Technological Applied Sciences
ISSN: 1308 7258
Article ID: 2A0208

Status : Research Article
Received: 27.03.2025
Accepted: 25.07.2025

Büşra Er
Erzincan Binali Yıldırım University
busraer7@gmail.com, Erzincan-Türkiye
Volkan Kaya
Erzincan Binali Yıldırım University
vkaya@erzincan.edu.tr, Erzincan-Türkiye

DOI	http://dx.doi.org/10.12739/NWSA.2025.20.3.2A0208	
ORCID ID	0009-0009-4255-2800	0000-0001-6940-3260
Corresponding Author	Büşra Er	

A COMPARATIVE STUDY OF CNN AND TRANSFORMER-BASED DEEP LEARNING MODELS FOR TEA LEAF DISEASE RECOGNITION

ABSTRACT

This study presents a comparative analysis of six deep learning models for the automatic classification of eight different disease categories found in tea leaves. The dataset used in the study was divided into three parts with 70% training, 15% validation, and 15% testing ratios. As part of the experimental evaluation, five convolutional neural network (CNN) based architectures (ResNet50, DenseNet121, EfficientNet-B0, MobileNetV3-Large, and ConvNeXt-Tiny) and one Transformer-based model (Vision Transformer, ViT-Small) were tested using the same training strategies. The models were trained using a transfer learning and fine-tuning approach; performance metrics were reported based on accuracy, precision, recall, and F1-score values. In addition, the number of parameters and the prediction time per image were calculated for each model. Experimental results show that the DenseNet121 model achieved the highest success rate in the validation dataset, while the ConvNeXt-Tiny architecture achieved the highest accuracy and F1-score values in the standalone test dataset. The findings indicate that modern CNN-based architectures offer high generalization capabilities in the classification of tea leaf diseases. The results obtained serve as a comparative reference for future studies in the field of agricultural image analysis.

Keywords: Tea Leaf Disease Classification, Deep Learning, Convolutional Neural Networks, Vision Transformer, Plant Disease Detection

1. INTRODUCTION

Tea (Camellia sinensis) is a strategic agricultural product that forms the raw material for one of the most widely consumed beverages globally. With daily consumption reaching billions of cups, tea has become both a product of high economic value and a primary source of income for millions of producers. Tea production is highly sensitive to biotic stress factors, particularly those targeting the leaf tissue. Fungal pathogens, bacterial infections, and leaf pests reduce photosynthetic efficiency, leading to significant losses in both yield and product quality. Some studies in the literature have reported that tea diseases can reduce annual yields by approximately 20% in large-scale plantations [1].

Effective management of tea leaf diseases depends on the ability to diagnose them early and accurately. However, traditional diagnostic methods rely heavily on visual inspections carried out by specialists in the field. This process has limitations, such as being open to

How to Cite:
Er, B. and Kaya, V., (2025). A comparative study of cnn and transformer-based deep learning models for tea leaf disease recognition. Technological Applied Sciences, 20(3):79-93, DOI: 10.12739/NWSA.2025.20.3.2A0208.

subjective interpretation, requiring high labor costs, and having low applicability in large agricultural areas. In field conditions, variable light intensity, overlapping leaves, shade, rain marks, and complex background structures can negatively affect the accuracy of identification by the human eye [2 and 3]. For these reasons, many symptoms are overlooked in the early stages, leading to rapid disease spread and consequently greater productivity losses [1 and 2].

In recent years, artificial intelligence and image processing-based approaches have become a significant research area in the automated detection of agricultural diseases. Deep learning models, particularly convolutional neural networks (CNNs), significantly improve classification performance by extracting features from raw images [1, 2, and 3]. Furthermore, Transformer-based architectures such as Vision Transformer (ViT) offer an alternative approach to plant disease detection thanks to their ability to model long-range dependencies [2]. Studies on the detection of tea leaf diseases show that lightweight CNN architectures such as EfficientNet, DenseNet, and MobileNet provide high accuracy rates [3]. Furthermore, various hybrid approaches are presented in the literature. This approach has resulted in higher accuracy and computational efficiency compared to standalone models [2].

These studies demonstrate that deep learning offers a powerful solution for classifying tea leaf diseases, that segmentation-supported models improve accuracy by reducing background noise, that lightweight architectures are suitable for mobile and field-based applications, and that hybrid models combine performance and cost in a balanced way. In this context, systematic comparison of different architectural approaches serves as an important reference for future methods in agricultural image analysis. Therefore, there is growing interest in deep learning-based methods for the automated and highly accurate detection of tea leaf diseases. Deep learning models overcome the limitations of classical expert-based diagnostic processes, enabling real-time scanning of large areas and reducing yield losses through early diagnosis [1, 2, and 3].

This study comprehensively evaluates six different deep learning models for classifying eight disease categories observed in tea leaves. The analysis compares five CNN-based architectures ResNet50, DenseNet121, EfficientNet-B0, MobileNetV3-Large, and ConvNeXt-Tiny and the Vision Transformer (ViT-Small) model. All models were evaluated using the same training, validation, and testing separation; accuracy, recall, precision, and F1-score were reported as key performance metrics. In addition, computational characteristics such as model complexity, number of parameters, and estimation time were also analyzed. The study aims to reveal the performance differences of different architectural approaches in classifying tea leaf diseases and to serve as a comparative reference for future agricultural image analysis applications.

While current studies show that different deep learning strategies are applicable in the detection of tea leaf diseases, comprehensive comparisons between models appear to be limited. The majority of studies in the literature report the performance of a single architecture or evaluate it based on a limited number of models. Furthermore, the comparability of results is reduced because the dataset, hyperparameters, training protocols, and data augmentation strategies vary within the scope of the study. This situation makes it difficult to clearly identify the strengths and weaknesses of different architectural approaches in tea leaf disease classification.

On the other hand, while segmentation-supported methods and hybrid structures have been shown to have positive effects on accuracy enhancement, the relationship between mobile device applications, real-time prediction performance, and model complexity with classification

success has not yet been systematically analyzed. Although Transformer-based visual models (Vision Transformers) are known to yield positive results in object classification, the literature does not clearly report how these architectures perform compared to CNN-based models in datasets with natural field conditions such as tea leaf diseases. In this context, evaluating CNN and Transformer architectures on the same dataset under equal training conditions offers an important opportunity for comparison regarding which architectural structure is more suitable for agricultural image analysis.

In line with these requirements, this study aims to evaluate six different modern deep learning architectures for the classification of tea leaf diseases on the same dataset, with the same hyperparameter settings and training protocol. Model performance is examined not only with singular metrics such as accuracy, but also with multidimensional criteria such as F1-score, class-based error analysis, model parameter size, and prediction time. In this way, a comparative reference framework is provided for both accuracy-oriented and computationally cost-oriented decision-making processes.

The findings of this study are expected to contribute to the design of real-time diagnostic systems at the field scale, the development of mobile farming applications, and the optimization of future tea disease detection models.

This study offers the following contributions to the literature on image-based classification of tea leaf diseases:

- A comprehensive comparison of six modern architectures representing different deep learning approaches was performed on the same dataset: five CNN-based models (ResNet50, DenseNet121, EfficientNet-B0, MobileNetV3-Large, ConvNeXt-Tiny) and one Transformer-based model (Vision Transformer – ViT-Small).
- All models were evaluated using a uniform training protocol. Dataset splitting (70% training, 15% validation, 15% testing), hyperparameters, and optimization method steps were kept the same across all models. This approach provides a fair and reproducible evaluation infrastructure.
- Model performance was reported using multiple metrics such as accuracy, class-based precision, sensitivity, and F1-score. In addition, confusion matrices were visualized to show the error distributions
- Parameter size and prediction time per image were calculated for each architecture, thus examining the relationship between model complexity and classification performance. This evaluation presents findings aimed at establishing a balance between model performance and computational cost.
- To improve the explainability of the model's decision-making mechanisms, Grad-CAM-based visual descriptions were generated, and the image regions underlying the classification decisions were analyzed.
- The results provide a comparative reference framework for tea leaf disease detection; It provides descriptive and measurable outputs that can guide agricultural image analysis studies.

The remaining sections of the study are structured as follows: The second section summarizes and evaluates current studies on disease detection and classification in tea leaves. The third section details the materials and methods of the study, including the dataset used, preprocessing steps, and deep learning-based models. The fourth section presents the findings obtained from the experiments and discusses the relevant analyses. Finally, the fifth section includes the main conclusions of the study and offers suggestions for future research.

2. RESEARCH SIGNIFICANCE

This study addresses the need for reliable and objective identification of tea leaf diseases under real field conditions, where traditional expert-based diagnosis is time-consuming, subjective, and difficult to scale. The main objective is to comparatively evaluate CNN-based and Transformer-based deep learning architectures for tea leaf disease recognition using the same dataset and identical training conditions. The study hypothesizes that modern CNN architectures provide stronger generalization than Transformer-based models in agricultural image classification tasks with limited data. Experimental results support this hypothesis, showing that CNN-based models, particularly ConvNeXt-Tiny, outperform the Vision Transformer in terms of test accuracy and F1-score. The findings offer practical guidance for selecting efficient deep learning models for real-time agricultural monitoring and precision farming applications.

Highlights:

- A fair and systematic comparison of CNN and Transformer-based models for tea leaf disease recognition is presented.
- CNN-based architectures demonstrate stronger generalization performance than Vision Transformer under limited data conditions.
- The results provide practical guidance for real-time and mobile agricultural disease detection systems.

3. RELATED WORKS

Research on the automated detection of tea leaf diseases has recently focused on deep learning-based approaches.

Ahmed et al., in their comprehensive study on tea leaf disease classification, reported that the combined use of segmentation and classification methods contributed to improved performance. The study presented a method in which the leaf region was separated using the Segment Anything Model (SAM), and the symptom areas were then processed by a CNN-based classifier. The authors reported that this method of noise reduction improved the learning process and increased the accuracy in EfficientNet-based classification from 82% to 95.58%. The findings show that segmentation-supported models can provide an accuracy increase of approximately 10–13 points compared to raw image input [2].

Lightweight and mobile-focused methods are also being investigated in tea leaf disease detection studies. In this study, a wavelet-based lightweight CNN architecture called WaveLiteNet was proposed. Within the scope of the research, it was stated that the model achieved an accuracy rate of 98.7%, obtained high sensitivity values and offered stable prediction performance even in complex scenarios. In addition, it was stated that the model is suitable for real-time operation on mobile devices thanks to its low computational cost [1].

Approaches based on multi-model integration are also found in the literature. The combination of YOLOv7-based object detection and CNN classifier was evaluated as a hybrid approach and reported to have achieved the highest accuracy value on the dataset. The study indicates that the balance between performance and computational cost favors hybrid architectures [4].

Similarly, Transformer-based approaches have also begun to be applied in tea leaf disease detection. In the IEM-ViT-based model proposed by Zhang et al., the ViT architecture was combined with masking and extensive data augmentation strategies, achieving 93.78% accuracy and an approximate F1 score of 0.94 in classifying seven different tea diseases. The study reports that the ViT-based model provides significant advantages in both accuracy and F1-score compared to ResNet18 and VGG-derived CNN architectures [5].

The YOLO-Tea model was designed to improve tea leaf disease and pest detection by adding ACmix, CBAM, RFB, and GCNet modules to the YOLOv5 architecture. Through the integration of attention mechanisms and receptive field expansion strategies, the model's feature extraction success was enhanced while maintaining low resource consumption. Experimental results show that YOLO-Tea provides higher performance than YOLOv5 and other common object detection approaches, supporting the model's suitability for real-world applications [6].

Finally, in studies aimed at the automated identification of tea leaf diseases, segmentation and generative contrast network-based approaches are used to remove noise and reduce data imbalance in raw field images obtained under complex plantation conditions. In this context, methods combining conditional generative adversarial networks (IC-GAN) enhanced with two-stage image segmentation allow for the separation of disease regions from the background and the expansion of the dataset with synthetic samples. Segmentation strategies integrating graph cuts and support vector machines (SVMs) report significant improvements in recognition accuracy. Furthermore, generating synthetic disease images with IC-GAN improves classification performance by increasing the coverage of the training dataset. Studies using specialized deep learning architectures such as Inception Embedded Pooling Convolutional Neural Network (IDCNN) in the disease recognition task show that high accuracy, recall, and F1 values are obtained in three different tea disease types. Comparisons across different datasets reveal that segmentation and GAN-based data augmentation approaches contribute to the development of robust, highly accurate diagnostic systems adapted to field conditions [7].

These studies demonstrate that different deep learning strategies are applicable for disease detection in tea leaves, and that segmentation, lightweight architectures, and hybrid approaches stand out in terms of performance and efficiency.

4. MATERIALS AND METHODS

4.1. Dataset and image pre-processing

In this study, an online accessible image dataset was used to classify eight different disease categories seen in tea leaves [8]. The dataset consists of leaf images collected under natural light conditions and in a field environment, with each class labeled according to symptom type. The classes included in the dataset are determined as follows: Algal leaf, Anthracnose, Bird eye spot, Brown blight, Gray light, Healthy, Red leaf spot and White spot. These classes represent common pathological symptoms observed on the surface of tea leaves. Table 1 shows the number of images of tea leaf classes included in the dataset, and Figure 1 shows sample images of tea leaf classes included in the dataset.

Table 1. Number of images per tea leaf class included in the dataset

Sınıf Adı	Görüntü Sayısı (#)
Algal leaf	113
Anthracnose	100
Bird eye spot	100
Brown blight	113
Gray light	100
Healthy	74
Red leaf spot	143
White spot	142
Toplam	885

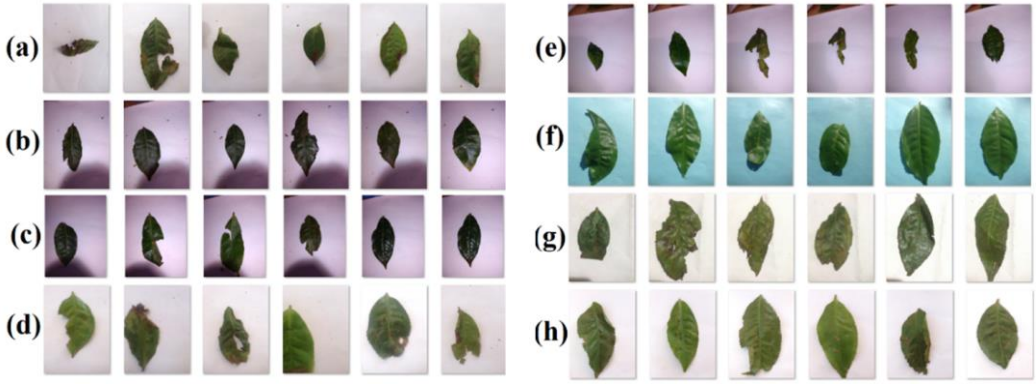


Figure 1. Sample images of tea leaf classes included in the dataset. (a) Algal leaf (b) Anthracnose (c) Bird eye spot (d) Brown blight (e) Gray light (f) Healthy (g) Red leaf spot (h) White spot

The dataset was divided into three sub-categories—training, validation, and testing to enable objective model evaluation. The data splitting ratio was applied as 70% training, 15% validation, and 15% testing. Data splitting was performed using a random selection method, ensuring a balanced distribution of each class across all subsets. Although some studies in the literature have evaluated the dataset using a two-part method with 80% training and 20% testing, this study preferred a three-part split strategy to observe the model generalization performance more stably. This approach allows for the optimization of hyperparameter settings through the validation set and ensures the test set represents the final performance. All images have been standardized to ensure dimensionally consistent model inputs. Each image in the dataset was rescaled to 224×224 pixels according to the format required by the trained models. Image data were normalized for model inputs, and pixel values were converted to a 0–1 range.

4.2. Deep Learning Models

ResNet50 is an architecture designed to mitigate the gradient fading problem seen in deep networks by using residual connections. The model enables the efficient training of deeper networks by creating direct links between successive convolutional layers [9].

DenseNet121 enhances feature sharing and provides parameter efficiency by using dense connectivity between layers. In this structure, each layer receives feature maps from all preceding layers as input, thus strengthening the flow of information [10].

The EfficientNet-B0 model is based on a combined scaling approach that performs scaling of depth, width, and resolution dimensions through a common coefficient, aiming to achieve high accuracy with fewer parameters [11].

MobileNetV3-Large is a lightweight CNN model optimized for mobile and embedded systems. This architecture, which includes depth-separable convolutional structures and Squeeze-and-Excitation modules, provides a balance between low computational cost and classification performance [12].

The ConvNeXt-Tiny model is an architecture developed by restructuring convolutional networks according to modern design principles. This structure combines traditional CNN architecture with current optimization techniques, using block design and scaling strategies inspired by Transformer-based models [13].

The Vision Transformer (ViT-Small) model is an architecture that uses a multi-headed attention mechanism instead of convolution in visual classification tasks. In this approach, images are broken down into

fixed-size segments, and each segment is processed as a series of elements by the Transformer encoder. The model performs representation learning that takes global context into account and can simultaneously evaluate the relationships between all regions in the image. This structure, unlike CNN-based methods, is fundamentally based on attention-based global feature interactions instead of local feature extraction [14].

4.3. Training Strategy

In this study, deep learning models for the classification of tea leaf diseases were trained using a transfer learning approach based on the reuse and adaptation of pre-trained network weights to the target dataset. All models were initialized with initial parameters trained on the ImageNet dataset, and the classification layer was reconfigured to represent the eight disease classes in the dataset. The training process aims to preserve the overall representational capacity of the model while learning about the distinctive features specific to each class.

4.3.1. Transfer Learning

Transfer learning is a method that allows models trained on large-scale datasets to be adapted to new problems with limited data. According to Pan and Yang's comprehensive definition, previously learned information is transferred from a source task to a target task, thus improving learning performance in the new task [15]. In this study, CNN-based architectures (ResNet50, DenseNet121, EfficientNet-B0, MobileNetV3-Large, ConvNeXt-Tiny) and the Transformer-based ViT-Small model were initialized with ImageNet pre-trained weights. While the pre-trained layers provided a representation of general image features, the final classification layer was retrained in line with the target labels of the study. This method improves the parameter efficiency of the model, reduces training time, and enhances performance in scenarios with limited data.

4.3.2. Fine-Tuning

A fine-tuning method was applied to adapt the pre-trained models to the new dataset. In the first stage, only the classification layer was randomly initialized, while the other layers were kept constant. This strategy ensures the preservation of pre-trained filters, enabling rapid adaptation to the target mission from the outset. In the second stage, the final blocks of the model were gradually re-trained to ensure that the feature representations were aligned with disease classes. The gradual opening of the layers aims to preserve the pre-trained filters and reduce the risk of overfitting [16]. Thus, the model learned task-specific features without losing the general visual information acquired in large-scale data.

4.3.3. Hyperparameters

All models were trained with the same hyperparameter settings to provide comparable performance evaluations. During training, the batch size was set to 16, and input images were processed at a resolution of 224×224 pixels. The learning rate was initially set to 1×10^{-4} and remained constant in all experiments. The data loader is divided to include 70% of the training set, 15% of the validation set, and 15% of the test set.

4.3.4. Loss Function, Optimizer ve Epoch Sayısı

In the training process, multi-class cross-entropy loss was used as the loss function. This function measures the difference between the predicted class distribution and the actual class labels, and ensures the updating of model weights. The Adam optimization algorithm was

applied for the optimization process, and all parameters were updated using this method. Adam accelerates the gradient-based learning process by offering a momentum-based adaptive learning rate strategy. The training cycle was run for 20 epochs for all models, and loss and accuracy values were calculated on the validation data after each epoch. The weights recorded in the epoch with the highest validation accuracy were used for performance evaluation in the testing phase.

5. RESULTS

This section presents the performance results of six deep learning models used to classify eight disease classes found in tea leaves. All models were evaluated under the same training–validation–test split (70–15–15). Accuracy, precision, recall, F1-score values, and inference time per image were reported as performance metrics.

5.1. Accuracy and F1-score comparison

The validation accuracy and F1 scores obtained for each model are summarized in Table 2. All models were evaluated using the same training/validation/testing section (70-15-15). Validation accuracy, the F1-Score representing the average across classes, and inference time results were obtained as performance metrics. The CNN-based ConvNeXt-Tiny model demonstrates the highest performance on the validation data with the highest accuracy (0.94) and F1-Score (0.94). DenseNet121 and ResNet50 showed high performance with similar levels of accuracy (0.93) and F1-Score values. The Vision Transformer (ViT-Small) model performed worse in terms of validation accuracy compared to CNN-based models. When the average prediction time was evaluated, the MobileNetV3-Large model achieved the fastest inference time.

Table 2. Numerical performance results obtained from each model construct using the validation dataset

Model	Accuracy	Precision	Recall	F1-Score	Inference Time (ms/image)
ResNet50	0.93	0.93	0.93	0.93	86.16
DenseNet121	0.93	0.93	0.93	0.93	89.41
EfficientNet-B0	0.90	0.91	0.90	0.90	64.34
MobileNetV3-Large	0.93	0.92	0.92	0.92	55.97
ConvNeXt-Tiny	0.94	0.94	0.94	0.94	78.36
ViT-Small	0.90	0.90	0.90	0.90	157.87

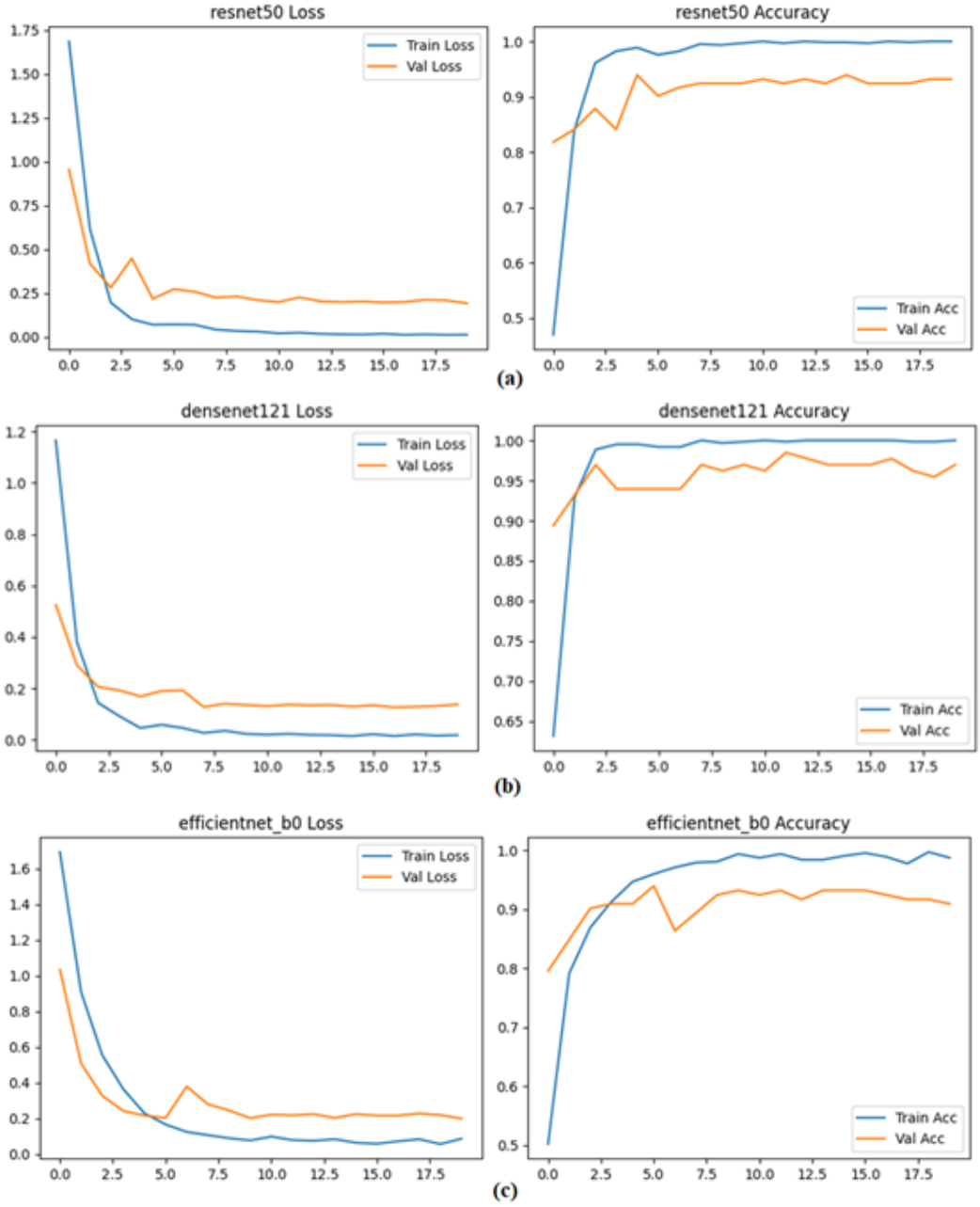
5.2. Training–Validation Curves

The loss and accuracy curves obtained throughout the training process demonstrate the learning dynamics of the models (Figure 2). All models were trained using a transfer learning approach, and early saturation in validation performance was observed throughout the training. The low difference between the training and validation curves of the ConvNeXt-Tiny model indicates a more stable learning process. In particular, DenseNet121 and ConvNeXt-Tiny produced more stable validation curves and did not exhibit overfitting.

When the training loss curves of the models are examined, it is seen that DenseNet121 and ConvNeXt-Tiny show rapid convergence behavior in the first epochs and limited fluctuation in loss values in subsequent epochs. ResNet50 and MobileNetV3-Large models, on the other hand, exhibited a more fluctuating appearance with increases in validation loss in certain epochs. In the EfficientNet-B0 and ViT-Small models, while the validation loss curves showed a plateau effect after certain epochs, the trend of gradually decreasing training loss continued. When comparing the accuracy curves, it is seen that the ConvNeXt-Tiny and DenseNet121 models brought the validation accuracy to a certain level

in the early stages and maintained this value in subsequent epochs. In the ResNet50 and MobileNetV3-Large graphs, it was observed that the validation accuracy fluctuated in a wider range; and in the EfficientNet-B0 and ViT-Small models, a periodic divergence occurred between training and validation accuracy. This situation reveals the level of adaptation that different architectures show to the validation set in the learning process.

There is a clear tendency towards convergence in the training and validation accuracy curves for all models. In the DenseNet121 and ConvNeXt-Tiny graphs, it is seen that the validation accuracy remained at high values from the early epochs; The EfficientNet-B0 and ViT-Small models appear to have lower validation accuracy. Additionally, the fact that the validation curve of ConvNeXt-Tiny shows a similar trend to the training curve indicates that the overall performance is stable during the training process.



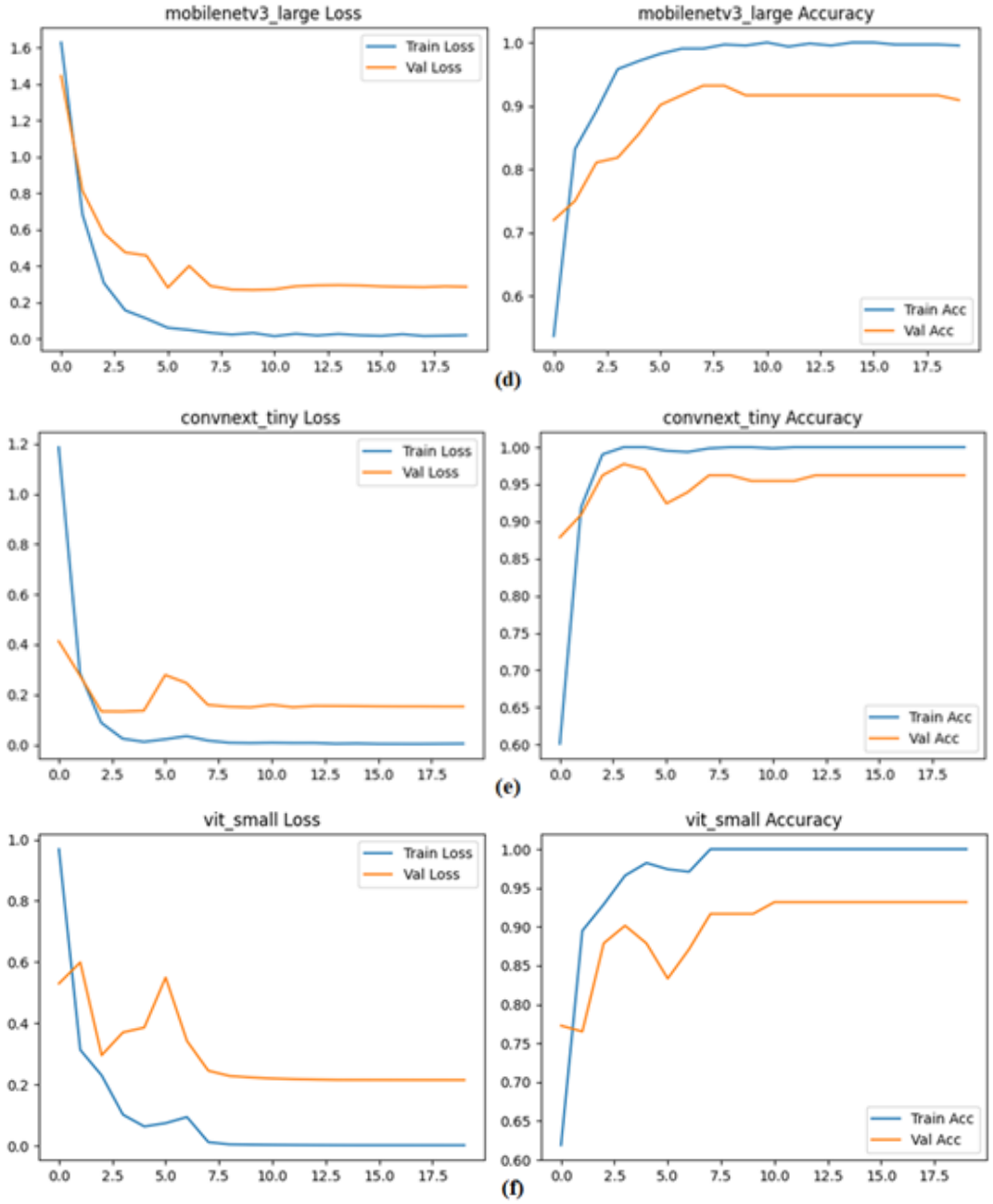
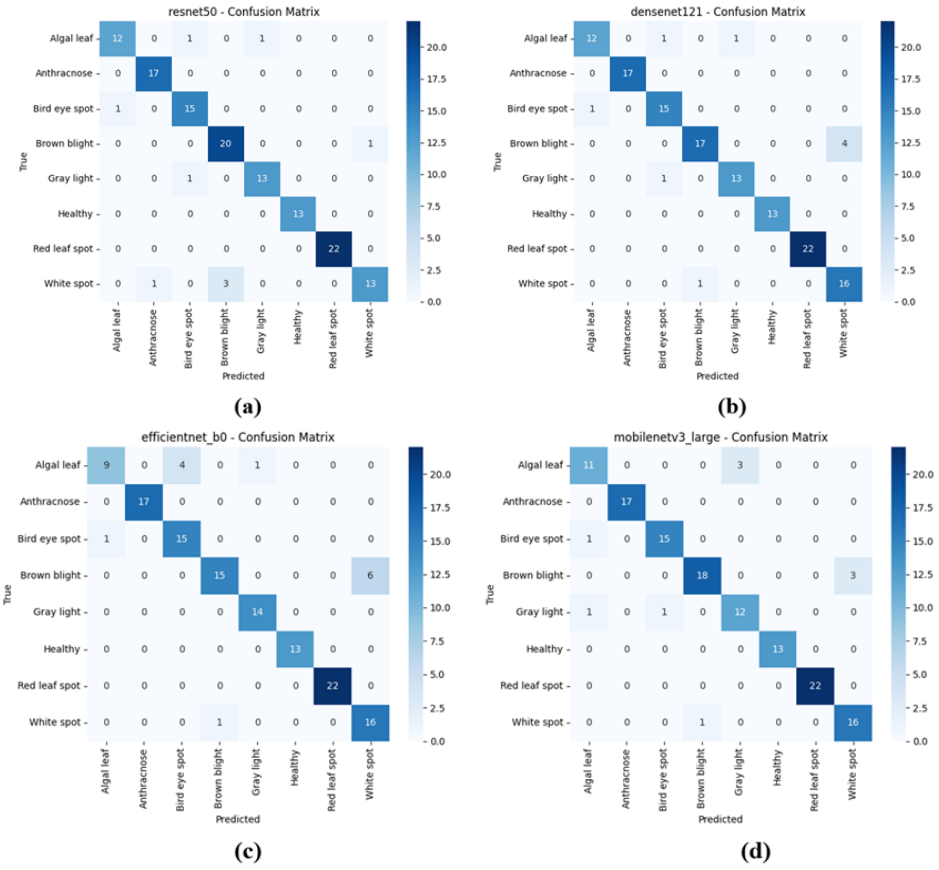


Figure 2. Train-validation loss graphs and success-accuracy graphs for each deep learning method. (a) Resnet50 (b) DenseNet121 (c) EfficientNet-B0 (d) MobileNetV3 (e) ConvNeXt-Tiny (f) Vit-Small

5.3. Confusion Matrix

The class-based performance of the models was evaluated using confusion matrices (Figure 3). For each model, the correct and incorrect classification distributions of eight classes were analyzed and visualized in the matrix. The ConvNeXt-Tiny and DenseNet121 models offer more balanced performance across their classes. Both models show that the "Healthy", "Red leaf spot", and "Anthracnose" classes have high accuracy rates. Relatively lower success rates were observed in some classes in the EfficientNet-B0 and ViT-Small models. When the confusion matrices are examined, it is seen that in the ConvNeXt-Tiny and DenseNet121 models, errors are not concentrated in specific categories among classes, and incorrect classifications generally remain at a low

level. In both models, the "Healthy" class stands out with its high number of accurate classifications. In addition, the "Red leaf spot" and "Anthracnose" classes were observed to have the highest correct classification rates among all models. The EfficientNet-B0 and ViT-Small models contain examples of cross-classification within specific classes. In the EfficientNet-B0 matrix, the misclassification rate is higher in the "Algal leaf" and "Brown blight" classes compared to other classes. In the ViT-Small matrix, it is observed that errors are concentrated in the "White spot" and "Brown blight" categories. Conversely, high accuracy levels are maintained in the "Healthy" class in both models. Comparing the matrices of the ResNet50 and MobileNetV3-Large models, it is observed that the overall accuracy levels are similar, and misclassifications are concentrated in limited categories. In the ResNet50 matrix, the "White spot" class is sometimes confused with the "Gray light" class. In the MobileNetV3-Large model, a relatively balanced distribution is observed among the classes, but errors are more frequent in the "Gray light" category compared to other classes.Overall, the confusion matrices show that all models exhibit different levels of success across the eight classes. ConvNeXt-Tiny and DenseNet121 presented a more homogeneous result in the distribution of success across classes. The EfficientNet-B0 and ViT-Small models contain errors concentrated in certain classes. This indicates that classification performance can vary depending on the number of class samples, symptom similarity, and the architectural features of the model.



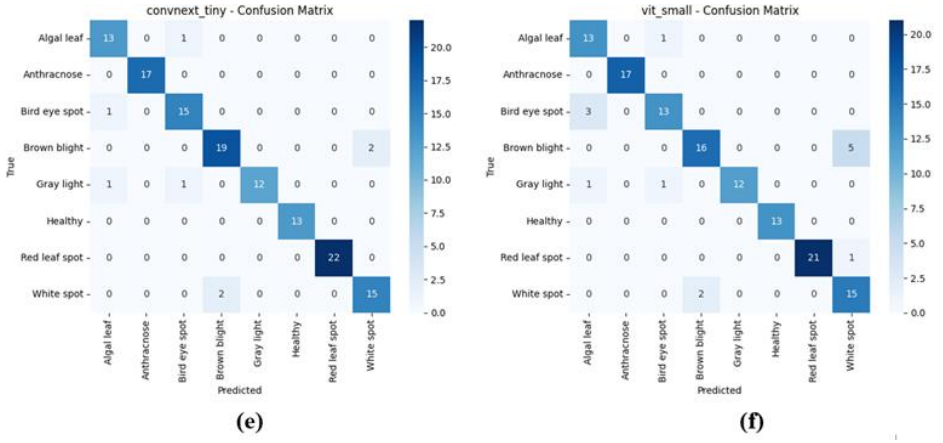


Figure 3. Confusion matrix graph for each deep learning method. (a) Resnet50 (b) DenseNet121 (c) EfficientNet-B0 (d) MobileNetV3 (e) ConvNeXt-Tiny (f) Vit-Small

5.4. Explainability (Grad-CAM)

The Grad-CAM method was used to visualize model decisions and evaluate the explainability of classification decisions [29]. With this approach, leaf regions that contribute most to classification decisions were shown via heat maps. Sample images representing each class were generated for the ConvNeXt-Tiny model; Correlation was observed between symptom regions and network activations. Grad-CAM visualizations increase the explainability of the results by revealing which morphological structures the model uses in the prediction process.

The Grad-CAM method was applied to visualize the decision processes of the ConvNeXt-Tiny model. The obtained heat maps show that the model focuses particularly on symptom regions on the leaf in its classification decisions. Active focus areas are mostly concentrated on necrotic tissues, regions experiencing pigment loss, and irregular tissue structures. In healthy leaf samples, it was observed that the model's activations were distributed at a lower level along the leaf vein structure. These results show that deep feature maps can capture disease symptoms at the semantic level and that the model performs class differentiation based on visual anomalies in symptom regions. Grad-CAM outputs demonstrate that the model's predictive behavior is consistent with symptomatic domains and enhances the explainability of the decision-making process. Figure 4 shows Grad-CAM examples of the ConvNeXt-Tiny model, the best deep learning model obtained in this study.

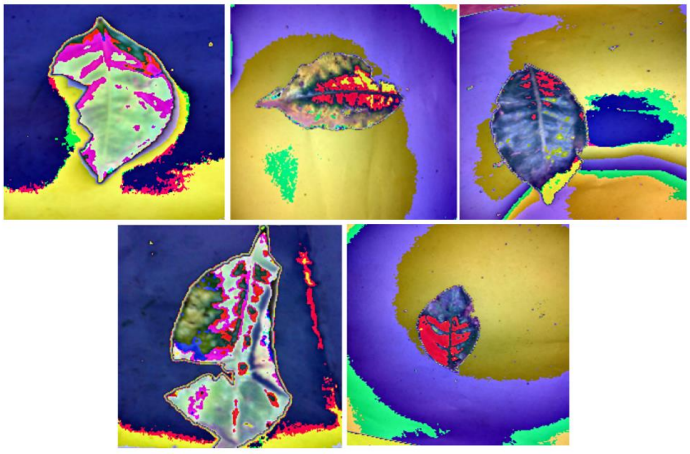


Figure 4. Grad-CAM examples (ConvNeXt-Tiny)

6. DISCUSSION

Experimental findings show that the DenseNet121 architecture achieved higher accuracy and F1 scores than other models during the validation phase. This result may be due to DenseNet's dense interconnect structure. Direct information flow between layers reduces gradient loss and increases feature reuse, resulting in more stable learning under limited data conditions. The fact that the visual symptoms of leaf diseases are quite similar to each other has made it advantageous to transfer fine details from different layers. The superior performance achieved during the validation phase reveals that DenseNet demonstrates high fit across the samples in its training distribution, but this success is not reflected to the same extent on the test set. This supports the findings that validation performance does not directly reflect generalization ability in agricultural imaging problems studied with limited data.

According to the test results, ConvNeXt-Tiny achieved higher macro-F1 and accuracy values than all other models. Although DenseNet appeared to lead during the training and validation phases, ConvNeXt's better performance on an independent test set indicates that this model has a stronger generalization capacity. ConvNeXt is an architecture that follows modern CNN design principles; Normalization strategies, thanks to wider convolution windows and Transformer-like block arrangement, have been able to more effectively distinguish complex textural differences. The results show that ConvNeXt can deliver high performance on new samples without overfitting to the training data. When aiming for real-time detection of tea leaf diseases in field applications, test performance becomes a more decisive criterion compared to validation performance.

The dataset was divided using the commonly used 70% training, 15% validation, and 15% testing ratios in the literature. This distribution aims to provide sufficient samples for the training process while allowing for the use of an independent validation section to reduce the risk of overfitting. However, the limited sample size of the tea leaf disease data led to performance differences between validation and test scores. The inability of models that achieved high performance in the validation set to maintain the same level in the test set suggests that the samples in the validation section may be more similar to the training distribution. This indicates that performance changes can be observed when data diversity is increased or when different splitting strategies (e.g., stratified split) are applied. The results reveal that data splitting strategies in agricultural image analysis have a direct impact on model selection.

The models used in this study represent two main architectural approaches: CNN-based models (ResNet, DenseNet, EfficientNet, MobileNet, ConvNeXt) and a Transformer-based model (ViT-Small). Test results have shown that CNN-based models offer higher classification performance than the Transformer model, especially under limited data conditions. While ViT enables powerful representation learning on large-scale datasets, it typically requires additional pre-training or data augmentation mechanisms for high performance on smaller datasets. In contrast, CNNs, thanks to their principle of hierarchically capturing local spatial features, have been able to more effectively represent agricultural symptoms such as point spots, color changes, and textural anomalies on leaf surfaces. The findings reveal that the Transformer-based approach requires further improvement compared to CNNs in this problem.

Model comparison results show that architectures with a high number of parameters do not always offer the highest accuracy. Models that are more compact in terms of parameters (e.g., MobileNetV3-Large) achieved high success in some classes but lagged behind ConvNeXt-Tiny and DenseNet

in the overall average. On the other hand, Transformer-based models such as ViT-Small, despite requiring higher computational costs, failed to deliver the expected performance on small datasets. This situation demonstrates that in problems where features are limited but semantically dense, such as leaf diseases, the feature extraction strategy, rather than model complexity, is the determining factor. In conclusion, it is understood that performance evaluation requires a combined interpretation of factors such as accuracy, computational cost, estimation time, and model size.

7. CONCLUSION

This study presents a comparative evaluation of six deep learning models aimed at classifying eight different disease categories in tea leaves. Experiments conducted on the same dataset revealed the performance of the models in terms of accuracy, F1-score, macro-average F1, and computational efficiency. The results show that DenseNet121 demonstrated the highest performance in the validation phase; However, when evaluated on an independent test set, the ConvNeXt-Tiny model demonstrates stronger generalization capabilities than all other architectures. This finding suggests that modern CNN-based architectures can offer more stable performance compared to traditional approaches in agricultural imaging problems with limited data.

The results support the real-world applicability of deep learning-based systems for classifying visual signs of tea leaf diseases. High success rates, partial elimination of class imbalances, and fast prediction times demonstrate that these methodologies can be integrated with mobile applications, artificial vision-assisted drone systems, and smart agricultural robots in the field of agriculture. In particular, the ConvNeXt-Tiny model offers high accuracy despite a lower number of parameters, providing potential for real-time diagnostics even in devices with limited hardware capabilities. Such a system can reduce crop losses by enabling early disease detection, optimize pesticide use, and contribute to agricultural productivity.

Future studies will focus on increasing data diversity and model optimization. Expanding the dataset with samples from different geographic regions, different climatic conditions, and various imaging devices can positively impact generalization performance. Furthermore, re-evaluating Vision Transformer architectures with larger-scale data pre-training, self-supervised learning, few-shot learning, or domain adaptation approaches can improve the success of Transformer-based models in small data scenarios. In addition, model compression, quantization, and hardware acceleration methods are suggested as future research areas to reduce computational costs for real-time agricultural applications.

Overall, this study provides a comprehensive analysis of different deep learning architectures in the automated classification of tea leaf diseases, creating a comparative reference framework for future studies in the field of agricultural artificial vision.

CONFLICT OF INTEREST

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

FINANCIAL DISCLOSURE

The authors did not receive any financial support in conducting this study.

DECLARATION OF ETHICAL STANDARDS

The authors of this article declare that the materials and methods used in this study do not require ethical committee permission and/or legal-special permission.

REFERENCES

[1] Yang, J., Xu, G., Yang, M., and Lin, Z., (2025). Lightweight wavelet-CNN tea leaf disease detection. PLoS One, 20(5):e0323322.

[2] Ahmed, F., Ahad, M.T., and Emon, Y.R., (2023). Machine learning-based tea leaf disease detection: A comprehensive review. arXiv preprint arXiv:2311.03240.

[3] Ozturk, O., Sarica, B., and Seker, D.Z., (2025). Interpretable and robust ensemble deep learning framework for tea leaf disease classification. Horticulturae, 11(4):437.

[4] Ye, R., Gao, Q., and Li, T., (2024). BRA-YOLOv7: improvements on large leaf disease object detection using FasterNet and dual-level routing attention in YOLOv7. Frontiers in Plant Science, 15, 1373104.

[5] Zhang, J., Guo, H., Guo, J., and Zhang, J., (2023). An information entropy masked vision transformer (iem-vit) model for recognition of tea diseases. Agronomy, 13(4):1156.

[6] Xue, Z., Xu, R., Bai, D., and Lin, H., (2023). YOLO-tea: A tea disease detection model improved by YOLOv5. Forests, 14(2):415.

[7] Li, J. and Liao, C., (2025). Tea disease recognition based on ımage segmentation and data augmentation. IEEE Access.

[8] Kaggle, "Identifying Disease in Tea leaves", kaggle.com, [Online]. Available: https://www.kaggle.com/datasets/shashwatwork/identifying-disease-in-tea-leafs. [Accessed: Nov. 11, 2025].

[9] He, K., Zhang, X., Ren, S., and Sun, J., (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp:770-778).

[10] Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K.Q., (2017). Densely connected convolutional networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp:4700-4708).

[11] Tan, M. and Le, Q., (2019, May). Efficientnet: Rethinking model scaling for convolutional neural networks. In International conference on machine learning (pp:6105-6114). PMLR.

[12] Howard, A., Sandler, M., Chu, G., Chen, L.C., Chen, B., Tan, M., ... and Adam, H., (2019). Searching for mobilenetv3. In Proceedings of the IEEE/CVF international conference on computer vision (pp:1314-1324).

[13] Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., and Xie, S., (2022). A convnet for the 2020s. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp:11976-11986).

[14] Dosovitskiy, A., (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

[15] Pan, S.J. and Yang, Q., (2009). A survey on transfer learning. IEEE Transactions on knowledge and data engineering, 22(10):1345-1359.

[16] Yosinski, J., Clune, J., Bengio, Y., and Lipson, H., (2014). How transferable are features in deep neural networks? Advances in neural information processing systems, 27.