PLANT SCIENCE TODAY
Vol 13(1): 1-9
https://doi.org/10.14719/pst.11513

eISSN 2348-1900

RESEARCH ARTICLE

Multi class tea leaf disease classification using feature level and output level ensemble strategies with Grad-CAM visualization

Kandarpa Kalita[1], Kishore Medhi[2], Sanjeet Kumar Borah[3], Aniruddha Deka[1] & Sunandan Baruah[1*]

[1]Faculty of Computer Technology, Assam down town University, Guwahati 781 026, Assam, India
[2]Department of Computer Applications, Assam Don Bosco University, Guwahati 781 017, Assam, India
[3]Faculty of Agricultural Sciences and Technology, Assam down town University, Guwahati 781 026, Assam, India

*Correspondence email - sunandan.baruah@adtu.in

Received: 28 August 2025; Accepted: 07 December 2025; Available online: Version 1.0: 07 January 2025; Version 2.0: 19 January 2026

Cite this article: Kandarpa K, Kishore M, Sanjeet KB, Aniruddha D, Sunandan B . Multi class tea leaf disease classification using feature level and output level ensemble strategies with Grad-CAM visualization. Plant Science Today. 2026; 13(1): 1-9. https://doi.org/10.14719/pst.11513

Abstract

Tea is one of the most widely consumed drinks in the world and plays an important role in the economy of tea-growing regions. However, leaf diseases are a major problem for the tea industry because they lower both yield and quality, directly affecting tea growers' livelihoods and the overall supply chain. Therefore, detecting these diseases in their early stages is important for healthy crop growth and good yield. In traditional practice, disease identification is carried out through field inspection or laboratory testing by expert farmers and plant pathologists. However, these methods are slow, require a lot of manual work and may sometimes lead to errors, making them difficult to adopt for large-scale cultivation. This study presents a novel deep learning-based methodology for eight-class tea leaf disease classification, exploring feature-level and output-level ensemble strategies. Four widely used convolutional neural networks (CNNs)-ResNet-18, VGG-16, InceptionV3 and MobileNetV2-pretrained on ImageNet were utilized. In the feature-based approach, deep features extracted from these models were compressed using principal component analysis (PCA) and classified using a Random Forest classifier, achieving an accuracy of 95.6 %. In the output-based approach, probability predictions from the above CNNs were combined, resulting in a higher accuracy of 98.3 %. Grad-CAM visualizations confirmed that the models consistently highlighted symptomatic leaf regions, improving interpretability and user trust.

Keywords: convolutional neural networks; deep learning; ensemble learning; Grad-CAM; precision agriculture; tea leaf disease detection

Introduction

Tea (*Camellia sinensis* L.), an evergreen plant native to Asia, thrives in warm and humid climates. India is the world's second-largest tea producer, with Assam contributing ap proximately 630–700 million kilograms annually as of 2025, according to the Tea Board of India (1). Assam alone accounts for more than half of the total tea-cultivated area in the country. However, the same climatic conditions that support tea growth also encourage the spread of diseases. The continuous expansion of tea plantations in recent years has led to increasing incidences of leaf diseases such as blister blight, red rust, grey blight and other foliar infections, all of which negatively affect both yield and quality. Due to the vast geographical spread of plantations, timely detection and management of diseases remain challenging, often resulting in significant economic losses (2). Consequently, early detection of tea leaf diseases is essential to ensure sustainable production and minimize crop loss.

Early disease diagnosis plays a key role in maintaining leaf quality, reducing yield decline and preventing excessive or indiscriminate pesticide use (3). Recent advancements in image-based disease classification have demonstrated that convolutional neural networks (CNNs) can efficiently learn discriminative spatial features from leaf images, outperforming traditional hand-crafted descriptors across a wide variety of horticultural crops (4). CNNs have also shown strong generalization performance under field conditions involving variations in illumination, background noise and environmental complexity (5). An earlier study proposed a novel deep convolutional neural network (DCNN) using inverted residuals and linear bottleneck layers for the automatic detection of grey blight disease on tea leaves and achieved a test accuracy of 98.99 % (6). Researchers introduced a model called YOLO-tea, which is an enhanced version of YOLOv5 designed specifically for detecting multiple tea diseases by improving feature extraction and small-object recognition capability (7). Previous studies evaluated several deep learning models for real-time tea leaf disease detection, focusing on achieving fast and reliable performance suitable for field applications (8). An earlier study proposed a deep neural network that uses a hybrid pooling strategy to improve feature extraction for automatic identification of tea leaf diseases (9). Researchers extracted deep learning features and use a kernelized SVM to improve the accuracy of tea leaf disease prediction (10). Their hybrid approach showed better performance than standalone CNN or SVM models. Another study introduced an improved YOLOv7-

MobileNeXt model designed for efficient and accurate classification of multiple tea diseases, particularly optimized for lightweight deployment (11). Researchers also proposed a hybrid approach that combines CNN-based feature extraction with a Random Forest classifier for multiclass tea leaf disease identification (12).

Despite these advantages, standalone CNN models still face challenges such as over fitting, inter-class confusion and domain shifts between laboratory and real-world environments.

Ensemble learning has emerged as a promising solution to these limitations, offering improved robustness by combining the complementary strengths of multiple models (13). Two ensemble paradigms have shown particular effectiveness: feature-level fusion, which integrates deep feature representations from different CNNs and output-level fusion, which aggregates model predictions into a unified decision (14). Studies in plant disease classification show that ensembles often surpass single-model approaches, especially when datasets exhibit overlapping symptoms, subtle inter-class variations and limited training samples (15).

Previous research presented a multi-model ensemble methodology named PlantDet, which combines several deep learning architectures to improve the reliability of plant disease detection across diverse datasets (16). Researchers proposed a hybrid deep learning approach for image-based plant disease detection that combines CNN feature extraction with additional learning layers to enhance classification performance (17).

Other researchers presented a hybrid deep multistacking integrated model designed to improve the accuracy and stability of plant disease detection by combining multiple deep learning architectures (18). Another study proposed a hybrid model for leaf disease classification that combines modified deep transfer learning with an ensemble approach to enhance performance in agricultural AIoT-based monitoring systems (19). An earlier study presented an ensemble hybrid framework that combines CNN-based features with various metaheuristic optimization algorithms for plant disease classification (20). Researchers proposed a hybrid ensemble model that combines CNN and RNN features for multimodal cotton plant disease detection, applying both spatial and sequential information (21).

In addition, explainable AI (XAI) techniques such as Gradient-weighted class activation mapping (Grad-CAM) provide visual insights into model decision-making by highlighting the discriminative regions contributing to classification (22). Such interpretability frameworks support model transparency and enable agronomists to assess whether computational predictions align with biologically meaningful symptom patterns. While deep learning-based disease detection has been widely studied for various crops, research focusing on multiclass tea leaf disease classification using both feature-level and output -level ensembles remains limited (23). Furthermore, many existing approaches rely on single CNN models, lack interpretability, or are restricted by dataset size and diversity, limiting their generalizability (24). There is also a shortage of integrated systems that combine ensemble learning with visual explanation tools for real-world decision support in tea plantations (25).

To address these gaps, this paper proposes a hybrid ensemble framework that com bines feature-level CNN fusion with output-level probabilistic ensemble decision making. Grad-CAM visualization is incorporated to highlight disease-specific lesion regions and enhance interpretability. This comprehensive approach aims to improve classification ac curacy, transparency and robustness, ultimately supporting scalable precision agriculture practices for tea growers. By integrating deep feature fusion, ensemble strategies and explainable AI, this study contributes to advancing AI-enabled plant health monitoring and promoting sustainable crop management through data-driven technologies.

Materials and Methods

Proposed methodology

This study presents two ensemble learning strategies aimed at enhancing the performance, stability and generalization capability of tea leaf disease classification systems. Four pre-trained CNN architectures, namely ResNet-18, VGG-16, MobileNetV2 and InceptionV3, are employed as base learners due to their proven effectiveness in extracting diverse and discriminative feature representations from image data. The first strategy, referred to as the output-level ensemble, integrates the SoftMax probability scores generated by each CNN through element-wise averaging, thereby reducing prediction variance and overcoming the limitations of individual models. The second strategy, termed the Feature-Level RF ensemble, aggregates deep features extracted from the four CNNs, applies PCA to reduce feature dimensionality and subsequently classifies the compressed representations using a Random Forest classifier.

Dataset description

The dataset employed in this study is publicly available on Mendeley data and comprises a total of 885 high-quality RGB images of *C. sinensis* leaves (26). The dataset includes eight distinct classes, representing seven common tea leaf diseases along with healthy leaf specimens, with a near-balanced distribution across categories. The distribution of images in the dataset is shown in Table 1. This balanced representation is critical for preventing bias during model training and ensuring reliable classification performance.

Table 1. Number of images per class in the dataset

Class name	Image count	Condition type
Algal leaf	113	disease
Gray blight	100	disease
Anthracnose	100	disease
Healthy	74	normal
Bird eye spot	100	disease
Red leaf spot	143	disease
Brown blight	113	disease
White spot	142	disease
Total	885	—

Representative samples from a few classes are provided in Fig. 1, demonstrating the characteristic visual symptoms used for classification. The healthy class specimens show optimal leaf morphology without discoloration or textural abnormalities.

Augmentation

To help the models cope with natural variation in field images, we applied a simple but effective set of image augmentations (27, 28). Images were randomly rotated, shifted, sheared, zoomed and flipped to simulate changes in leaf pose and camera angle; brightness was varied to reflect different lighting; Gaussian noise was added to mimic sensor disturbances; and random erasing (50×50 px cutout) was used to increase resilience to partial occlusion. All images were resized to 224 × 224 pixels and normalized prior to training. Most transforms were implemented using Keras's Image Data Generator, with Gaussian noise and cutout applied via small custom preprocessing routines, producing a richer and more varied training set.

Preprocessing

Preprocessing is essential for converting raw images into a consistent format suitable for deep learning. In this study, three preprocessing steps were applied: image resizing, pixel normalization and dataset partitioning. All images were resized to 224 × 224 pixels to match the input requirements of the CNN models while maintaining sufficient visual detail. Pixel intensities were then normalized from the range [0, 255] to [0, 1] to stabilize training and reduce sensitivity to illumination variations. The dataset was divided into training (70 %), validation (15 %) and testing (15 %) subsets. The training set was used to learn model parameters, the validation set to tune hyperparameters and the test set to provide an unbiased evaluation of the final performance. The preprocessing steps are summarized in Fig. 2.

Implementation

Following preprocessing, each of the four pretrained CNN architectures mentioned above was fine-tuned on the tea leaf dataset using a transfer learning approach. Transfer learning enables the convolutional feature extractor pretrained on ImageNet to be reused while adapting the final layers to the target classification task, thereby reducing training time and minimizing overfitting in limited data scenarios (29).

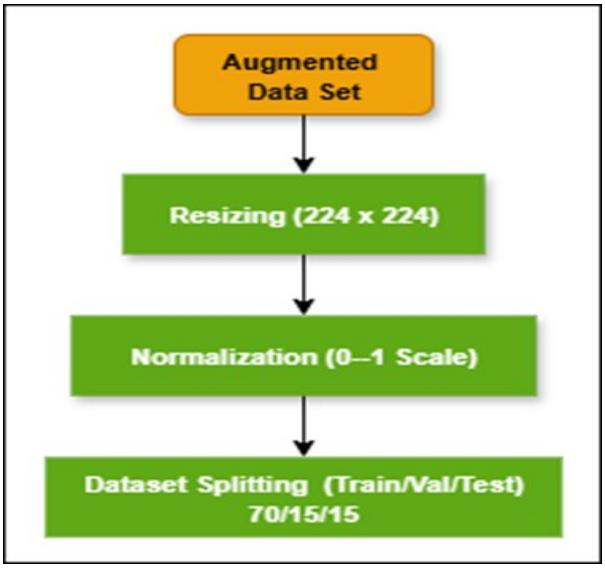

Fig. 2. Preprocessing flow chart.

For each model, the pretrained feature extraction layers were initialized with ImageNet weights and the original fully connected layer was replaced with a new classifier corresponding to the number of tea leaf disease classes (i.e., 8). During the initial training phase, the early feature extraction layers were frozen to retain generic visual representations, while deeper layers were gradually unfrozen for fine-tuning to capture task-specific patterns.

Model training employed the Adam optimizer with a learning rate of 1×10^{-4} and categorical cross-entropy loss. Adam was selected for its adaptive learning rate mechanism, which supports stable and efficient convergence. A batch size of 32 was used and early stopping was applied, halting training if validation accuracy did not improve for 50 consecutive epochs. This strategy prevented overfitting and eliminated unnecessary computation.

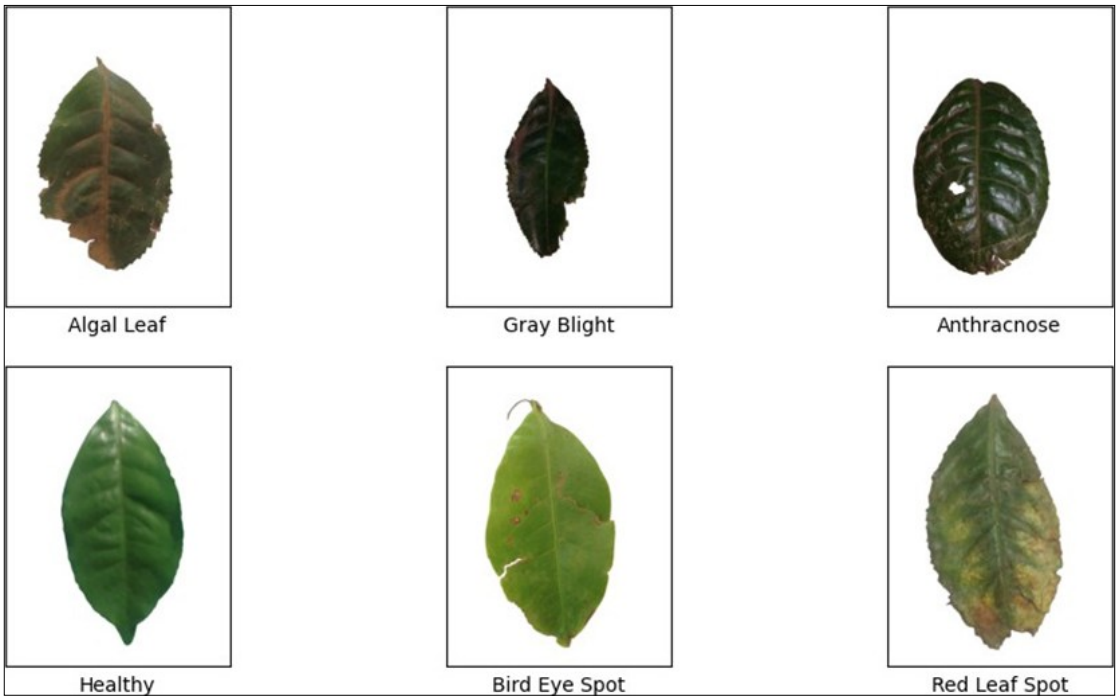

Fig. 1. Sample tea leaf disease images collected from the source mentioned above.

Output-level ensemble with SoftMax classification

In the output-level ensemble strategy, each CNN model independently processes the input image and produces a probability distribution over all disease classes through the SoftMax activation in its final classification layer. For a model output logit vector Z = (z_1, z_2, . . . , z_K), the SoftMax probability assigned to class *c* is given by

$$P(c) = \frac{\exp(z_c)}{\sum_{k=1}^{K} \exp(z_k)} \qquad (1)$$

where, *K* denotes the total number of disease classes. Thus, each CNN generates a probability vector.

$$P_i = [\, P_i(1), P_i(2), \ldots, P_i(K) \,], \qquad (2)$$

where $i \in \{1, 2, 3, 4\}$ corresponds to the four base models (ResNet-18, VGG-16, Mo- bileNetV2 and InceptionV3).

To obtain a unified prediction, the probability vectors from all models are fused using element-wise averaging:

$$P_{\text{final}}(c) = \frac{1}{N} \sum_{i=1}^{N} P_i(c) \qquad (3)$$

where $N = 4$ is the number of CNN models and $P_{\text{final}}(c)$ is the aggregated probability for class *c*. The final predicted label is then selected as the class with the highest fused probability:

$$y_{pred} = arg \max_{c} \; P_{final}(c) \qquad 4$$

By averaging the probability distributions from multiple models, this ensemble approach reduces the impact of misclassification by any single network, compensates for individual model biases and improves the overall stability and robustness of the classification system. The workflow is summarized in the Fig. 3 and the Algorithm 1.

Algorithm 1: Probability averaging ensemble

1: Input: Pre-trained models $M = \{M_1, M_2, M_3, M_4\}$; test dataset

$$D = \{x_i\}_{i=1}^{N}$$

2: Output: Predicted labels $\hat{y}_i$

3: Preprocessing: Resize and normalize all input images.

4: for each sample $x_i \in D$ do

5: Obtain class probability vectors $P_k(x_i)$ from each model M_k.

6: Compute the averaged probability: $P_{avg}(x_i) = \frac{1}{4}\sum_{k=1}^{4} P_k(x_i)$

7: Assign the predicted label: $\hat{y}_i = \arg\max_c P_{avg}(x_i)$

8: end for

Feature-level RF ensemble with deep feature extraction and PCA

In the feature-level Random Forest (RF) ensemble, deep features were extracted from the final fully connected layer of each CNN model-ResNet-18 (512), VGG-16 (4096), MobileNetV2 (1280) and InceptionV3 (2048)-and these feature vectors were concatenated to form a single high-dimensional representation.

$$F = [\, f_1 \,||\, f_2 \,||\, f_3 \,||\, f_4 \,] \qquad (5)$$

Before dimensionality reduction, each feature block was standardized using the training-set mean and standard deviation to ensure uniform scaling across the different architectures.

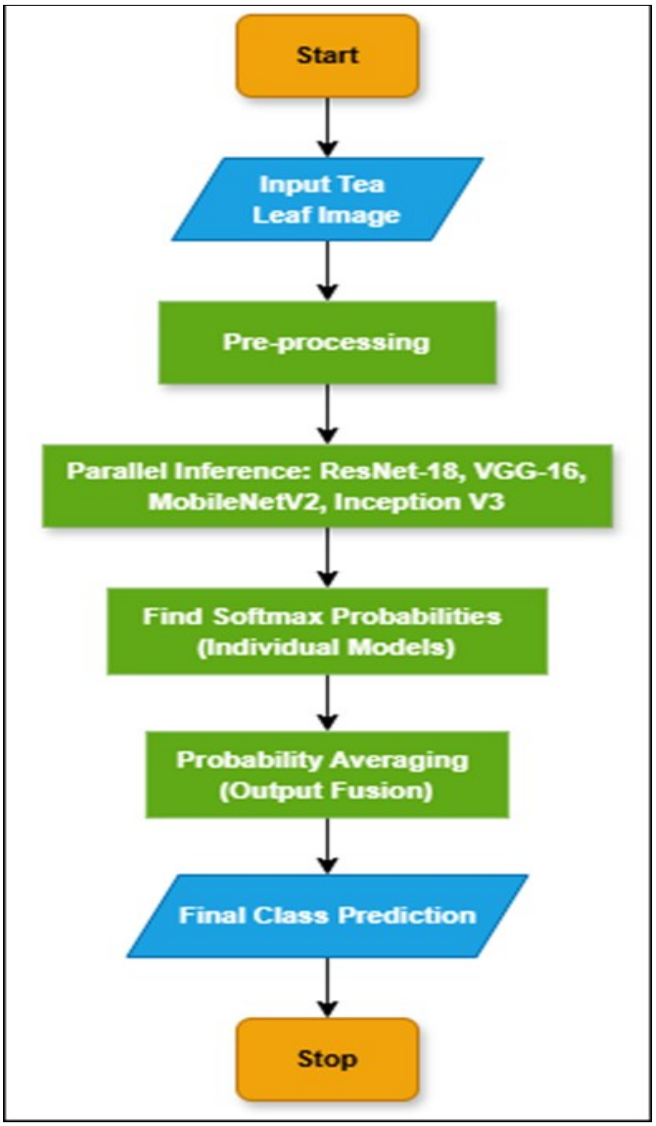

Fig. 3. Workflow of probability averaging ensemble.

PCA was then applied to remove redundant information and reduce noise. PCA was implemented using the SVD solver with $n_{\text{components}} = 300$, selected based on achieving more than 90 % cumulative explained variance. Whitening was disabled (whiten=False) to preserve the natural variance structure and a fixed random state=42 ensured reproducibility. The resulting compressed feature representation is denoted as $F_{PCA} \in R300$,

which served as the input to the RF classifier.

The Random Forest model was optimized using stratified 5-fold cross-validation on the training set. The hyperparameter search space included: n estimators ∈ {100, 200, 300, 500}, max depth ∈ {None, 10, 20, 30, 50}, min samples split ∈ {2, 5, 10}, min samples leaf ∈ {1, 2, 4}, max features ∈ {'sqrt', 'log2'} and bootstrap = True. Balanced accuracy and macro-F1 were used as the selection metrics to address potential class imbalance and out-of-bag (OOB) estimates (oob score=True) provided an internal generalization check. The best configuration was found to be n estimators = 300, max depth = 30, min samples split = 2, min samples leaf = 1, max features = 'sqrt' and bootstrap=True, with random state = 42. The final RF model was trained on the complete training set using these hyperparameters and predictions for the test images were generated through majority voting across the trees:

$$y_{\text{pred}} = \text{mode}\{\text{Tree}_1(F_{PCA}), \text{Tree}_2(F_{PCA}), \ldots, \text{Tree}_T(F_{PCA})\} \qquad (6)$$

where T = 300. By integrating complementary deep features, PCA-based compression and a robust ensemble classifier, the feature-level RF approach enhances both accuracy and generalization performance. The workflow is summarized in Fig. 4 and Algorithm 2.

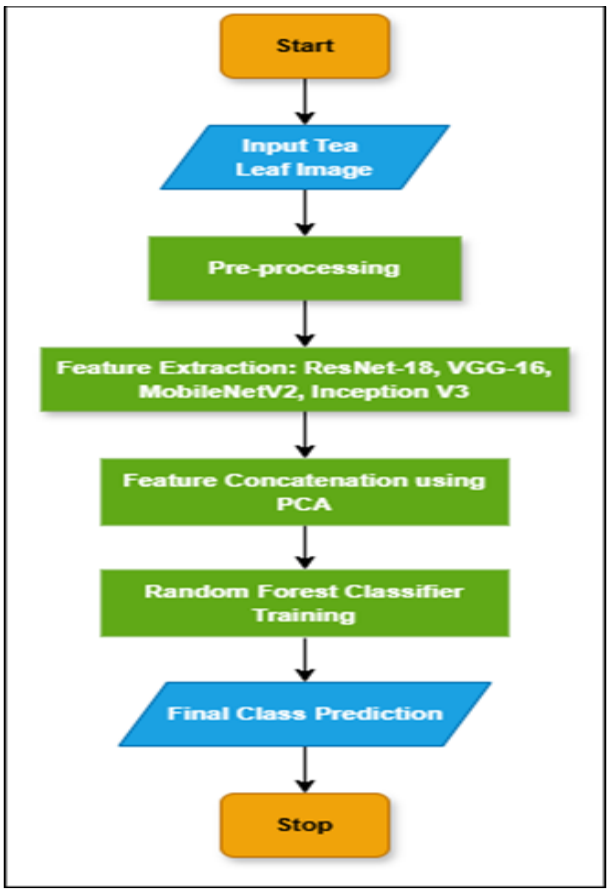

Fig. 4. Workflow of the feature-level RF ensemble.

Performance evaluation metrics

The proposed dual-strategy tea leaf disease classification framework, comprising an output- level ensemble and a feature-level RF ensemble, was comprehensively evaluated to measure its effectiveness in accurately identifying eight tea leaf categories (seven diseases and one healthy class).

Given the multi-class nature of the problem, the evaluation metrics were calculated using macro-averaging, ensuring that all classes contributed equally to the overall performance regardless of the number of samples in each class. This prevents dominant classes from disproportionately influencing the evaluation and allows for a fair assessment across all disease categories.

Algorithm 2. Feature-Level ensemble with PCA and random forest

1. Input: Pre-trained models $M = \{M_1, M_2, M_3, M_4\}$; training dataset $D=\{(x_i, y_i)\}^N i=1$; PCA target dimension d = 300

1. Output: Predicted labels $\hat{y}_i$

2. Feature Extraction:

4. for each $x_i \in D$ do

5. Extract deep features $f^{(1)} \in R^{512}$ from M_1 (ResNet-18)

6. Extract deep features $f^{(2)} \in R^{4096}$ from M_2 (VGG-16)

7. Extract deep features $f^{(3)} \in R^{1280}$ from M_3 (MobileNet V2)

8. Extract deep features $f^{(4)} \in R^{2048}$ from M_4 (Inception V3)

10. Concatenate features: $F_i = [f_i^{(1)} || f_i^{(2)} || f_i^{(3)} || f_i^{(4)} ||] \in R7936$

11. end for

12. Dimensionality Reduction:

13. Apply PCA: $F^{PCA} = PCA(F_i, d)$, where d = 300

14. Training:

15. Train Random Forest classifier RF on reduced features F^{PCA}

16. Inference:

17. Predict labels $\hat{y}_i$ for test samples using trained RF

To obtain a comprehensive performance matrix of the proposed classification procedure, a combination of standard and diagnostic evaluation metrics was used. Accuracy provides a global measure of correct predictions, while Precision, Recall and F1-Score offer deeper insights into the trade-offs between false positives and false negatives for each category. In addition, an analysis of the confusion matrix was performed to visually inspect class-level prediction patterns, allowing the identification of specific misclassifications trends, particularly between diseases with similar visual symptoms. This multifaceted evaluation approach ensures that the reported performance reflects not only the overall accuracy of the model's predictions but also its reliability in classifying all eight tea leaf categories with balanced sensitivity and specificity.

Quantitative classification metrics - accuracy, precision, recall and F1-score

Accuracy quantifies the proportion of correctly classified samples out of the total.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (7)$$

where, TP, TN, FP and FN denote true positives, true negatives, false positives and false negatives respectively.

Precision measures the fraction of predicted positive cases that are actually correct:

$$Precision = \frac{TP}{TP+FP} \quad (8)$$

Recall (or sensitivity) indicates the proportion of actual positive cases that are correctly identified:

$$Recall = \frac{TP}{TP+FN} \quad (9)$$

F1-Score, the harmonic mean of precision and recall, balances both metrics:

$$\text{F1-Score} = 2 \times \frac{Precesio \times \text{Recall}}{Precesio + \text{Recall}} \quad (10)$$

For the four-class classification task, the macro-averaged versions of these measures are expressed as:

$$Macro\_Precision = \frac{1}{C} \sum_{i=1}^{C} \text{Precision}_i \quad (11)$$

$$Macro\_Recall = \frac{1}{C} \sum_{i=1}^{C} Recall_i \quad (12)$$

$$Macro\_F1 = \frac{1}{C} \sum_{i=1}^{C} F1 - Score_i \quad (13)$$

where C = 8 is the total number of classes in the dataset.

The comparative performance of individual CNN architectures and the two proposed ensemble strategies was quantitatively assessed using macro-averaged accuracy, precision, recall and F1-score (Table 2).

From the results, it is evident that the ensemble strategies significantly outperform individual base learners. The output-level ensemble achieved the highest overall performance, with a macro-accuracy of 98.3 %, macro-precision of 98.1 %, macro-recall of 98.2 % and macro F1-score of 98.2 %. The feature-level RF ensemble also delivered strong results, surpassing all single CNN models.

Confusion matrix analysis

Fig. 5(a) and 5(b) show that the output-level ensemble consistently outperforms the feature-level RF ensemble across all eight disease categories, with the largest gains in Red Leaf Spot and White Spot and at least a +2 improvement for Healthy. class-wise accuracy remains 2-3 % higher for the Output-Level model in most cases. The confusion matrices illustrate this advantage: predictions from the output-level ensemble are concentrated along the diagonal, indicating strong classification confidence, whereas the feature-level ensemble shows more off-diagonal errors, especially between visually similar diseases such as Gray Blight and Anthracnose.

Results and Discussion

Comparative analysis of ensemble strategies

In this study, two ensemble strategies were evaluated for the 8-class tea leaf disease classification task: Feature-level fusion using a RF classifier and output- level fusion by probability averaging of individual CNN model predictions. The goal was to identify the approach that delivers the best trade-off between precision, robustness and computational efficiency for agricultural disease detection.

The results revealed that the output-level ensemble approach consistently outperformed the feature-level RF method across all performance metrics. Specifically, accuracy increased from 95.6 % to 98.3 % (+2.7 %), with corresponding improvements in precision (+2.9 %), recall (+2.8 %) and F1-score (+2.9 %). These gains indicate a stronger ability to generalize to unseen data, attributed to the averaging mechanism's capacity to reduce individual model bias and prediction variance.

As shown in Table 3, the output-level ensemble achieves the highest performance across all evaluation metrics.

The findings highlight that the output-level ensemble is the most effective strategy, offering higher predictive accuracy, greater robustness and lower computational complexity. While feature-level RF can be valuable for feature interpretability, its lower recall and higher processing requirements make it less suitable for real-time or embedded agricultural applications. In contrast, probability averaging provides an optimal combination of performance and efficiency, making it ideal for mobile-based or on -field disease detection systems.

Fig. 6 shows a grouped bar chart comparing accuracy, precision, recall and F1- score for all individual CNN base models as well as the two ensemble strategies. It clearly illustrates the performance advantage of the output-Level probability averaging, demonstrating consistently superior performance across all four evaluation metrics.

Table 2. Performance comparison of individual CNN models and the proposed ensemble methods based on macro-averaged metrics

Method	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
ResNet-18	88.0	87.5	87.9	87.7
VGG-16	84.5	83.2	83.7	83.4
MobileNetV2	86.1	85.7	85.4	85.6
Inception V3	87.3	86.9	87.0	87.0
Feature-level ensemble (RF)	95.6	95.2	95.4	95.3
Output-level ensemble (Avg)	98.3	98.1	98.2	98.2

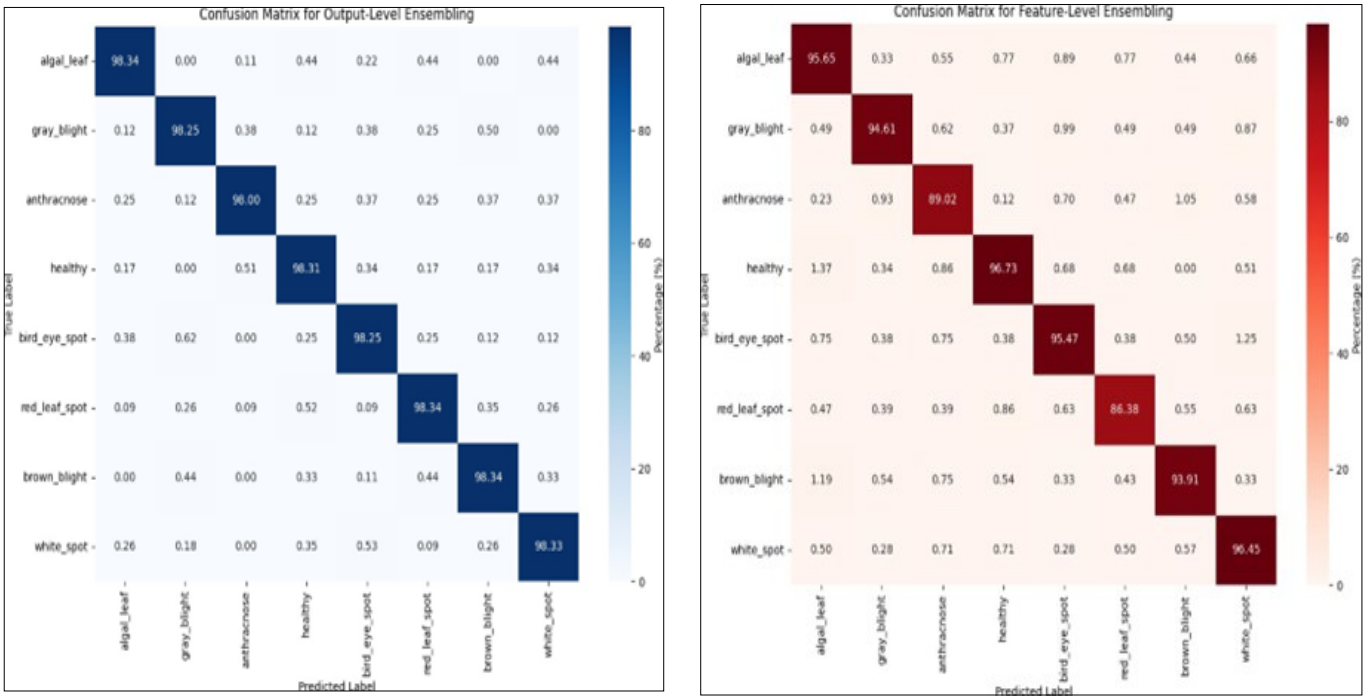

Fig. 5. Performance comparison of the proposed ensembles: (a) Confusion matrix for Output-Level Ensemble, (b) Confusion matrix for Feature-Level Ensemble.

Table 3. Performance comparison between ensemble strategies

Metric	Feature-level RF	Output-level Avg	Absolute Gain
Accuracy	95.6	98.3	+2.7
Precision	95.2	98.1	+2.9
Recall	95.4	98.2	+2.8
F1-score	95.3	98.2	+2.9

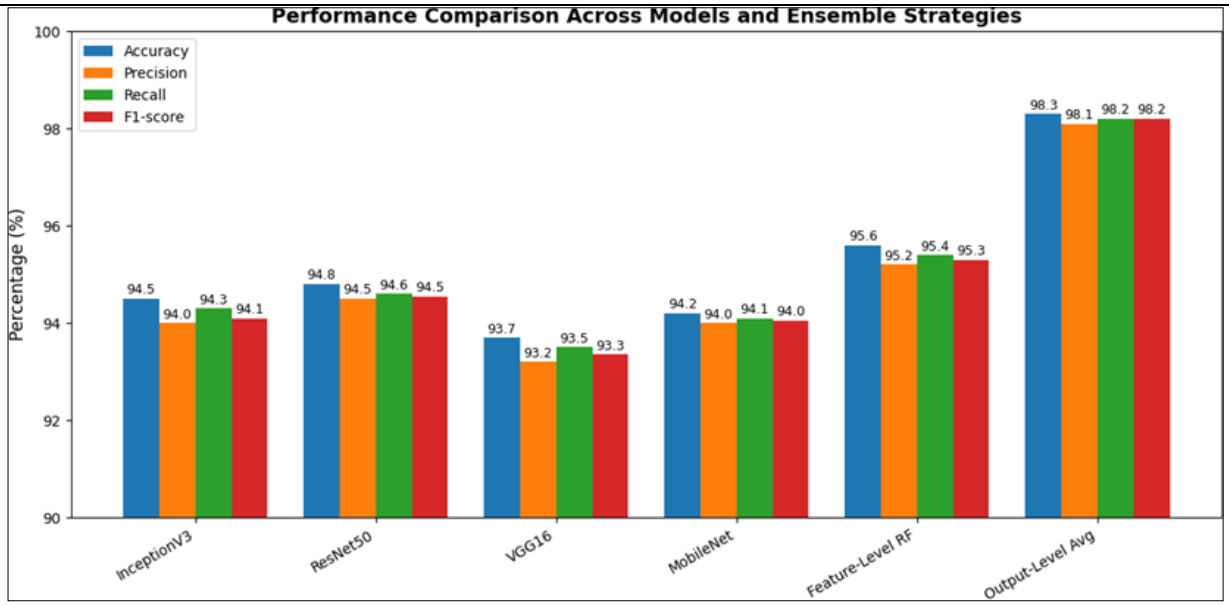

Fig. 6. Grouped bar chart comparing Accuracy, Precision, Recall and F1-score across base models and ensemble strategies.

Ablation study and model interpretability

To evaluate the impact of different combinations of models in the probability averaging ensemble, an ablation study was conducted. Table 4 summarizes the accuracy achieved by different ensemble combinations. Results showed that while two-model ensembles achieved accuracies above 91 %, performance consistently improved with the addition of more models. The best result of 98.3 % accuracy was obtained by combining all four models-ResNet-18, VGG-16, MobileNetV2 and InceptionV3-highlighting the complementary strengths of each architecture. This indicates that a diversified ensemble significantly boosts classification performance.

Table 4. Ablation study of probability averaging ensemble with different model combinations

Model combination	No. of models	Accuracy (%)
ResNet-18 + VGG-16	2	91.2
ResNet-18 + MobileNetV2	2	92.5
ResNet-18 + InceptionV3	2	92.0
VGG-16 + MobileNetV2 + InceptionV3	3	94.6
ResNet-18 + VGG-16 + MobileNetV2	3	95.1
ResNet-18 + MobileNetV2 + InceptionV3	3	96.2
ResNet-18 + VGG-16 + MobileNetV2 + InceptionV3	4	98.3

Visualization of model decisions

In addition to performance evaluation, interpretability was enhanced through Grad-CAM visualizations (30, 31). Fig. 7 presents the heatmaps generated using Grad-CAM implementation. These heatmaps revealed that the model consistently focuses on the diseased areas of the leaf, with red regions indicating the most critical features for decision-making. Correct classifications show clear alignment between the highlighted zones and visible symptoms, while even in misclassified cases, the model's attention remains on relevant leaf areas. This interpretability not only increases trust in the model but also supports its practical deployment in tea plantations for real-time disease monitoring.

Performance comparison with existing work

Table 5 compares the proposed model with recent approaches for tea leaf disease classification. An earlier study reported an accuracy of 96 % using an ensemble approach on the Tea Sickness dataset (23). Another study achieved 92.47 % with a hybrid pooling-based CNN strategy (9), while a separate work obtained 97 % using a transfer learning approach (30). Other researchers used a hybrid CNN–RF methodology and reported 96 % accuracy (12). Additionally, 96.67 % accuracy was achieved using an SVM classifier with VGG-16 deep features (10). In comparison, the proposed model using output-level ensemble attained an accuracy of 98.30 %, outperforming the existing methods, as summarized in Table 5.

Conclusion

The tea leaf disease classification system developed in this work shows that combining deep learning models with feature reduction techniques can offer practical and reliable results. By aligning features extracted from different CNN models, applying PCA to remove redundant information and using a Random Forest classifier, the system achieved an accuracy of 95.6 %. This performance further improved to 98.3 % when a probability-based ensemble was applied. Such high accuracy is valuable in real agricultural settings, where early and accurate disease detection can help prevent crop losses and support tea growers in taking timely and effective actions. The outcomes also highlight the potential of the system to support precision agriculture by assisting farmers and field experts in monitoring large tea estates and reducing dependence on manual inspection. This, in turn, lowers the time and cost associated with traditional disease diagnosis procedures. The interpretability provided by Grad-CAM further enhances the trust and usability of the system, making it suitable for integration into mobile-based tools, drone-surveillance applications and other plantation-monitoring platforms.

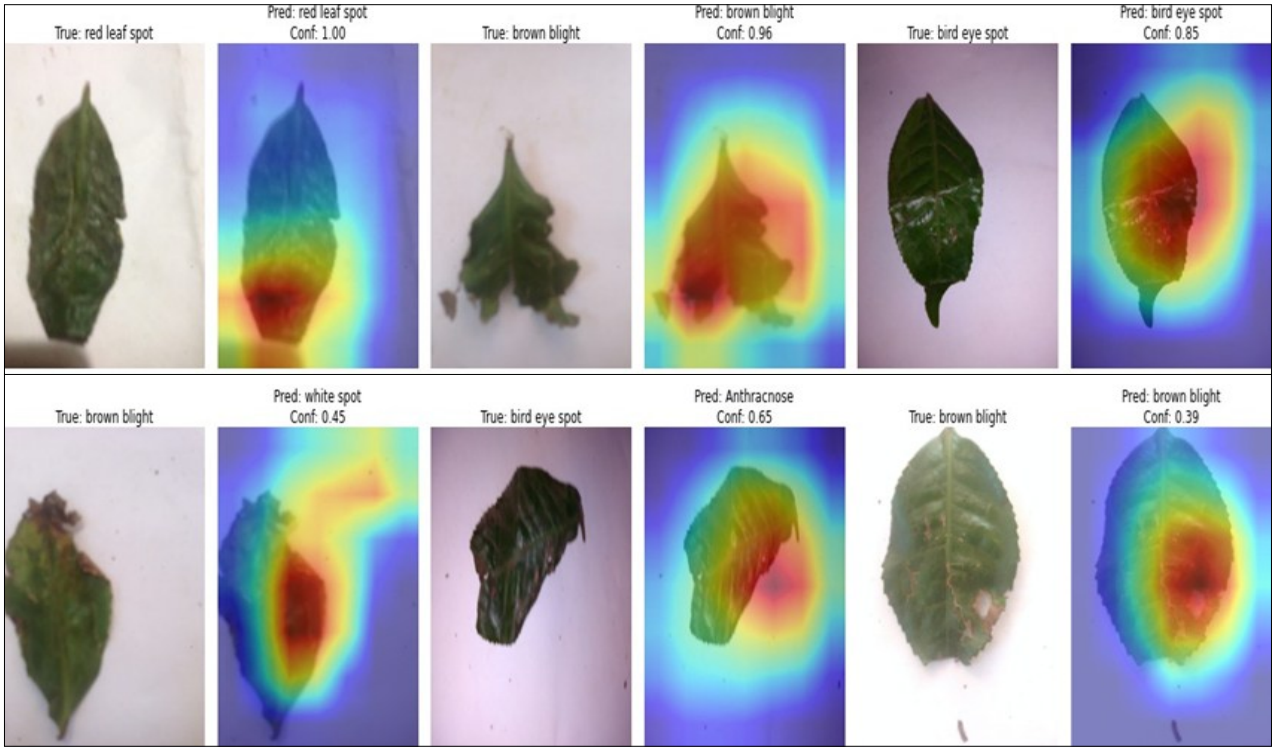

Fig. 7. Grad-CAM visualizations highlighting key regions influencing classification decisions. Red areas indicate high importance, while blue areas are less influential.

Table 5. Comparison of the proposed model with recent tea leaf disease classification approaches

Sl. No.	Authors & Ref	model	Accuracy (%)
1	Ozturk et al. (23)	Ensemble	96
2	Heng et al. (9)	Hybrid Pooling-based CNN	92.47
3	Dipti et al. (30)	Transfer Learning	97
4	Raj et al. (12)	CNN + RF	96
5	Bhagat et al. (10)	SVM + VGG16	96.67
6	Proposed Model	Output-Level Ensemble	98.30

However, some challenges remain. The dataset lacked balance, as certain disease categories had fewer samples, which was addressed through targeted augmentation techniques. Training and fine-tuning multiple deep learning models in parallel also required high computational resources, but efficient scheduling and resource management helped manage the process. Additionally, adapting the proposed methodology to other crops will require adjustments to account for variations in leaf morphology and disease characteristics.

Future work will focus on expanding and diversifying the dataset, developing lightweight models suited for field deployment and extending the methodology for multi-crop disease diagnosis. These efforts aim to improve scalability and make the system more practical, field-ready and robust as a digital decision-support tool for modern agriculture.

Acknowledgements

The authors gratefully acknowledge the Faculty of Computer Technology and the Faculty of Agricultural Sciences and Technology, Assam down town University, Guwahati, for providing invaluable research facilities and technical support, which significantly contributed to the successful completion of this study.

Authors' contributions

KK contributed to the conceptualization of the study, experimental design, data analysis and preparation of the manuscript. KM was involved in the development of the methodology and analysis of the results. SKB conducted the literature review and data collection. AD assisted in interpreting the results and drafting the figures and tables. SB carried out critical revisions, provided editorial input. All authors read and approved the final version of the manuscript.

Compliance with ethical standards

Conflict of interest: Authors do not have any conflict of interest to declare.

Ethical issues: None

References

1. Hajra NG Hajra NG. Indian tea: robust growth amid rising challenges. J Tea Sci Res. 2021;11:1.
2. Pandey AK, Sinniah GD, Babu A, Tanti A. Global tea industry response to fungal diseases: challenges and opportunities. Plant Dis. 2021;105(7):1868-79. https://doi.org/10.1094/PDIS-09-20-1945-FE
3. Soeb MJA, et al. Tea leaf disease detection and identification based on YOLOv7. Sci Rep. 2023;13:6078. https://doi.org/10.1038/s41598-023-33270-4
4. Lu J, Tan L, Jiang H. Convolutional neural networks for plant leaf disease classification: a review. Agriculture. 2021;11(8):707. https://doi.org/10.3390/agriculture11080707
5. Guth FA, Ward S, McDonnell K. Generalization of CNNs for crop disease detection from lab to field. Eur J Eng Technol Res. 2023;8 (2):33-40. https://doi.org/10.24018/ejeng.2023.8.2.2773
6. Pandian JA, Nisha SN, Kanchanadevi K, Pandey AK, Rima SK.

Grey blight disease detection on tea leaves using deep CNN. Comput Intell Neurosci. 2023;2023:7876302. https://doi.org/10.1155/2023/7876302

7. Xue Z, Xu R, Bai D, Lin H. YOLO-tea: tea disease detection using improved YOLOv5. Forests. 2023;14(2):415. https://doi.org/10.3390/f14020415

8. Sarker SS, Islam A, Raktim RT, Roshni SA, Joy SK, Shah FM. Real-time tea leaf disease detection using deep learning models. Proc ICCIT. 2024:197-202. https://doi.org/10.1109/ICCIT64611.2024.11021943

9. Heng Q, Yu S, Zhang Y. AI-based tea leaf disease identification using hybrid pooling deep networks. Heliyon. 2024;10(5):e26465. https://doi.org/10.1016/j.heliyon.2024.e26465

10. Bhagat M, Kumar D. Kernelized SVM with deep learning features for tea leaf disease prediction. Multimed Tools Appl. 2024;83(13):39117-34. https://doi.org/10.1007/s11042-023-17172-1

11. Xia Y, Yuan W, Zhang S, Wang Q, Liu X, Wang H, et al. Tea disease identification using improved YOLOv7-MobileNeXt. Sci Rep. 2024;14:11799. https://doi.org/10.1038/s41598-024-62451-y

12. Raj M, Jha P, Magar MG, Kukreja V. CNN-random forest hybrid model for tea leaf disease classification. Proc AUTOCOM. 2024:52-56. https://doi.org/10.1109/AUTOCOM60220.2024.10486177

13. Zubair F, et al. Robust ensemble deep learning model for plant disease detection. AgriEngineering. 2025;7(5):159. https://doi.org/10.3390/agriengineering7050159

14. Alhichri H. RS-DeepSuperLearner: CNN ensemble fusion for remote sensing classification. Ann GIS. 2023;29(1):121-42. https://doi.org/10.1080/19475683.2023.2165544

15. Kumar Prusty CA, et al. Plant disease detection from leaf images using ensemble CNN. Proc ICAIHC. 2025:1-6. https://doi.org/10.1109/ICAIHC64101.2025.1095748

16. Shovon MS, Mozumder SJ, Pal OK, Mridha MF, Asai N, Shin J. PlantDet: multi-model ensemble deep learning for plant disease detection. IEEE Access. 2023;11:34846-59. https://doi.org/10.1109/ACCESS.2023.3264835

17. Chug A, Bhatia A, Singh AP, Singh D. Hybrid deep learning framework for image-based plant disease detection. Soft Comput. 2023;27(18):13613-38. https://doi.org/10.1007/s00500-022-07177-7

18. Khalid M, Talukder MA. Hybrid deep multistacking model for plant disease detection. IEEE Access. 2025;13:1-14. https://doi.org/10.1109/ACCESS.2025.3647467

19. Saberi Anari M. Hybrid transfer learning and ensemble model for leaf disease classification in agricultural AIoT. Comput Intell Neurosci. 2022;2022:6504616. https://doi.org/10.1155/2022/6504616

20. Taji K, Sohail A, Shahzad T, Khan BS, Khan MA, Ouahada K. Ensemble hybrid CNN features for plant disease classification. IEEE Access. 2024;12:61886-906. https://doi.org/10.1109/ACCESS.2024.3389648

21. Shrotriya A, Sharma AK, Bairwa AK, Manoj R. Hybrid CNN-RNN ensemble for cotton plant disease detection. IEEE Access. 2024;12:1-15. https://doi.org/10.1109/ACCESS.2024.3515843

22. Fayyaz AM, et al. Grad-CAM: a systematic review. Comput Biol Med. 2025;198:111200. https://doi.org/10.1016/j.compbiomed.2025.111200

23. Ozturk O, Sarica B, Seker DZ. Interpretable ensemble deep learning for tea leaf disease classification. Horticulture. 2025;11(4):437. https://doi.org/10.3390/horticulturae11040437

24. Barbiero P, Zarlenga ME, Termine A, Jamnik M, Marra G. Foundations of interpretable models. arXiv. 2025;2508:00545.

25. Oad A, et al. Plant leaf disease detection using ensemble learning and explainable AI. IEEE Access. 2024;12:156038-49. https://doi.org/10.1109/ACCESS.2024.3484574

26. Kimutai G, Forster A. Tea sickness dataset. Mendeley Data. 2022;V2.

27. Shorten C, Khoshgoftaar TM. Survey on image data augmentation for deep learning. J Big Data. 2019;6:1-48. https://doi.org/10.1186/s40537-019-0197-0

28. Zheng Q, Yang M, Tian X, Jiang N, Wang D. Full-stage data augmentation in CNNs for image classification. Discrete Dyn Nat Soc. 2020;2020:4706576. https://doi.org/10.1155/2020/4706576

29. Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L. ImageNet: large-scale hierarchical image database. Proc CVPR. 2009:248-55. https://doi.org/10.1109/CVPR.2009.5206848

30. Dipty I, Assaduzzaman M, Fahad N, Hossen MJ, Haider MF, Rahman F. TeaNet8: Android-based tea leaf disease detection using transfer learning and Grad-CAM. Results Control Optim. 2025;100577. https://doi.org/10.1016/j.rico.2025.100577

31. Ahammed F, Shikdar OF, Alam BS, Jahan S, Kibria G, Ali NY. Classification of nutritional deficiencies in coffee leaf using transfer learning and Grad-CAM. Proc ICCCT. 2025:1-5. https://doi.org/10.1109/ICCCT63501.2025.11019090

Additional information

Peer review: Publisher thanks Sectional Editor and the other anonymous reviewers for their contribution to the peer review of this work.

Reprints & permissions information is available at https://horizonepublishing.com/journals/index.php/PST/open_access_policy

Publisher's Note: Horizon e-Publishing Group remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Indexing: Plant Science Today, published by Horizon e-Publishing Group, is covered by Scopus, Web of Science, BIOSIS Previews, Clarivate Analytics, NAAS, UGC Care, etc
See https://horizonepublishing.com/journals/index.php/PST/indexing_abstracting

Publisher information: Plant Science Today is published by HORIZON e-Publishing Group with support from Empirion Publishers Private Limited, Thiruvananthapuram, India.