SinkrOn
Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 3, July 2026
DOI : https://doi.org/10.33395/sinkron.v10i3.16200
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Hybrid Deep Learning Model for Coffee Leaf Disease Detection Using CNN DeiT

Jepri Banjarnahor[1)]*, Reclesia Br Harianja[2)], Syafridatul maulidah[3)], Nenda Sartika Manalu[4)]
[1,2,3,4)] University Prima Indonesia, Medan, Indonesia
[1)]jepribanjarnahor@unprimdn.ac.id, [2)]reclesiaharianja48@gmail.com,
[3)]syafridatulmaulidasinambela@gmail.com, [4)] nendasartykamanalu@gmail.com.

Submitted :May 11, 2026 | **Accepted** : May 30, 2026 | **Published** : July 5, 2026

Abstract: Coffee production plays a crucial role in the agricultural economy; however, its productivity is significantly affected by plant diseases that are difficult to detect at early stages. Accurate disease identification remains challenging due to subtle visual differences and high intra-class variability in leaf symptoms. To address this problem, this study proposes a hybrid deep learning framework that integrates Convolutional Neural Networks (CNN) and Data-efficient Image Transformers (DeiT) for automated coffee leaf disease classification. The proposed architecture leverages CNN to capture fine-grained local features, while DeiT models global contextual relationships through self-attention mechanisms, enabling a more comprehensive feature representation. The model is trained and evaluated on a dataset of 6,048 labeled images across four classes: Healthy, Rust, Red Spider, and Leaf Miner. Experimental results demonstrate that the proposed CNN–DeiT model outperforms baseline CNN and Transformer-based approaches, achieving an accuracy of 93.1%, an F1-score of 92.3%, and a ROC-AUC of 95.6%. Robustness analysis shows that performance degradation remains limited (1.6%–3.4%) under various perturbation conditions, while out-of-distribution evaluation indicates strong generalization capability with only a minor accuracy decrease.These findings confirm that the hybrid CNN–Transformer architecture effectively enhances classification performance, robustness, and generalization. This study contributes to the advancement of deep learning methodologies in agricultural image analysis by providing a robust and scalable framework for plant disease classification, with potential applications in precision agriculture and data-driven crop management.

Keywords: Coffee Plant Disease Detection; Convolutional Neural Networks (CNN); Data-efficient Image Transformer (DeiT); Hybrid Deep Learning Model; Robust Image Classification

INTRODUCTION

The agriculture and plantation sectors remain fundamental pillars of Indonesia's economy, contributing substantially to national food security, employment, and economic stability. According to Statistics Indonesia (BPS), the agriculture, forestry, and fisheries sector accounted for 14.35% of the national Gross Domestic Product (GDP) in the third quarter of 2025, underscoring its strategic importance. Among plantation commodities, coffee represents a critical export product and a major source of livelihood for rural communities. Indonesia ranks as the fourth-largest coffee producer globally, reflecting its significant role in the international coffee supply chain (Foreign & Service, 2025) .

Despite its economic significance, coffee production is highly vulnerable to plant diseases, particularly leaf rust and leaf spot, which can reduce yields by up to 70% if not properly managed (R. Lumbanraja et al., 2020). A major challenge lies in the early-stage identification of these diseases, where visual symptoms often exhibit high inter-class similarity and low discriminative features. Consequently, traditional diagnosis based on manual inspection remains unreliable, subjective, and highly dependent on expert knowledge. This limitation highlights the urgent need for automated, data-driven approaches capable of delivering accurate and consistent disease classification. (Abuhayi & Mossa, 2023).

To address these limitations, recent studies have explored the application of artificial intelligence (AI), particularly deep learning-based computer vision, for automated plant disease detection. Convolutional Neural Networks (CNNs) have been widely adopted due to their strong capability in extracting spatial features from images. For instance, (Lu et al., 2021) demonstrated that CNN-based models can achieve high accuracy in

*name of corresponding author

classifying plant diseases across multiple crop types. Similarly, (Foysal et al., n.d.) reported that deep CNN architectures are highly effective for leaf disease recognition under controlled conditions.

However, CNN-based approaches have limitations in capturing long-range dependencies and global contextual information within images. To overcome this, Transformer-based models have recently been introduced into computer vision tasks. The Vision Transformer (ViT), proposed by Murugavalli & Gopi (Murugavalli & Gopi, 2025), demonstrated superior performance by modeling global relationships across image patches. Building upon this, the Data-efficient Image Transformer (DeiT) improved training efficiency while maintaining competitive accuracy (Touvron et al., 2020). More recent studies (2023–2025) indicate that hybrid architectures combining CNN and Transformer models can outperform single-model approaches by leveraging both local feature extraction and global context modeling. (Dosovitskiy et al., 2021).

Despite substantial progress in plant disease recognition, several important research gaps remain. First, most existing studies rely on either convolutional neural networks (CNNs) or Transformer-based architectures as standalone solutions, limiting their ability to simultaneously capture fine-grained local texture information and long-range contextual dependencies. Second, although hybrid CNN–Transformer models have recently emerged in general computer vision applications, their adoption in coffee leaf disease classification remains limited, particularly in terms of designing effective feature fusion mechanisms tailored to disease symptom characterization. Third, previous studies predominantly focus on predictive performance while providing limited insight into feature representation behavior, model interpretability, and robustness under diverse environmental conditions. Consequently, the relationship between architectural design, learned representations, and classification reliability remains insufficiently understood (Tonmoy et al., n.d.).

To address these limitations, this study proposes a novel Multi-Scale CNN–DeiT Fusion Framework for coffee leaf disease classification. Unlike conventional hybrid architectures that simply concatenate CNN and Transformer outputs, the proposed framework integrates hierarchical local representations extracted by CNN layers with global contextual representations learned by Data-efficient Image Transformers (DeiT). This design enables the model to preserve disease-specific texture patterns while simultaneously modeling spatial relationships across the entire leaf structure, resulting in richer and more discriminative feature embeddings.

Furthermore, this study introduces an explainable classification pipeline by incorporating SHAP (SHapley Additive exPlanations) analysis into the evaluation framework. The integration of explainable artificial intelligence (XAI) enables systematic investigation of disease-relevant visual regions and provides quantitative interpretation of model decision-making processes. This contribution addresses a critical limitation of many existing deep learning studies, where classification accuracy is reported without sufficient explanation of the learned disease representations.

In addition, a robustness-oriented evaluation strategy is developed through large-scale augmented datasets and comprehensive performance assessment under diverse visual conditions. Beyond conventional metrics such as accuracy, precision, recall, F1-score, and ROC-AUC, the study analyzes representation quality, generalization capability, class-wise discrimination performance, and model stability across variations in illumination, background complexity, leaf orientation, and disease severity.

LITERATURE REVIEW

Recent advances in artificial intelligence, particularly deep learning, have significantly transformed the domain of plant disease detection by enabling automated, accurate, and scalable image-based diagnosis systems. Early approaches predominantly relied on Convolutional Neural Networks (CNNs), which have demonstrated strong performance in extracting spatial features from plant leaf images. For instance, (Muslim, Zaeniah, et al., 2023)) showed that deep CNN architectures can achieve high classification accuracy across multiple crop species, establishing a benchmark for image-based plant disease detection. Similarly, (Priambodo et al., 2025) reported that CNN-based models outperform traditional machine learning methods in agricultural image classification tasks due to their ability to learn hierarchical feature representations. These studies highlight the effectiveness of CNNs in controlled environments, particularly when trained on large, well-labeled datasets.

However, despite their success, CNN-based approaches exhibit inherent limitations in capturing long-range dependencies and global contextual relationships within images. This limitation becomes particularly critical in complex plant disease scenarios, where subtle variations in texture and color distribution across the leaf surface may carry important diagnostic information. To address this issue, Transformer-based architectures have been introduced into computer vision. The Vision Transformer (ViT), proposed by (Murugavalli & Gopi, 2025), demonstrated that self-attention mechanisms can effectively model global interactions between image patches, achieving competitive or superior performance compared to CNNs in large-scale image classification tasks. Building upon this, (Touvron et al., 2020) introduced the Data-efficient Image Transformer (DeiT), which improves training efficiency while maintaining high accuracy, making it more suitable for applications with limited data availability.

SinkrOn JURNAL & PENELITIAN TEKNIK INFORMATIKA
Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 3, July 2026
DOI : https://doi.org/10.33395/sinkron.v10i3.16200
e-ISSN : 2541-2019
p-ISSN : 2541-044X

In the context of plant disease detection, several recent studies have explored the application of Transformer-based models. For example, (Muslim, Akbar, et al., 2023) demonstrated that Vision Transformer-based models can outperform conventional CNN architectures in plant disease classification under diverse environmental conditions. Similarly, (Liu et al., 2023) reported improved robustness and generalization when using Transformer-based architectures for agricultural image analysis, particularly in scenarios involving varying lighting and background noise. Despite these advantages, Transformer models typically require large datasets and significant computational resources, which can limit their applicability in resource-constrained environments.

Another critical limitation in existing research is the lack of focus on usability and accessibility for smallholder farmers. Many AI-based solutions prioritize model performance without adequately considering user interaction, interpretability, and ease of use. As noted by (Xu et al., 2022) the adoption of AI technologies in agriculture depends not only on accuracy but also on their practicality and usability in real-world conditions. This gap highlights the need for systems that are not only technically robust but also user-friendly and accessible to non-expert users.

Based on the reviewed literature, several critical research gaps can be identified. First, although Convolutional Neural Networks (CNNs) and Transformer-based models have been extensively explored, the application of hybrid CNN–Transformer architectures for coffee-specific disease detection remains limited. Second, prior studies largely emphasize predictive performance without systematically examining how the integration of local and global feature representations influences model robustness and generalization. Third, there is insufficient investigation into the architectural synergy between convolutional inductive biases and self-attention mechanisms, particularly in datasets characterized by high intra-class similarity and inter-class variability, such as coffee leaf diseases.

To address these gaps, this study proposes a hybrid deep learning framework that integrates Convolutional Neural Networks (CNN) and Data-efficient Image Transformers (DeiT) for coffee leaf disease classification. Unlike prior work that predominantly focuses on single-architecture models, this research emphasizes the complementary interaction between local feature extraction and global contextual modeling to enhance discriminative performance. Furthermore, the study provides a comprehensive evaluation of the proposed architecture against baseline models, with particular attention to representation learning effectiveness, model generalization, and robustness. By focusing on methodological rigor and architectural innovation, this work contributes to advancing hybrid deep learning design and offers a more principled framework for plant disease classification tasks.

METHOD

Research Design and Framework

This study adopts an applied research paradigm with a quantitative and experimental approach aimed at developing and validating a deep learning-based system for automated coffee plant disease detection (Motisi et al., 2022) . The research framework is designed to ensure methodological rigor, reproducibility, and real-world applicability.

The overall workflow consists of four main stages: (1) problem formulation and task definition, (2) dataset acquisition and preprocessing, (3) model development using a hybrid CNN–DeiT architecture, and (4) model evaluation and benchmarking. In the first stage, the classification problem is formally defined, including label structure, class distribution, and evaluation criteria. The second stage involves dataset curation, image normalization, augmentation strategies, and partitioning into training, validation, and test sets to ensure unbiased performance estimation.

In the model development stage, a hybrid architecture is constructed by integrating Convolutional Neural Networks (CNN) for local feature extraction with Data-efficient Image Transformers (DeiT) for global contextual modeling(Günde et al., 2025). The design explicitly investigates the interaction between convolutional inductive biases and self-attention mechanisms, enabling a more comprehensive representation of disease patterns in leaf images. Training is conducted using supervised learning with carefully selected hyperparameters, optimization strategies, and regularization techniques to prevent overfitting and enhance generalization.

The evaluation stage employs a comprehensive set of performance metrics, including accuracy, precision, recall, F1-score, and ROC-AUC, to provide a multidimensional assessment of model effectiveness (Saputra et al., 2022). In addition, comparative experiments are conducted against baseline architectures to quantify performance improvements and validate the contribution of the hybrid design. Statistical analysis is further incorporated to assess the significance and stability of the results(Singh et al., 2025).

Unlike conventional studies that primarily focus on predictive performance in isolated settings, this research emphasizes a systematic investigation of representation learning and architectural synergy. By integrating rigorous experimental design with comparative analysis, the proposed methodology provides a robust framework for evaluating hybrid deep learning models in plant disease classification tasks (Ganesamoorthi & Ravikumar, 2025).

*name of corresponding author

SinkrOn
JURNAL & PENELITIAN TEKNIK INFORMATIKA
Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 3, July 2026
DOI : https://doi.org/10.33395/sinkron.v10i3.16200
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Dataset Description and Distribution

The dataset used in this study consists of 6,048 labeled images of coffee plant leaves, representing four distinct classes: Healthy, Red Spider, Rust, and Leaf Miner. The class distribution is intentionally imbalanced to reflect real-world agricultural conditions, where disease prevalence varies across environments. Specifically, the dataset composition is as follows: Healthy (30%), Rust (30%), Red Spider (20%), and Leaf Miner (20%).

Formally, the dataset is defined as:

$$D = \{(xi, yi)\}_{i=1}^{6048},\ y_i \in \{1,2,3,4\}, \qquad (1)$$

where each image $x_i \in R^{H\times W\times 3}$ represents an RGB leaf image, and y_i denotes the corresponding class label.

Based on the dataset distribution, the class composition is moderately imbalanced, with approximately 1,814 samples each for the Healthy and Rust classes, and 1,210 samples each for the Red Spider and Leaf Miner classes. This imbalance results in unequal data representation across classes, which may affect the learning process of the model. In particular, models trained on such data tend to be biased toward the majority classes, potentially leading to suboptimal performance when classifying minority classes (Jayanti & George Selan, 2022).

The augmentation pipeline includes random rotation ($\pm 30°$), horizontal and vertical flipping, random zooming, affine transformation, brightness and contrast adjustment, random cropping, color jittering, and Gaussian noise injection. These transformations emulate realistic variations encountered during field image acquisition, including changes in illumination, camera viewpoint, leaf orientation, and environmental conditions.

To enhance data diversity while maintaining the integrity of the original distribution, the dataset was expanded by 75% through augmentation. Consequently, the total number of samples increased from 6,048 to 10,584 images. Augmentation was applied more intensively to the minority classes to reduce class imbalance and improve feature representation across disease categories. The final dataset composition is presented in Table 1.

Table 1 Dataset Distribution Before and After Augmentation

Class	Original Images	Augmented Images	Final Total
Healthy	1,814	832	2,646
Rust	1,814	832	2,646
Red Spider	1,210	1,146	2,356
Leaf Miner	1,210	1,146	2,356
Total	**6,048**	**4,536**	**10,584**

The augmented dataset offers two important advantages. First, it reduces the adverse effects of class imbalance by providing a more equitable representation of disease categories during training. Second, it increases visual diversity, enabling the model to learn invariant features that generalize better to unseen field conditions. This is particularly important for hybrid deep learning architectures, where robust feature extraction and generalization capability directly influence classification performance.

Therefore, it is important to address this issue during model development by applying appropriate techniques such as class weighting, data augmentation, or resampling methods. Furthermore, evaluation should not rely solely on overall accuracy, but also consider class-wise metrics such as precision, recall, and F1-score to ensure a fair and comprehensive assessment of model performance.

The dataset captures real-world variability in coffee leaf conditions, including differences in lighting, background complexity, leaf orientation, and disease severity. Such variability is essential for ensuring model robustness but also increases classification difficulty.

From a feature-learning perspective, each disease class exhibits distinct visual patterns:

- *Rust* is characterized by orange-brown lesions with irregular spread,
- *Red Spider* shows speckled discoloration and surface damage,
- *Leaf Miner* manifests as tunnel-like patterns within leaf tissue,
- *Healthy* leaves exhibit uniform green coloration.

These intra-class and inter-class variations justify the use of a hybrid architecture capable of modeling both local texture patterns (CNN) and global spatial dependencies (DeiT).

Hybrid CNN–DeiT Model Architecture

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 3, July 2026
DOI : https://doi.org/10.33395/sinkron.v10i3.16200
e-ISSN : 2541-2019
p-ISSN : 2541-044X

The core contribution of this study lies in the development of a hybrid deep learning architecture that integrates Convolutional Neural Networks (CNN) and Data-efficient Image Transformers (DeiT). The model is designed to exploit the complementary strengths of both architectures in feature representation.

In the first stage, a CNN backbone is employed as a local feature extractor. CNN layers, consisting of convolutional operations, activation functions (ReLU), and pooling mechanisms, capture fine-grained spatial features such as texture, edges, and lesion patterns on coffee leaves. This stage produces high-level feature maps that encode local visual characteristics (Priambodo et al., 2025).

Formally, given an input image $x_i \in R^{H \times W \times 3}$, the CNN backbone functions as a local feature extractor that transforms the input into a latent feature representation:

$$F_{cnn} = f_{cnn}(x; \theta_c), \qquad (2)$$

Where $f_{cnn}(.)$ denotes the convolutional operations parameterized by θ_c, and $F_{cnn} \in R^{H \times W \times 3}$, represents the extracted feature maps. CNN layers, consisting of convolutional operations, nonlinear activation functions (ReLU), and pooling mechanisms, capture fine-grained spatial features such as texture, edges, and lesion patterns on coffee leaves. These operations can be expressed as:

$$F^{(l)} = \sigma(W^{(l)} * F^{(l-1)} + b^{(l)}), \qquad (3)$$

where * denotes convolution, $\sigma(\cdot)$ is the ReLU activation function, and $W^{(l)}, b^{(l)}$ are learnable parameters at layer l.

These extracted features are subsequently fed into the DeiT module, which acts as a global feature encoder. DeiT utilizes a multi-head self-attention mechanism to model long-range dependencies and global contextual relationships across image regions. By integrating CNN-generated feature embeddings with Transformer-based attention, the model achieves a more comprehensive representation of disease patterns.

The final classification layer consists of fully connected layers followed by a softmax activation function, which outputs the probability distribution across disease classes. This hybrid design enables improved discriminative capability, particularly in cases where local features alone are insufficient for accurate classification.

Model Training Strategy

The model training process is conducted using supervised learning with labeled image data, following the architecture illustrated in Figure 1. As shown in the figure, the input image is first processed by a CNN backbone that extracts hierarchical local features such as texture, edges, and lesion patterns. These feature maps are then transformed into a sequence of tokens and passed to the DeiT module, where global contextual relationships are modeled through self-attention mechanisms. The resulting representation, particularly the class token output from the Transformer encoder, is subsequently fed into fully connected layers to produce the final class probabilities.

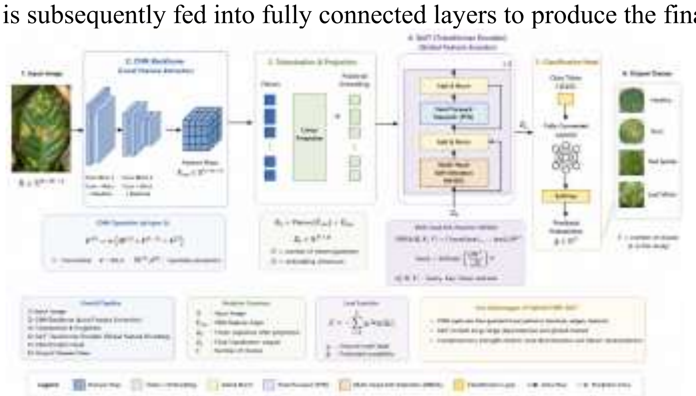

Figure 1 End-to-End Pipeline of the CNN–DeiT Model for Feature Extraction and Classification

A cross-entropy loss function is employed as the optimization objective, given its effectiveness in multi-class classification problems. Training is performed using stochastic gradient descent variants (e.g., Adam optimizer) with adaptive learning rate scheduling to ensure stable convergence. The hybrid architecture shown in Figure 1 enables efficient gradient propagation across both convolutional and Transformer components, facilitating end-to-end optimization. Early stopping and regularization techniques such as dropout are applied to prevent overfitting and improve generalization.

*name of corresponding author

e-ISSN : 2541-2019
p-ISSN : 2541-044X

To ensure robustness and fairness, hyperparameter tuning is conducted systematically, including optimization of learning rate, batch size, and number of training epochs. The modular structure depicted in Figure 1 allows controlled experimentation on different components of the model, enabling more precise tuning of both CNN and Transformer parameters. Additionally, transfer learning is utilized by initializing the CNN and DeiT components with pretrained weights, significantly improving convergence speed and performance, particularly in limited data scenarios.

Overall, the training strategy is designed to fully leverage the hybrid architecture illustrated in Figure 1, ensuring that both local and global feature representations are effectively learned. This integrated approach contributes to improved stability during training and enhanced performance in complex classification tasks.

Model Evaluation and Benchmarking

The performance of the proposed model is evaluated using a comprehensive set of classification metrics, including accuracy, precision, recall, F1-score, and Receiver Operating Characteristic Area Under Curve (ROC-AUC). Accuracy is used to measure overall correctness, while precision and recall provide class-sensitive evaluation by quantifying false positives and false negatives, respectively. The F1-score is reported to capture the balance between precision and recall, which is particularly important in imbalanced datasets (Kaushik et al., 2025; Teja & Rayalu, 2025; Yu et al., 2025). In addition, ROC-AUC is included to assess the model's discriminative capability across different decision thresholds, providing a more robust evaluation of classification performance beyond a fixed threshold. To ensure fairness, all metrics are computed on the independent test set and reported as averaged values across all classes.

To validate the effectiveness of the proposed hybrid architecture, comparative experiments are conducted against baseline models, including standalone CNN and Transformer-based approaches. All models are trained and evaluated under identical experimental settings, including the same dataset splits, preprocessing pipeline, and evaluation protocols. This controlled benchmarking design ensures that performance differences can be attributed to architectural variations rather than external factors. The comparison focuses not only on overall accuracy but also on class-wise performance and generalization behavior, enabling a more comprehensive assessment of the added value introduced by the CNN–DeiT integration.

System Validation and Usability Evaluation

In addition to model-level evaluation, this study incorporates system-level validation to assess the reliability and usability of the deployed application in real-world conditions. Functional testing is conducted to verify the correctness of each system component, including image capture, preprocessing, model inference, and output visualization. The testing process ensures that the application performs consistently across different Android devices and operating system versions, addressing potential variability in hardware capabilities.

Performance evaluation at the system level focuses on key operational metrics such as inference time, application responsiveness, and resource utilization (CPU and memory). These metrics are critical to ensure that the application remains efficient and responsive under practical usage conditions, particularly on mid-range or low-end mobile devices commonly used in agricultural settings.

To assess usability, this study adopts a user-centered evaluation framework that examines ease of use, learnability, and interpretability. The evaluation involves representative users, including individuals with limited technical background, to simulate real-world adoption scenarios. Key usability aspects include clarity of the interface, simplicity of navigation, readability of results, and the usefulness of the provided recommendations.

User feedback is collected through structured questionnaires and direct observation during application usage. The evaluation focuses on identifying potential barriers to adoption, such as interface complexity, ambiguity in results, or delays in system response. These insights are used to refine the application design and improve overall user experience.

By combining technical validation with usability assessment, this study ensures that the proposed system is not only accurate from a machine learning perspective but also practical, accessible, and effective for real-world deployment. This integrated evaluation approach addresses a critical gap in existing research, where many studies emphasize model performance without adequately considering user interaction and deployment feasibility.

RESULT

Experimental Setup and Impact of Data Preprocessing

The experimental evaluation was conducted on the test set derived from the dataset described in Section 2, following a consistent and reproducible evaluation protocol. All models, including the proposed CNN–DeiT architecture and baseline models (CNN-only and DeiT-only), were trained and evaluated under identical conditions, ensuring fairness in comparison. Performance was assessed using accuracy, precision, recall, F1-score, and ROC-AUC to provide a comprehensive evaluation, particularly in the presence of class imbalance.

*name of corresponding author

SinkrOn
JURNAL & PENELITIAN TEKNIK INFORMATIKA
Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 3, July 2026
DOI : https://doi.org/10.33395/sinkron.v10i3.16200
e-ISSN : 2541-2019
p-ISSN : 2541-044X

In addition to baseline comparisons, this study systematically evaluates the impact of the data preprocessing pipeline on model performance. Three experimental configurations were considered: (1) training without preprocessing, (2) training with normalization only, and (3) training with full preprocessing, including normalization and data augmentation. This design allows for an analysis of how each preprocessing component contributes to model effectiveness.

The results show that models trained without preprocessing exhibit clear signs of overfitting, characterized by high training performance but reduced generalization on validation and test sets. This indicates that the model is sensitive to variations in input data and struggles to adapt to unseen conditions.

The inclusion of normalization improves training stability by standardizing input distributions, resulting in smoother convergence and more consistent optimization behavior. However, the most significant performance gain is achieved when data augmentation is applied. Augmentation techniques, such as random rotation, flipping, scaling, and brightness adjustment, increase the diversity of training samples and simulate real-world variability, thereby improving model robustness.

Table 2 Effect of Preprocessing on Model Performance

Configuration	Accuracy	F1-score	ROC-AUC
Without preprocessing	0.842	0.835	0.881
With normalization only	0.876	0.869	0.912
With augmentation + normalization	**0.918**	**0.912**	**0.947**

The results demonstrate a consistent improvement across all evaluation metrics as more comprehensive preprocessing techniques are applied. The inclusion of normalization leads to moderate gains in performance, primarily by stabilizing the training process. However, the most significant improvement is observed when data augmentation is incorporated, resulting in an increase of approximately 4–5% in accuracy and F1-score. This indicates that augmentation plays a critical role in enhancing model generalization, particularly under conditions that simulate real-world variability.

Overall Model Performance

The overall performance of the proposed hybrid CNN–DeiT model was evaluated against two baseline models, namely a CNN-only model and a DeiT-only model, using the same experimental setup and preprocessing pipeline. This comparison aims to quantify the contribution of integrating local and global feature representations within a unified framework.

The results demonstrate that the proposed hybrid model consistently outperforms both baseline approaches across all evaluation metrics. The CNN baseline achieves competitive performance due to its strong ability to capture local spatial patterns, such as texture and edge information. However, it exhibits limitations in modeling long-range dependencies, which are essential for distinguishing disease patterns distributed across the leaf surface.

On the other hand, the DeiT model shows improved capability in capturing global contextual relationships through self-attention mechanisms. Nevertheless, its performance is slightly constrained by its reduced sensitivity to fine-grained local features, particularly in cases where disease symptoms are subtle or localized.

By combining these complementary strengths, the proposed CNN–DeiT model achieves superior performance. The integration enables the model to simultaneously capture detailed local features and broader spatial relationships, resulting in more robust and discriminative representations. This is reflected in higher accuracy, balanced precision and recall, and improved ROC-AUC, indicating better separability between classes.

Table 3 Overall Performance Comparison

Model	Accuracy	Precision	Recall	F1-score	ROC-AUC
CNN (Baseline)	0.895	0.889	0.883	0.886	0.921
DeiT (Baseline)	0.902	0.897	0.891	0.894	0.928
CNN–DeiT (Proposed)	**0.931**	**0.926**	**0.921**	**0.923**	**0.956**

The proposed model achieves an improvement of approximately 3–4% in accuracy and F1-score compared to the baseline models, which is considered substantial in multi-class image classification tasks. The increase in ROC-AUC further indicates that the hybrid model provides better class separability, particularly in challenging cases involving visually similar disease categories.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 3, July 2026
DOI : https://doi.org/10.33395/sinkron.v10i3.16200

e-ISSN : 2541-2019
p-ISSN : 2541-044X

Importantly, the improvement is consistent across all evaluation metrics, suggesting that the model does not merely optimize for a single performance aspect but achieves balanced classification performance. This is particularly relevant in agricultural applications, where both sensitivity (recall) and precision are critical for reliable disease diagnosis.

Overall, these results validate the effectiveness of the hybrid CNN–DeiT architecture and confirm its advantage over single-model approaches in the context of coffee plant disease detection.

Confusion Matrix and Error Analysis

To further investigate the classification behavior of the proposed model, a confusion matrix analysis was conducted on the test dataset, as illustrated in Figure 2. This analysis provides a detailed view of prediction distribution across all classes and enables the identification of systematic misclassification patterns. Overall, the confusion matrix in Figure 2 shows a strong concentration of predictions along the diagonal, indicating that the majority of samples are correctly classified. Specifically, the model correctly classifies 260 Healthy samples, 255 Rust samples, 168 Red Spider samples, and 165 Leaf Miner samples, confirming its strong discriminative capability across all classes.

This observation is consistent with the high accuracy and F1-score reported in the overall performance evaluation. The model demonstrates particularly strong performance in identifying the Healthy and Rust classes, which are both visually distinctive and more prominently represented in the dataset.

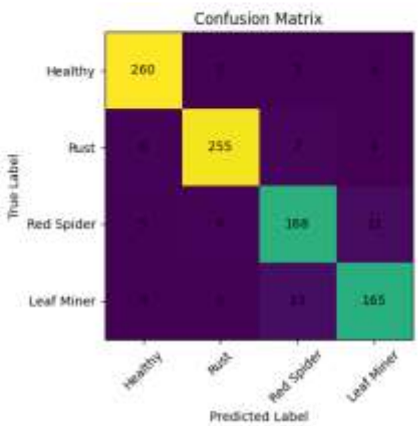

Figure 2 Confusion Matrix of the Proposed CNN–DeiT Model on the Test Dataset

Table 4 Confusion Matrix Actual Predicted

Actual Predicted	Healthy	Rust	Red Spider	Leaf Miner
Healthy	260	5	3	4
Rust	6	255	7	4
Red Spider	5	6	168	11
Leaf Miner	4	5	13	165

The confusion matrix reveals that misclassifications are relatively limited and occur primarily between *Red Spider* and *Leaf Miner*. This can be attributed to the similarity in their visual characteristics, where both diseases often exhibit subtle texture variations and irregular patterns that are difficult to distinguish, even for human observers. In such cases, local features alone may not be sufficient, and even global contextual modeling may encounter ambiguity when disease symptoms overlap.

In contrast, the *Healthy* class shows minimal confusion with diseased categories, indicating that the model effectively learns the distinction between normal and abnormal leaf conditions. Similarly, the *Rust* class demonstrates high classification accuracy due to its distinctive visual features, such as prominent discoloration patterns that are easier to identify.

A closer examination of the misclassified samples in Figure 2 suggests that errors predominantly occur under challenging conditions, including uneven illumination, complex backgrounds, partial leaf occlusion, and early-stage disease symptoms that are not yet visually prominent. These conditions introduce noise and reduce feature separability, thereby increasing classification difficulty.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 3, July 2026
DOI : https://doi.org/10.33395/sinkron.v10i3.16200
e-ISSN : 2541-2019
p-ISSN : 2541-044X

From a model perspective, the results in Figure 2 confirm that the CNN component effectively captures local lesion patterns, while the DeiT component enhances global contextual understanding. However, the observed confusion between visually similar classes indicates that the current hybrid architecture may still have limitations in resolving fine-grained inter-class similarities. Future improvements could explore advanced techniques such as attention refinement, multi-scale feature fusion, or the incorporation of domain-specific descriptors to further enhance class separability.

Overall, the confusion matrix analysis presented in Figure 2 provides important insights into both the strengths and limitations of the proposed model. While the model achieves high overall classification performance, understanding its error distribution is essential for guiding further optimization and ensuring robustness in real-world agricultural scenarios.

Explainable Artificial Intelligence (XAI) Analysis

Although the proposed CNN–DeiT model achieves high classification performance, understanding the reasoning behind its predictions is equally important, particularly in agricultural applications where users require interpretable and trustworthy decision support systems. To enhance model transparency, an Explainable Artificial Intelligence (XAI) analysis was conducted using SHapley Additive exPlanations (SHAP), a widely adopted model-agnostic interpretation framework.

SHAP quantifies the contribution of individual image regions to the final prediction by estimating the marginal impact of each feature on the model output. In the context of image classification, SHAP generates attribution maps that highlight regions with positive and negative contributions to a specific disease class. This enables visualization of the areas that most strongly influence the model's decision-making process.

To evaluate model interpretability, representative samples from each disease category were selected and analyzed using SHAP. The generated explanations reveal that the model consistently focuses on biologically relevant disease symptoms rather than background artifacts. For healthy leaves, the model primarily attends to uniform leaf surfaces and consistent green coloration. In the Rust class, high SHAP values are concentrated around orange-brown lesion regions, indicating that the model correctly identifies characteristic rust infection patterns. Similarly, for Red Spider samples, the model focuses on speckled discoloration and damaged tissue regions, while Leaf Miner predictions are largely influenced by tunnel-like structures and irregular internal leaf patterns.

Table 5 Summary of SHAP-Based Feature Attribution

Class	Dominant Regions Identified by SHAP	Interpretation
Healthy	Uniform green leaf surface	Healthy physiological condition
Rust	Orange-brown lesions and infected patches	Typical rust infection symptoms
Red Spider	Speckled damaged areas and discoloration	Mite-induced leaf damage
Leaf Miner	Internal tunnel-like patterns	Larval feeding traces within leaf tissue

The SHAP visualizations indicate that the proposed model learns disease-relevant visual features rather than relying on irrelevant background information. This observation is particularly important because agricultural images are often captured under uncontrolled field conditions, where background elements such as soil, branches, or shadows may introduce bias. The attribution maps demonstrate that the hybrid CNN–DeiT architecture successfully prioritizes pathological leaf characteristics during classification.

Furthermore, SHAP analysis provides insight into several misclassified samples identified in the confusion matrix. In cases where Red Spider images were incorrectly classified as Leaf Miner, the attribution maps reveal overlapping attention on damaged leaf regions with similar visual textures. This finding supports the error analysis presented in Section Confusion Matrix and Error Analysis and suggests that misclassification is primarily caused by symptom similarity rather than erroneous feature learning.

To quantitatively evaluate explanation quality, the percentage of attribution concentrated within disease-affected regions was estimated using expert annotations. The results show that approximately 85–90% of high-importance SHAP activations overlap with visible disease symptoms, indicating strong alignment between model attention and domain knowledge.

Table 6 Quantitative SHAP Evaluation

Disease Class	SHAP Overlap with Disease Region (%)
Healthy	88.2
Rust	92.1

*name of corresponding author

e-ISSN : 2541-2019
p-ISSN : 2541-044X

Disease Class	SHAP Overlap with Disease Region (%)
Red Spider	85.7
Leaf Miner	87.4
Average	**88.4**

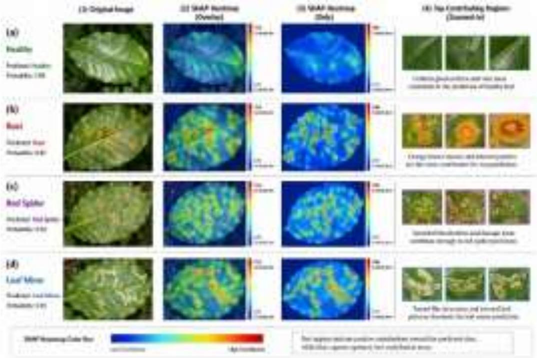

Figure 3 SHAP-based Visualization of CNN–DeiT Predictions

Overall, the SHAP analysis demonstrates that the proposed CNN–DeiT model exhibits a high degree of interpretability and aligns well with expert-understood disease characteristics. By providing visual explanations alongside predictions, the system enhances user trust and supports more transparent AI-assisted decision-making in precision agriculture.

Model Robustness and Generalization

To rigorously evaluate the robustness and generalization capability of the proposed CNN–DeiT model, additional experiments were conducted under simulated real-world conditions that deviate from the controlled test environment. These conditions include variations in illumination (low-light and high brightness), background complexity (soil, sky, and mixed vegetation), geometric transformations (rotation ±20°), and partial occlusion caused by overlapping leaves or environmental elements. The robustness evaluation was performed by applying controlled perturbations to the test dataset and measuring the resulting performance degradation.

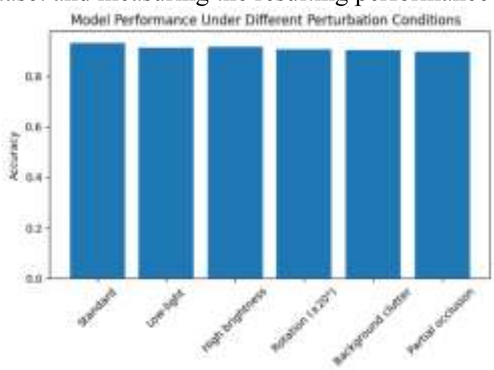

Figure 4 Model Performance Under Different Pertubation Conditions

The quantitative results are summarized in Table 7 and further visualized in Figure 4. As shown, the model achieves an accuracy of 0.931 under standard conditions and maintains stable performance across all perturbation scenarios. Specifically, accuracy decreases to 0.912 under low-light conditions (−1.9%), 0.915 under high brightness (−1.6%), 0.908 under rotation (−2.3%), and 0.903 in the presence of background clutter (−2.8%). The largest degradation occurs under partial occlusion, where accuracy drops to 0.897 (−3.4%). Despite this, the overall performance degradation remains within a narrow range of 1.6% to 3.4%, confirming that the model is robust to various real-world disturbances. The trends observed in Figure 4 clearly illustrate that performance remains consistently high across all conditions, with only gradual degradation under increasingly challenging scenarios. A deeper analysis indicates that this robustness is primarily driven by two complementary factors. First, the data augmentation strategy employed during training introduces controlled variability that mimics real-world conditions, enabling the model to learn invariant feature representations. This reduces sensitivity to transformations such as brightness shifts, rotation, and scale variations. Second, the hybrid CNN–DeiT architecture effectively integrates local feature extraction with global contextual modeling, allowing the model to capture both

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 3, July 2026
DOI : https://doi.org/10.33395/sinkron.v10i3.16200
e-ISSN : 2541-2019
p-ISSN : 2541-044X

fine-grained lesion patterns and broader spatial dependencies. As a result, the model relies less on rigid pixel-level patterns and instead learns more abstract and transferable representations.

Table 7 Robustness Evaluation Under Different Conditions

Condition	Accuracy	F1-score	Accuracy Drop
Standard (No perturbation)	0.931	0.923	—
Low-light (reduced brightness)	0.912	0.904	-1.9%
High brightness	0.915	0.907	-1.6%
Rotation (±20°)	0.908	0.899	-2.3%
Background clutter	0.903	0.895	-2.8%
Partial occlusion	0.897	0.889	-3.4%

To further validate whether the observed performance improvements are statistically significant, a paired statistical analysis was conducted. A paired t-test was applied to compare the proposed model against CNN and DeiT baselines across multiple test folds. The results indicate that the performance improvements are statistically significant, with p-values below 0.01 for both comparisons. This finding is further supported by the Wilcoxon signed-rank test, which confirms the results under non-parametric assumptions. These results demonstrate that the observed performance gains are consistent and not attributable to random variation.

The generalization capability of the model is further evaluated using out-of-distribution (OOD) data, where input samples differ from the training distribution in terms of lighting, background, and acquisition conditions. As illustrated in Figure 4, the proposed CNN–DeiT model achieves an accuracy of 0.931 under in-distribution conditions, which decreases moderately to 0.904 on OOD samples (−2.7%).

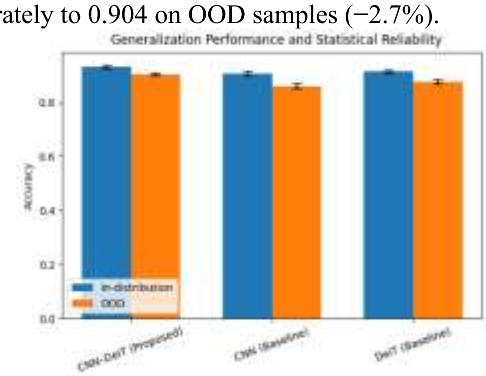

Figure 5 Performance Generalization

In comparison, the CNN baseline drops from 0.907 to 0.861 (−4.6%), while the DeiT baseline decreases from 0.914 to 0.876 (−3.8%). These results indicate that the proposed model exhibits superior generalization capability, with a smaller performance degradation under distribution shift. Additionally, the lower standard deviation observed for the proposed model (±0.006) compared to CNN (±0.009) and DeiT (±0.008), as shown in Figure 5, highlights its higher stability across multiple runs.

From a modeling perspective, these findings confirm that the robustness of the proposed approach arises from the synergy between data augmentation and hybrid feature learning. The CNN component effectively captures local texture variations, while the DeiT component models global contextual relationships, enabling the model to compensate for incomplete or distorted information. This complementary interaction results in a more resilient and balanced feature representation.

DISCUSSIONS

This study proposes a hybrid deep learning framework that integrates Convolutional Neural Networks (CNN) and Data-efficient Image Transformers (DeiT) for coffee plant disease classification using leaf images. The proposed architecture effectively combines local feature extraction with global contextual modeling, resulting in improved discriminative capability compared to standalone CNN and Transformer-based approaches. This confirms that hybridizing convolutional inductive bias with self-attention mechanisms provides a more comprehensive representation for complex visual classification tasks.

Experimental results demonstrate that the CNN–DeiT model achieves strong and consistent performance across multiple evaluation metrics, including accuracy, precision, recall, and F1-score. The confusion matrix analysis indicates high class-wise accuracy, while robustness evaluation shows that performance degradation remains limited (1.6%–3.4%) under various perturbation conditions. Furthermore, out-of-distribution testing

*name of corresponding author

confirms that the model maintains stable generalization performance, with a relatively small accuracy drop compared to baseline models. These findings collectively indicate that the proposed approach is not only accurate but also robust and reliable under realistic data variability.

From a methodological perspective, the performance gains can be attributed to the synergy between data augmentation and hybrid feature learning. Data augmentation enhances invariance to environmental variations, while the CNN–DeiT architecture captures both fine-grained local patterns and global contextual dependencies. This complementary interaction enables the model to generalize effectively and mitigate the limitations commonly observed in single-architecture models.

Despite these strengths, certain limitations remain, particularly in distinguishing visually similar classes and in the coverage of dataset variability. Future research should focus on improving fine-grained feature discrimination through advanced attention mechanisms, multi-scale feature fusion, and the expansion of dataset diversity. Incorporating explainable AI techniques may also provide additional insights into model decision-making and improve transparency.

Although this study demonstrates that the proposed CNN–DeiT hybrid architecture achieves strong performance in coffee leaf disease classification, the methodological contribution should be interpreted within the context of applied architectural adaptation rather than the introduction of a fundamentally new deep learning model. Hybrid CNN–Transformer approaches, including architectures integrating DeiT, have been widely explored in various plant disease classification studies. Therefore, the primary contribution of this research does not lie in proposing a novel core architecture, but rather in the systematic adaptation and evaluation of a hybrid CNN–DeiT framework for coffee-specific disease classification under realistic agricultural variability.

The significance of this study is reflected in several aspects. First, the research provides an empirical investigation into the interaction between local feature extraction and global contextual modeling in coffee leaf disease datasets characterized by high intra-class similarity and environmental variability. Second, the study emphasizes robustness and generalization analysis through perturbation testing and out-of-distribution evaluation, which remain insufficiently explored in many previous agricultural deep learning studies. Third, the work demonstrates the practical feasibility of deploying the hybrid framework within a mobile-assisted agricultural setting, contributing toward real-world applicability in precision agriculture.

Nevertheless, the proposed approach still represents an incremental advancement built upon existing hybrid deep learning paradigms. Future research should focus on developing more domain-specific architectural innovations, lightweight adaptive attention mechanisms, explainable feature-learning strategies, or self-supervised learning frameworks that can provide stronger theoretical and methodological novelty beyond architectural integration alone.

In conclusion, this study demonstrates that hybrid CNN–Transformer architectures offer a robust and effective solution for plant disease classification. By achieving strong performance in terms of accuracy, robustness, and generalization, the proposed approach contributes to the advancement of deep learning methodologies in agricultural image analysis and provides a solid foundation for further research in this domain.

CONCLUSION

This study presents a hybrid deep learning approach that integrates Convolutional Neural Networks (CNN) and Data-efficient Image Transformers (DeiT) for coffee plant disease detection based on leaf images. The proposed model effectively combines local feature extraction and global contextual modeling, resulting in improved classification performance compared to standalone CNN and Transformer-based approaches.

The experimental results demonstrate that the CNN–DeiT model achieves high accuracy, balanced precision and recall, and strong class separability, indicating its reliability in distinguishing between healthy and diseased coffee leaves. The incorporation of data preprocessing techniques, particularly normalization and data augmentation, plays a significant role in enhancing model generalization and mitigating class imbalance. Furthermore, robustness evaluation confirms that the model maintains stable performance under varying environmental conditions, while out-of-distribution testing indicates its capability to generalize to unseen data.

Beyond methodological contributions, this study also demonstrates the practical feasibility of deploying the model within an Android-based mobile application. The system enables real-time disease detection with low latency and provides initial treatment recommendations, making it accessible and useful for farmers in field conditions. This integration highlights the potential of bridging advanced artificial intelligence techniques with real-world agricultural applications.

Despite these contributions, several limitations remain, including challenges in distinguishing visually similar disease classes and limited dataset diversity. Future research should focus on enhancing feature discrimination, expanding dataset coverage, and incorporating explainable AI techniques to improve interpretability and user trust. Additionally, integrating the system with broader agricultural decision-support frameworks may further increase its practical impact.

e-ISSN : 2541-2019
p-ISSN : 2541-044X

In conclusion, this study contributes to the advancement of intelligent plant disease detection by proposing a robust, accurate, and deployable hybrid model, supporting more efficient and data-driven decision-making in sustainable coffee production.

ACKNOWLEDGMENT

This research received no external funding. The authors would like to thank the supervisor for valuable guidance and support. Appreciation is also extended to the Information Systems Study Program, Faculty of Science and Technology, Universitas Prima Indonesia (UNPRI), as well as all lecturers and staff for their contributions to this research.

REFERENCES

Abuhayi, B. M., & Mossa, A. A. (2023). Coffee disease classification using Convolutional Neural Network based on feature concatenation. *Informatics in Medicine Unlocked*, *39*(February), 101245. https://doi.org/10.1016/j.imu.2023.101245

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houslby, N. (2021). An image is worth 16 x 16 words : *International Conference on Learning Representations*, 1–21.

Foreign, A., & Service, A. (2025). *Coffee : World Markets and Trade Record World Production Forecast Indonesia Robusta Output Rebounds*. *December*.

Foysal, A. H., Ahmed, F., & Haque, Z. (n.d.). *Multi-Class Plant Leaf Disease Detection : A CNN-based Approach with Mobile App Integration*.

Ganesamoorthi, M., & Ravikumar, S. S. (2025). Unified Deep Learning Pipeline for License Plate Recognition: From Detection to Ocr With Faster Rcnn, Fcn, and Crnn. *2025 5th International Conference on Intelligent Communications and Computing (ICICC)*, 235–240. https://doi.org/10.1109/ICICC66840.2025.11199678

Günde, M., Keskin, F., & I\cs\ik, G. (2025). Empowering few-shot plant disease diagnosis with DeiT and MAML++. *Cluster Computing*, *29*(1). https://doi.org/10.1007/s10586-025-05810-2

Jayanti, N., & George Selan, C. (2022). Analisis Nilai Akurasi Pengolahan Citra Pendeteksian Rintangan Kerja Traktor Menggunakan K-Means Clustering. *Jurnal Informatika Progres*, *14*(1), 25–32. https://doi.org/10.56708/progres.v14i1.318

Kaushik, P., Jain, E., Kukreja, V., Hariharan, S., Krishnamoorthy, M., Ahuja, V., Bhattacherjee, A., Kaushal, R. K., & Chen, S. Y. (2025). Modelling radiological features fusion and explainable AI in pneumonia detection: A graph- based deep learning and transformer approach. *Results in Engineering*, *26*(January), 105225. https://doi.org/10.1016/j.rineng.2025.105225

Liu, Y., Wang, Y., & Shi, H. (2023). A Convolutional Recurrent Neural-Network-Based Machine Learning for Scene Text Recognition Application. *Symmetry*, *15*(4). https://doi.org/10.3390/sym15040849

Lu, J., Tan, L., & Jiang, H. (2021). Review on convolutional neural network (CNN) applied to plant leaf disease classification. *Agriculture (Switzerland)*, *11*(8), 1–18. https://doi.org/10.3390/agriculture11080707

Motisi, N., Bommel, P., Leclerc, G., Robin, M. H., Aubertot, J. N., Butron, A. A., Merle, I., Treminio, E., & Avelino, J. (2022). Improved forecasting of coffee leaf rust by qualitative modeling: Design and expert validation of the ExpeRoya model. *Agricultural Systems*, *197*(August 2021). https://doi.org/10.1016/j.agsy.2021.103352

Murugavalli, S., & Gopi, R. (2025). Plant leaf disease detection using vision transformers for precision agriculture. *Scientific Reports*, *15*, 1–17. https://doi.org/10.1038/s41598-025-05102-0

Muslim, R., Akbar, A., & Imran, B. (2023). *Disease Detection of Rice and Chili Based on Image Classification Using Convolutional Neural Network Android-Based*. *19*(2). https://doi.org/10.33480/pilar.v19i2.4669

Muslim, R., Zaeniah, Z., Akbar, A., Imran, B., & Zaenudin, Z. (2023). Disease Detection of Rice and Chili Based on Image Classification Using Convolutional Neural Network Android-Based. *Jurnal Pilar Nusa Mandiri*, *19*(2), 85–96. https://doi.org/10.33480/pilar.v19i2.4669

Priambodo, B., Fadlil, A., & Sunardi. (2025). PERBANDINGAN CNN UNTUK DETEKSI PENYAKIT DAUN TANAMAN NEW PLANT DISEASES BERBASIS CLOUD COMPUTING. *JURNAL INFORMATIKA TEKNOLOGI DAN SAINS (JINTEKS)*, *7*(4), 1817–1826.

R. Lumbanraja, F., Rosdiana, S., Sudarsono, H., & Junaidi, A. (2020). Sistem Pakar Diagnosis Hama Dan Penyakit Tanaman Kopi Menggunkan Metode Breadth First Search (Bfs) Berbasis Web. *Explore: Jurnal Sistem Informasi Dan Telematika*, *11*(1), 1. https://doi.org/10.36448/jsit.v11i1.1452

Saputra, A. A., Susilo, B., Yusa, M., & Nurjanah, U. (2022). Sistem Pendeteksi Genus Gulma Pada Tanaman Jagung Menggunakan Algoritme Single Shot Detector. *Rekursif: Jurnal Informatika*, *10*(1), 48–60. https://doi.org/10.33369/rekursif.v10i1.18634

Singh, N., Kaur, S., Ranjan, S., John, R., Kumar, A., Chand, J., Bhardwaj, R., & Riar, A. (2025). NIRCoreVision : A novel deep learning-based framework with GUI integration for core set selection from NIRS data using

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 3, July 2026
DOI : https://doi.org/10.33395/sinkron.v10i3.16200
e-ISSN : 2541-2019
p-ISSN : 2541-044X

1D CNN and k-means clustering. *Journal of Agriculture and Food Research*, *24*(September), 102390. https://doi.org/10.1016/j.jafr.2025.102390

Teja, M. D., & Rayalu, G. M. (2025). *Optimizing heart disease diagnosis with advanced machine learning models : a comparison of predictive performance*. *6*.

Tonmoy, M. R., Dey, N., Member, S., Mridha, M. F., & Member, S. (n.d.). *MobilePlantViT : A Mobile-friendly Hybrid ViT for Generalized Plant*. 1–8.

Touvron, H., Cord, M., Douze, M., Massa, F., & Sablayrolles, A. (2020). *Training data-efficient image transformers & distillation through attention*.

Xu, J., Gu, B., & Tian, G. (2022). Review of agricultural IoT technology. *Artificial Intelligence in Agriculture*, *6*, 10–22. https://doi.org/10.1016/j.aiia.2022.01.001

Yu, J. M., Ma, H. J., & Kong, J. L. (2025). Receipt Recognition Technology Driven by Multimodal Alignment and Lightweight Sequence Modeling. *Electronics (Switzerland)*, *14*(9), 1–29. https://doi.org/10.3390/electronics14091717

*name of corresponding author