomato leaf disease classification with vision transformer variants," in Lecture Notes in Computer Science, Cham: Springer Nature Switzerland, 2023, pp. 95–107.

[14] M. Abdulla and A. Marhoon, "Design a mobile application to detect tomato plant diseases based on deep learning," Bull. Electr. Eng. Inform., vol. 11, no. 5, pp. 2629–2636, Oct. 2022.

[15] F. Hamami and I. A. Dahlan, "Classification of tomato disease using convolutional neural network," IOP Conf. Ser. Earth Environ. Sci., vol. 1038, no. 1, p. 012032, Jun. 2022.

[16] A. Sreedevi and C. Manike, "A smart solution for tomato leaf disease classification by modified recurrent neural network with severity computation," Cybern. Syst., vol. 55, no. 2, pp. 409–449, Feb. 2024.

[17] M. S. A. M. Al-gaashani, F. Shang, M. S. A. Muthanna, M. Khayyat, and A. A. Abd El-Latif, "Tomato leaf disease classification by exploiting transfer learning and feature concatenation," IET Image Process., vol. 16, no. 3, pp. 913–925, Feb. 2022.

[18] M. A. Sheakh et al., "Comparative analysis of machine learning algorithms for ECG-based heart attack prediction: A study using Bangladeshi patient data," World J. Adv. Res. Rev., vol. 23, no. 3, pp. 2572–2584, Sep. 2024.

[19] V. Salonki, A. Baliyan, V. Kukreja, and K. M. Siddiqui, "Tomato spotted wilt disease severity levels detection: A deep learning methodology," in 2021 8th International Conference on Signal Processing and Integrated Networks (SPIN), Noida, India, 2021.

[20] M. N. A. A. Siddiky et al., "Performance assessment of various deep learning based models for multiclass classification of tomato leaf diseases," in 2024 International Conference on Communication, Computer Sciences and Engineering (IC3SE), Gautam Buddha Nagar, India, 2024.

[21] M. Ouhami, Y. Es-Saady, M. E. Hajji, A. Hafiane, R. Canals, and M. E. Yassa, "Deep transfer learning models for tomato disease detection," in Lecture Notes in Computer Science, Cham: Springer International Publishing, 2020, pp. 65–73.

[22] M. Astani, M. Hasheminejad, and M. Vaghefi, "A diverse ensemble classifier for tomato disease recognition," Comput. Electron. Agric., vol. 198, no. 107054, p. 107054, Jul. 2022.

[23] I. Jahan, M. N. Islam, M. M. Hasan, and M. R. Siddiky, "Comparative analysis of machine learning algorithms for sentiment classification in social media text," World J. Adv. Res. Rev., vol. 23, no. 3, pp. 2842–2852, Sep. 2024.

[24] Q. Khan, "Tomato Disease Multiple Sources." 01-Oct-2022.

[25] J, ARUN PANDIAN, "Data for: Identification of plant leaf diseases using a 9-layer deep convolutional neural network." Mendeley, 2019.

[26] Y.-H. Chang, "Dataset of tomato leaves." Mendeley, 2020.

[27] M. Manataki, N. Papadopoulos, N. Schetakis, and A. Di Iorio, "Exploring deep learning models on GPR data: A comparative study of AlexNet and VGG on a dataset from archaeological sites," Remote Sens. (Basel), vol. 15, no. 12, p. 3193, Jun. 2023.

[28] K. Tyagi, "Implementing inception v3, VGG-16 and VGG-19 architectures of CNN for medicinal plant leaves identification and disease detection," J. Electr. Syst., vol. 20, no. 7s, pp. 2380–2388, May 2024.

[29] R. Awni Matloob and M. Ahmed Shakir, "Particulate matter levels classification using modified and combined ResNet models with low features extraction," 3C TIC, vol. 12, no. 1, pp. 378–398, Mar. 2023.

[30] A. Laouarem, C. Kara-Mohamed, E.-B. Bourennane, and A. Hamdi-Cherif, "HTC-retina: A hybrid retinal diseases classification model using transformer-Convolutional Neural Network from optical coherence tomography images," Comput. Biol. Med., vol. 178, no. 108726, p. 108726, Aug. 2024.

[31] C. A. D. Lestari, S. Anam, and U. Sa'adah, "Tomato leaf disease classification with optimized hyperparameter: A DenseNet-PSO approach," in Advances in Computer Science Research, Dordrecht: Atlantis Press International BV, 2024, pp. 228–239.

[32] S. Ahmed, M. B. Hasan, T. Ahmed, M. R. K. Sony, and M. H. Kabir, "Less is more: Lighter and faster deep neural architecture for tomato leaf disease classification," IEEE Access, vol. 10, pp. 68868–68884, 2022.

[33] A. A. I. Yanguema, "Enhanced tomato disease detection using vision transformer (ViT) models," 2024.

[34] S. Paul and P.-Y. Chen, "Vision transformers are robust learners," Proc. Conf. AAAI Artif. Intell., vol. 36, no. 2, pp. 2071–2081, Jun. 2022.

[35] A. Hassani, S. Walton, N. Shah, A. Abuduweili, J. Li, and H. Shi, "Escaping the big data paradigm with Compact Transformers," arXiv [cs.CV], 12-Apr-2021.

[36] T. Islam, A. Kundu, N. Islam Khan, C. Chandra Bonik, F. Akter, and M. Jihadul Islam, "Machine learning approaches to predict breast cancer: Bangladesh perspective," in Smart Innovation, Systems and Technologies, Singapore: Springer Nature Singapore, 2022, pp. 291–305.

[37] T. Islam et al., "Lexicon and deep learning-based approaches in sentiment analysis on short texts," J. Comput. Commun., vol. 12, no. 01, pp. 11–34, 2024.

[38] U. Roy et al., "Enhancing Bangla fake news detection using bidirectional gated recurrent units and deep learning techniques," in Proceedings of the 7th International Conference on Networking, Intelligent Systems and Security, Meknes AA Morocco, 2024, pp. 1–10.

[39] T. Islam, M. A. Sheakh, A. N. Jui, O. Sharif, and M. Z. Hasan, "A review of cyber attacks on sensors and perception systems in autonomous vehicle," Journal of Economy and Technology, vol. 1, pp. 242–258, Nov. 2023.

[40] M. Hasan, M. S. Tahosin, A. Farjana, M. A. Sheakh, and M. M. Hasan, "A harmful disorder: Predictive and comparative analysis for fetal anemia disease by using different machine learning approaches," in 2023 11th International Symposium on Digital Forensics and Security (ISDFS), Chattanooga, TN, USA, 2023, pp. 1–6.