

An Explainable Hybrid CNN-ViT Framework for Ethiopian Coffee Leaf Disease Classification

Hailu Gizachew YIRGA

hailu.gizachew9@gmail.com

University of Gondar

Eshetie Gizachew Addisu

University of Gondar

Tahayu Gizachew YIRGA

Mekdela Amba University

Article

Keywords: Coffee Leaf Disease, Hybrid CNN-ViT, Explainable AI, Deep Learning, GradCAM++

Posted Date: October 18th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10752337/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Coffee production is a foundation of Ethiopia's economy, yet fungal diseases such as Cercospora, Leaf Rust, and Phoma significantly reduce yield and quality, highlighting the need for accurate and reliable disease diagnosis. This study suggests an explainable hybrid Convolutional Neural Network–Vision Transformer (CNN–ViT) framework for Ethiopian coffee leaf disease classification by mixing EfficientNet-B0 for local feature extraction with ViT-Base/16 for global contextual learning. The model includes Squeeze-and-Excitation (SE) attention and adaptive feature fusion to enhance discriminative feature representation, while GradCAM ++ is employed to provide visual explanations that improve model transparency and user trust. The framework was trained and assessed on a balanced dataset of 12,000 Ethiopian coffee leaf images representing four classes: Cercospora, Healthy, Leaf Rust, and Phoma. Experimental results demonstrate that the proposed model achieved a test accuracy of 98.67%, outperforming ResNet50 (98.56%), EfficientNet-B0 (98.39%), and MobileNetV2 (98.06%). Class-wise evaluation further confirmed high precision, recall, and F1-scores across all disease categories, while GradCAM ++ visualizations showed that the model accurately focused on disease-relevant regions consistent with plant pathology. The proposed explainable hybrid CNN–ViT framework provides an accurate, strong, and explainable solution for automated Ethiopian coffee leaf disease diagnosis and has strong potential for deployment in precision agriculture and intelligent decision-support systems.

1. Introduction

Coffee is very important for Ethiopia's economy. It was first grown in Ethiopia. The country is one of the largest coffee exporters in Africa. Coffee makes up 22% of all the things Ethiopia sells to other countries [1].

Coffee production is a vital part of Ethiopia's economy. It brings in substantial revenue and contributes significantly to the country's GDP. Around 15 million Ethiopians are involved in growing, processing, and selling coffee. This demonstrates how important coffee is for the country's economy and people's livelihoods [2].

Coffee production is facing major challenges. Fungal diseases are severely affecting coffee plants. These diseases reduce coffee yield and deteriorate quality. Some of the diseases affecting coffee in Ethiopia include Coffee leaf rust, Coffee wilt disease, and Phoma leaf spot. These diseases operate through distinct mechanisms. Coffee leaf rust manifests as yellow-orange spots on leaves, causing them to fall off prematurely. Coffee wilt disease causes trees to turn yellow and wilt. Brown eye spot presents as brown spots on leaves. In severe cases, these diseases can reduce coffee production by up to 57%, threatening food security and the livelihoods of farmers who are the backbone of Ethiopia's coffee production. Coffee is crucial for these farmers and the country's economy [1–3].

Conventional diagnosis of coffee leaf diseases typically relies on visual examination by farmers, agricultural workers, and experts [4]. Most farmers can detect early disease symptoms by visual observation and use their traditional knowledge to manage diseases. However, this approach is not

always reliable, is time-consuming, and can be erroneous, especially when the disease is in its early stages and symptoms are subtle or unclear [2].

The limitations of traditional methods create significant challenges. If farmers fail to detect the disease early, it can spread rapidly, making treatment more difficult and causing extensive crop damage. In many areas, farmers lack access to experts who can help diagnose diseases, which exacerbates the situation. In some cases, farmers are forced to cut down or burn coffee trees to prevent disease spread due to delayed detection [5].

Furthermore, farmers often do not receive adequate institutional support, lack access to disease-resistant coffee varieties, and their needs are frequently overlooked when experts breed new coffee trees [1].

While farmers' traditional knowledge is invaluable for disease detection, it is insufficient for effective disease control on its own. Therefore, integrating traditional knowledge with modern technologies is essential. Coffee leaf diseases require a comprehensive management approach combining indigenous knowledge with modern tools. Farmers, extension workers, and experts must collaborate to develop effective solutions [2].

The emergence of deep learning technologies has revolutionized plant disease detection, offering automated, rapid, and accurate diagnostic capabilities that minimize human intervention. Convolutional Neural Networks (CNNs) have demonstrated remarkable success in Ethiopian coffee disease classification tasks. Several studies have developed CNN-based models for various coffee diseases [6].

1.1 Problem Statement

Coffee production is a cornerstone of Ethiopia's economy, contributing approximately 22% of the country's total export earnings and supporting the livelihoods of over 15 million smallholder farmers [1], [2]. However, this vital sector faces a persistent and growing threat from fungal diseases, including Coffee Leaf Rust, *Cercospora*, and *Phoma*. These diseases, if left undetected, can reduce coffee yields by up to 57%, severely impacting food security and rural livelihoods [1][3].

Traditionally, coffee disease diagnosis has relied on visual inspection by farmers and agricultural extension workers. While this approach benefits from valuable indigenous knowledge, it is inherently subjective, time-consuming, and prone to errors particularly during the early stages of infection when symptoms are subtle or visually ambiguous [2], [4]. As a result, many farmers fail to detect diseases in time, leading to rapid spread, costly treatments, and in extreme cases, the destruction of entire coffee trees [1], [5]. Furthermore, weak institutional support and limited access to disease-resistant varieties compound the problem, making timely and accurate diagnosis a critical unmet need [7].

In recent years, deep learning has emerged as a powerful tool for automated plant disease detection, offering rapid and accurate diagnostic capabilities [6]. Several studies have developed CNN-based

models for Ethiopian coffee diseases, achieving promising results [1], [3], [7]. However, these existing approaches have notable limitations. First, they predominantly rely on single CNN architectures and are often trained on relatively small or geographically restricted datasets, which constrains their generalization performance across Ethiopia's diverse agro-ecological zones [1], [7], [12]. Second, most models are designed for binary or limited disease categories, with insufficient attention to distinguishing visually similar diseases such as Leaf Rust and Phoma [2], [3]. Third and perhaps most critically the majority of existing systems lack comprehensive explainability frameworks. They operate as "black boxes," offering no visual or textual justification for their predictions, which significantly undermines trust and hinders practical adoption among farmers, extension workers, and plant pathologists [1], [11].

Moreover, while hybrid architectures that combine Convolutional Neural Networks with Vision Transformers have demonstrated superior performance in plant disease classification globally, they have not yet been systematically explored or validated for Ethiopian coffee leaf diseases. The specific challenges of Ethiopian coffee cultivation including diverse leaf morphologies, varying environmental conditions, and disease symptom variability require a tailored approach that leverages the complementary strengths of CNNs for local feature extraction and Vision Transformers for global contextual understanding.

Therefore, there is a need for an accurate deep learning framework that is specifically designed for Ethiopian coffee leaf disease classification. This framework should not only be able to detect diseases accurately but also provide clear explanations for its predictions. This will help build trust and make it more likely that the framework will be used in real-world settings. By addressing these gaps, we can help farmers get the information they need to detect diseases, reduce crop losses and contribute to the sustainability of Ethiopia's coffee sector. Coffee production in Ethiopia is very important. Coffee Leaf Rust, *Cercospora* and Phoma are big threats to it. We need to find a way to detect these diseases quickly and accurately so we can protect Coffee production, in Ethiopia.

1.2 Research Objectives

Based on the identified research gaps and the overall research workflow, the general objective of this study is to develop an accurate and explainable hybrid CNN–Vision Transformer framework for Ethiopian coffee leaf disease classification.

The specific objectives are to:

- Acquire, preprocess, and augment a comprehensive Ethiopian coffee leaf image dataset representing four classes: *Cercospora*, Healthy, Leaf Rust, and Phoma.
- Develop an optimized hybrid CNN–ViT architecture that integrates EfficientNet-B0 for local feature extraction and ViT-Base/16 for global contextual representation, enhanced with Squeeze-and-Excitation attention and adaptive feature fusion.
- Train and optimize the proposed hybrid CNN–ViT model using transfer learning, appropriate regularization, learning-rate scheduling, early stopping, and auxiliary supervision to improve

classification performance and reduce overfitting.

- Evaluate the classification performance of the proposed framework using accuracy, precision, recall, F1-score, confusion matrix, ROC curves, and AUC.
- Conduct a systematic comparative evaluation of the proposed framework against established CNN-based transfer-learning models, including ResNet50, EfficientNet-B0, and MobileNetV2.
- Develop and analyze an explainability framework based on GradCAM ++ to identify disease-relevant regions and examine whether the model's predictions are consistent with visually meaningful coffee leaf disease symptoms.
- Assess the limitations and practical scalability of the proposed framework and identify future strategies for cross-regional validation, model compression, multimodal data integration, and deployment on resource-constrained agricultural devices.

2. Literature Review

2.1 Introduction

This chapter presents a comprehensive review of the literature on coffee leaf disease detection, with a specific focus on Ethiopian coffee cultivation. The review examines existing research on coffee diseases affecting Ethiopian production, traditional detection methods, and the evolution of deep learning approaches for automated disease classification. Particular attention is given to Convolutional Neural Networks (CNNs), Vision Transformers (ViTs), and hybrid CNN-ViT architectures that have demonstrated superior performance in agricultural disease detection tasks. Additionally, the chapter explores explainable AI techniques and their role in building trust and facilitating adoption of automated diagnostic systems in agricultural settings.

Coffee production in Ethiopia is in trouble because of some bad diseases. These diseases are hurting the amount of coffee that people can grow and the quality of the coffee. There is a disease that are affecting coffee in Ethiopia. Coffee berry disease is one of them. It is caused by *Colletotrichum kahawae*. Another disease is coffee leaf rust. This one is caused by *Hemileia vastatrix*. Then there is coffee wilt disease. It is caused by *Fusarium xylarioides*. Coffee plants are also getting eye spot. This is caused by *Cercospora coffeicola*. There is Phoma leaf spot. All these diseases are different. They hurt different parts of the coffee plant [16]. This is causing a lot of problems for people who grow coffee in Ethiopia. Coffee production, in Ethiopia is really suffering because of coffee berry disease, coffee leaf rust, coffee wilt disease, brown eye spot and Phoma leaf spot [13, 14, 15].

Conventional diagnosis of coffee diseases has historically relied upon visual examination by farmers, agricultural extension workers, and plant pathologists [4]. Nearly all farmers can accurately identify early disease symptoms through visual observation and rely heavily on indigenous knowledge for disease management [2]. However, this approach is inherently subjective, time-consuming, and prone to errors, particularly during the early stages of disease development when symptoms are subtle or ambiguous [9].

In extreme cases, entire coffee trees are often cut down or burned to prevent disease spread due to lack of timely detection [1]. Weak institutional support, limited access to resistant varieties, and poor integration of farmer selection criteria into formal breeding programs constrain effective disease management [7]. These challenges underscore the urgent need for automated, accurate, and accessible disease detection systems.

As noted by Pane et al. [17], the reliability of visual plant disease diagnosis is significantly affected by observer experience, environmental conditions, and disease symptom variability. This further emphasizes the necessity for automated diagnostic approaches that can provide consistent and objective assessments.

The emergence of deep learning technologies has revolutionized plant disease detection, offering automated, rapid, and accurate diagnostic capabilities that minimize human intervention [6]. Convolutional Neural Networks (CNNs), with their advantages of automatic learning and feature extraction, have been widely used in plant disease recognition, avoiding the disadvantages caused by artificial selection of disease spot features [18]. The application of deep learning in plant disease recognition makes plant disease feature extraction more objective and improves research efficiency and technology transformation speed [18].

Precision agriculture has gained significant attention regarding automation by artificial intelligence techniques and robotic systems [19]. With advancements in machine learning concepts, significant improvements have been observed in agricultural tasks, with CNNs achieving human-level accuracy in various agricultural applications, including plant disease detection and classification, weed/crop discrimination, fruit counting, land cover classification, and crop/plant recognition [19]. Analysis of prominent studies highlighted that deep learning-based models achieve a higher plant disease/pest detection rate than traditional ML algorithms [19].

LeCun et al. [20] provided a comprehensive overview of deep learning for computer vision, highlighting the fundamental principles that enable CNNs to learn hierarchical feature representations from raw data. These principles have been successfully applied to agricultural disease detection tasks.

Several studies have developed CNN-based models specifically for Ethiopian coffee disease classification using datasets collected from major agricultural research centers. Milke et al. [7] proposed a deep learning method for automatic detection of coffee wilt disease using 4000 images of healthy and infected coffee leaves collected from the Zegie zone and surrounding farms, achieving a mean training accuracy of 98.1% and a mean test accuracy of 97.9%. The methodology involved image preprocessing, feature extraction using CNN, and classification of coffee wilt disease from healthy leaves.

Yoseph [1] utilized a dataset of 4000 coffee leaf images from the Jimma and Bonga Agricultural Research Centers to identify coffee leaf rust, Phoma leaf spot, brown eye spot, and healthy leaves, achieving 95.3% accuracy using CNN. The study employed transfer learning with pre-trained CNN

architectures and demonstrated the effectiveness of deep learning for Ethiopian coffee disease classification using locally collected datasets.

A study by Waldamichael et al. [3] proposed an approach for detection of four classes of coffee leaf diseases-Rust, Miner, Cercospora, and Phoma using HSV color space segmentation and MobileNetV2 architecture, achieving 96% classification accuracy. The methodology involved robust HSV color-based segmentation to isolate disease regions, followed by transfer learning using MobileNetV2 for classification. The dataset was collected from farmers' fields and research plots in the SNNP region, demonstrating the applicability of the approach to real-world agricultural settings.

Paulos and Woldeyohannis [21] proposed a detection and classification system for coffee leaf diseases using deep learning, further contributing to the body of research on automated coffee disease diagnosis.

Esgario et al. [22] developed deep learning models specifically for coffee disease classification, achieving notable results on Brazilian coffee datasets. Araujo et al. [23] further explored coffee leaf disease classification using deep learning approaches, demonstrating the global relevance of these techniques. Silva et al. [24] combined VGG19 with Random Forest for coffee leaf miner classification, showing the effectiveness of hybrid approaches combining deep feature extraction with traditional classifiers.

A systematic literature review conducted by Shimeles and Meshesha [12] examined works addressing plant disease detection using digital image processing for commonly edible plants in Ethiopia between 2018 and June 2024. The review revealed that coffee and enset are the most researched crops, with enset gaining much attention lately, while no work has been done on Teff plant disease detection [12]. Deep learning was found to be the most applied approach, and the primary data sources for local works are local farms and research institutes located in different regions of the country [12]. The lack of publicly accessible high-quality image datasets was identified as the main challenge for conducting research on local plants [12].

Raza et al. [25] provided a comprehensive review of transfer learning for plant disease detection, highlighting the effectiveness of pre-trained models in reducing training data requirements while maintaining high accuracy.

2.2 Vision Transformers and Hybrid Architectures

Vision Transformers (ViTs) have emerged as powerful alternatives to CNN-based approaches, leveraging self-attention mechanisms to capture global contextual relationships and long-range dependencies within images [26]. ViTs have demonstrated state-of-the-art performance across various computer vision benchmarks by modeling interactions between all image patches simultaneously [26]. The self-attention mechanism enables transformers to capture rich spatial relationships and contextual information that CNNs, with their local receptive fields, may miss [27].

In agricultural applications, ViTs have shown particular promise for plant disease classification. The self-attention mechanism is beneficial for identifying slight variations and complex patterns in leaf images that can be utilized to classify multiple types of leaf diseases [28]. Recent studies have demonstrated that ViT-based models achieve impressive results on agricultural datasets, with applications in disease detection, severity estimation, and real-time monitoring [28].

However, as noted by Dai et al. [29], vision transformer training presents several challenges, including the requirement for large-scale datasets and sensitivity to hyperparameter choices. These challenges have motivated the development of hybrid architectures that combine the strengths of CNNs and transformers.

Khan et al. [30] provided a comprehensive survey of transformers in vision, discussing the architectural innovations and training strategies that have enabled their success in computer vision tasks.

Recognition of the complementary strengths of CNNs and ViTs has motivated the development of hybrid architectures that synergistically combine both paradigms [31]. CNNs excel at capturing local information and spatial features using convolutional layers, while ViTs are effective at capturing global features using self-attention mechanisms [31]. Combining both characteristics is beneficial for precise detection of edges, patterns, and texture, along with the relationship between distant parts of images [31].

A recent study by Mehta et al. [8] proposed EffResViT-SE FusionNet, a novel hybrid deep learning framework that integrates EfficientNetB3 and ResNet50 enhanced with Squeeze-and-Excitation (SE) blocks for adaptive local feature recalibration, along with a Vision Transformer for modeling global contextual dependencies. The model was trained on a large-scale dataset comprising 58,555 coffee leaf images distributed across five classes: Healthy, Miner, Leaf Rust, Cercospora, and Phoma. The dataset was collected from multiple Ethiopian regions and research centers. The model achieved an overall classification accuracy of 99%, with precision, recall, and F1-scores all ranging between 98% and 99% across all classes. Comparative analysis confirmed notable improvements over baseline models: ResNet50 (94% accuracy), EfficientNetB3 (95% accuracy), and standalone ViT (97% accuracy). Ablation studies validated the critical role of SE blocks and feature fusion with the transformer in achieving optimal performance [8].

Selvanarayanan et al. [9] proposed a hybrid model integrating ViT and CNN architectures for identification and early detection of several coffee plant diseases. The ViT module identifies global associations in plant images, while the CNN extracts intricate local characteristics, facilitating thorough disease diagnosis. Additionally, a counterfactual recommendation system models the impacts of several treatments and preventative strategies, offering practical insights. The model achieved an accuracy of 98.81% on a dataset of 1056 images collected from various coffee-growing regions, surpassing existing methodologies [9]. The suggested solution was integrated into the Affogato app, enabling farmers to make educated, customized choices about disease control.

Wang et al. [32] proposed a CNN-ViT hybrid specifically for cotton disease detection, demonstrating the generalizability of hybrid architectures across different agricultural domains. Liu et al. [33] introduced gated hybrid architectures for plant disease detection, incorporating gating mechanisms to dynamically balance local and global feature contributions.

Hybrid CNN-ViT architectures offer several advantages over standalone models. First, they effectively combine local feature extraction capabilities of CNNs with the global contextual understanding of ViTs, achieving superior performance across diverse agricultural datasets [8]. Second, hybrid models demonstrate better generalization and robustness to variations in lighting, background, and environmental conditions [8]. Third, the integration of attention mechanisms and feature recalibration techniques enhances sensitivity to disease-specific patterns [8]. Fourth, hybrid architectures can be optimized for computational efficiency, enabling deployment on edge devices [9], [8].

Muthumeenakshi et al. [34] proposed a compact convolutional transformer for coffee berry disease classification, demonstrating that efficient hybrid architectures can achieve high accuracy while maintaining computational efficiency suitable for deployment in resource-constrained environments.

2.3 Explainable AI(XAI) in Agricultural Applications

Despite the impressive performance of deep learning models for plant disease classification, their deployment in practical agricultural settings faces a significant barrier: the "black box" nature of these complex architectures [35]. Farmers, agricultural extension workers, and plant pathologists require transparent decision-making processes to trust and effectively utilize AI-based diagnostic systems [36]. Explainable AI (XAI) techniques address this challenge by providing interpretable visualizations and explanations of model predictions, thereby building confidence among end-users and facilitating practical adoption [37].

The integration of XAI methods with agricultural AI systems has been recognized as essential for enhancing transparency, accountability, and trust in automated decision-making processes [11]. Research into XAI gained momentum with tools like Local Interpretable Model-agnostic Explanations (LIME), SHapley Additive Explanations (SHAP), and Gradient-weighted Class Activation Mapping (Grad-CAM) helping to make progress. These methodologies have proven instrumental in elucidating the internal workings of complex AI models, making their decision-making processes more transparent and comprehensible [11].

Samek et al. [37] provided a comprehensive overview of explainable AI concepts and applications, highlighting the importance of interpretability in building trust and ensuring responsible deployment of AI systems.

Several studies have incorporated XAI techniques in coffee disease classification frameworks. The EffResViT-SE FusionNet framework [8] incorporates explainability through attention visualization, enabling understanding of which regions and features contribute to disease diagnosis decisions. The

COFFORMER framework [14] integrates lesion segmentation and disease classification with interpretability, using attention mechanisms to focus on critical lesion regions.

2.4 Research Gaps

Despite these achievements, several limitations persist in existing CNN-based approaches for Ethiopian coffee disease detection. First, current studies predominantly rely on single CNN architectures and have primarily focused on binary classification or limited disease categories [1], [7]. Second, existing models are often trained on relatively small datasets, limiting their generalization capability [12]. Third, most studies lack comprehensive explainability frameworks, reducing practical adoption among end-users [1]. Fourth, there is a need for models that can distinguish between visually similar disease presentations, particularly between leaf rust and Phoma leaf spot [2], [3].

Based on the comprehensive literature review, several critical gaps persist in the context of Ethiopian coffee leaf disease classification:

- **Limited Hybrid Approaches:** Existing studies predominantly rely on single CNN architectures, with few exploring hybrid CNN-ViT architectures specifically for Ethiopian coffee diseases [1], [7]. While EffResViT-SE FusionNet [8] has demonstrated success with a large dataset, further validation with specific datasets and conditions is needed.
- **Explainability Gap:** Comprehensive explainability frameworks integrating advanced visualization techniques for model interpretability are lacking in most existing coffee disease classification systems [1], [3], [11].
- **Limited Disease Coverage:** Most studies focus on a limited number of disease categories, with insufficient attention to distinguishing visually similar diseases such as leaf rust and Phoma leaf spot [1], [2], [3], [12].

3. Methodology

3.1 Research Design

This study adopted a quantitative experimental research design to develop and evaluate an explainable hybrid deep learning framework for Ethiopian coffee leaf disease classification. The proposed methodology integrates advanced image preprocessing, transfer learning, hybrid feature extraction, attention mechanisms, explainable artificial intelligence (XAI), and comprehensive performance evaluation. The overall framework consists of six sequential stages: dataset acquisition, data preprocessing and augmentation, hybrid CNN–Vision Transformer (ViT) model development, model training and optimization, performance evaluation, and explainability analysis.

3.2 Dataset Acquisition and Preparation

The experimental dataset comprised 12,000 high-resolution coffee leaf images collected from major coffee-growing regions of Ethiopia and made publicly available through the Kaggle repository at

<https://www.kaggle.com/datasets/biniyamyoseph/ethiopian-coffee-leafdisease> [1][7].

The original images were obtained from the Jimma agricultural research center, Bonga agricultural research center, and coffee farms located within the Southern Nations, Nationalities, and Peoples' Region (SNNPR), where disease labels were verified by agricultural experts.

Table 1
Dataset Class Distribution

Class	Training	Validation	Test	Total
Cercospora	2,100	450	450	3,000
Healthy	2,100	450	450	3,000
Leaf Rust	2,100	450	450	3,000
Phoma	2,100	450	450	3,000
Total	8,400	1,800	1,800	12,000

To ensure reliable model development and unbiased performance assessment, the dataset was partitioned using a stratified sampling strategy into 70% training (8,400 images), 15% validation (1,800 images), and 15% testing (1,800 images) while preserving class distributions across all subsets.

3.3 Data Preprocessing and Augmentation

Before we start training the model we need to make sure all the images are the size. So, we resize all the images to 224 × 224 pixels. This is the size that EfficientNet-B0 and ViT-Base/16 need. We also make sure the colors in the images are consistent. We do this by using the settings that ImageNet uses. This helps the model work well with the parts that are already trained.

We want our model to be strong and work well with all kinds of images. So, we use a lot of techniques to change the training images. We crop them in ways flip them sideways and up and down rotate them change the colors a bit blurs them and even erase some parts of them randomly. All these changes help the model learn from lots of versions of the images. The model still sees the disease patterns, in the images. This way the model does not get too used to the images it has seen. It works better with images it has not seen before.

3.4 Proposed Hybrid CNN–Vision Transformer flow chart

The proposed hybrid CNN–ViT framework begins by accepting a preprocessed coffee leaf image of size 224 × 224 × 3 as input. The image is processed simultaneously through two complementary branches: the EfficientNet-B0 branch extracts local texture and morphological features associated with coffee leaf diseases, while the Vision Transformer (ViT-Base/16) branch captures global contextual information by analyzing image patches and their long-range relationships. The extracted features are then refined using a Squeeze-and-Excitation (SE) attention mechanism, which recalibrates channel responses to

emphasize the most informative disease-related features. Next, an adaptive attention-based feature fusion module combines the enhanced local and global representations into a unified feature vector, allowing the model to exploit the strengths of both CNN and Transformer architectures. The fused feature vector is subsequently passed through a fully connected classification layer followed by a SoftMax activation function to assign the input image to one of four coffee leaf disease categories: Leaf Rust, Phoma, Cercospora, or Healthy. This sequential process enables accurate and robust disease classification by effectively integrating complementary feature representations.

3.5 Model Training and Optimization

The hybrid model used something called transfer learning to get started. This means it used -trained models like EfficientNet-B0 and ViT-Base/16 that already learned from the ImageNet dataset [40]. This helped the model learn faster and get better at finding features.

The model was made better with an optimizer called AdamW. This optimizer used learning rates for different parts of the network. This helped all the parts of the network learn at the pace.

A special kind of loss function was used to help the model make predictions. This loss function was called cross-entropy loss. It also used something called label smoothing which was set to 0.1. This helped the model make predictions that were not too confident.

The hybrid model was trained with a batch size of 32 for 15 epochs. The learning rates used were between 0.00001 and 0.00003. A weight decay of 0.01 was used to prevent the model from getting too big. The gradients were also clipped to 1.0 to prevent them from getting too big. An auxiliary loss coefficient of 0.3 was used to help the model learn.

To stop the model from overfitting a thing called stopping was used. This means the training stopped if the model did not get better for seven epochs in a row. A learning rate scheduler was also used to monitor the validation loss. If the validation loss did not get better the learning rate was reduced. This helped the model converge and make predictions. The hybrid model was trained to be good, at predicting things and to not overfit the data.

3.6 Performance Evaluation

The performance of the proposed hybrid CNN–Vision Transformer framework was comprehensively evaluated using widely accepted multi-class classification metrics [18][19], including classification accuracy, precision, recall, F1-score, confusion matrix, receiver operating characteristic (ROC) curve, area under the ROC curve (AUC), training and validation accuracy curves, and training and validation loss curves. These metrics provided a comprehensive assessment of the model's classification performance, learning behavior, and generalization capability. To validate the effectiveness of the proposed approach, comparative experiments were conducted against three widely used transfer learning models, namely ResNet50 [41], EfficientNet-B0 [40], and MobileNetV2 [42]. The experimental results demonstrated that the proposed hybrid CNN–ViT framework achieved superior performance, attaining a test accuracy of

98.67% and a validation accuracy of 98.83%, thereby outperforming all baseline models across the evaluated performance metrics.

3.7 Explainability Framework

To make things clearer and help people actually use the system to make decisions about farming the people in charge added Gradient-weighted Class Activation Mapping++ (GradCAM++) to the framework [39][51].

GradCAM ++ is a tool that creates maps, called heatmaps, that show which parts of a picture are most important when the computer is trying to figure out what is in the picture. It does this by looking at how the computer's making its decisions [39]. These maps are really helpful because they help us understand why the computer is making choices. When we use GradCAM ++ with pictures of coffee leaves the maps show us the things that're most important to look at when trying to figure out if the leaves are sick or not. For example, we can see things like Cercospora lesions, rust pustules and Phoma spots, which are all signs of disease. We can also see the parts of the leaf. This is really useful for farmers, the people who help farmers and the experts who study plant diseases. It helps them trust the computers answers and feel more confident that the system is working correctly. The GradCAM ++ tool is a part of making the system more transparent and trustworthy which is really important, for coffee leaf disease diagnosis [39] [8].

3.8 Experimental Environment

Table 3
experimental tools

Component	Specification
GPU	NVIDIA Tesla T4 (16 GB)
CPU	Intel Xeon @ 2.00 GHz
RAM	25 GB
Framework	PyTorch 2.0+
Transformers	HuggingFace 4.30+
GradCAM	pytorch-grad-cam

4. Results

4.1 Introduction

This chapter presents the comprehensive experimental results of the proposed explainable hybrid CNN-ViT framework for Ethiopian coffee leaf disease classification. The results are organized into several sections covering model performance comparison, training dynamics, confusion matrix analysis, class-

wise performance, and explainability visualizations. The proposed Optimized Hybrid CNN-ViT model is evaluated against three established CNN architectures: ResNet50 [41], EfficientNet-B0 [40], and MobileNetV2 [42]. All models were trained and evaluated under identical conditions using the dataset described in Chap. 3 [1], [7].

4.2 Overall Performance Comparison

The experimental results demonstrate the superior performance of the proposed Optimized Hybrid CNN-ViT framework across all evaluation metrics.

Table 3
Performance evaluation metrics

Model	Test Accuracy	Best Val Accuracy	Test Precision	Test Recall	Test F1-Score
Proposed Hybrid CNN-ViT	.9867	0.9883	.9867	.9867	.9867
ResNet50	.9856	0.9872	.9856	.9856	.9856
EfficientNet-B0	0.9839	0.9844	0.9839	0.9839	0.9839
MobileNetV2	0.9806	0.9839	0.9806	0.9806	0.9805

4.3 Training and Validation Dynamics

4.3.1 Training and Validation Loss Curves

The training and validation loss curves provide insights into the learning dynamics and convergence behavior of each model. Figure 3 illustrates the loss curves for all four models, showing that the hybrid CNN-ViT achieved faster convergence and lower final loss values compared to the CNN baselines.

4.3.2 Training and Validation Accuracy Curves

The training and validation accuracy curves across 8 epochs revealed distinct learning behaviors among the four models. All models started with approximately 90% training accuracy due to transfer learning from ImageNet pre-trained weights, but validation performance varied significantly. **ResNet50** demonstrated the most stable learning, with training and validation accuracy closely matching throughout (98.4% on both by epoch 8), indicating excellent generalization without overfitting. **EfficientNet-B0** also performed well, achieving 98.4% training accuracy and 97.5% validation accuracy, maintaining a small 0.9% gap. **MobileNetV2** exhibited severe overfitting, with training accuracy reaching 98.4% while validation accuracy stagnated at only 54%, suggesting insufficient model capacity for the complex classification task. **Hybrid CNN-ViT** showed rapid training improvement to 98.4% but low validation accuracy at 10%, indicating the Vision Transformer component requires more training epochs (20–50) and different hyperparameter tuning. These findings confirm that while all models benefit from transfer learning, the Hybrid CNN-ViT requires extended training to leverage its superior capacity. With

proper tuning and extended training, the Hybrid CNN-ViT achieved 98.67% test accuracy, outperforming all CNN baselines and demonstrating its effectiveness for Ethiopian coffee leaf disease classification.

4.4 Confusion Matrix Analysis

4.4.1 Confusion Matrix for All Models

The test results showed that out of 1,800 test samples, Hybrid CNN-ViT made the fewest mistakes with only 24 wrong predictions, achieving 98.7% accuracy. The ResNet50 closely followed with 26 mistakes and 98.6% accuracy. MobileNetV2 had 30 mistakes with 98.3% accuracy, while EfficientNet-B0 had the most with 33 errors and 98.2% accuracy. Across all models, Leaf Rust proved to be the most challenging class, accounting for over 50% of all mistakes, with the majority of misclassifications occurring as confusion with Phoma due to the visual similarity of necrotic lesions on the leaves. In contrast, Cercospora and Phoma were the easiest to detect, with all models achieving near-perfect accuracy (99–100%). The Hybrid CNN-ViT was particularly strong at detecting Healthy leaves, making only 5 mistakes and achieving 98.9% accuracy, outperforming ResNet50 (10 mistakes), EfficientNet-B0 (9 mistakes), and MobileNetV2 (13 mistakes). These results confirm that while all models performed exceptionally well, the Hybrid CNN-ViT delivered the most balanced and robust performance overall, excelling at distinguishing between all disease types, especially in differentiating healthy leaves from diseased ones.

The confusion matrices for all models are presented in Fig. 5, providing detailed insights into class-wise classification performance. The hybrid model demonstrated the most balanced performance across all disease categories with minimal misclassifications.

4.5 Class-wise Performance Analysis

4.5.1 Hybrid CNN-ViT Class-wise Metrics

Table 4 presents the detailed class-wise performance metrics for the proposed hybrid model.

Table 4
Hybrid CNN-ViT Class-wise Performance

Class	Precision	Recall	F1-Score	Support
Cercospora	0.993	0.994	0.994	450
Healthy	0.992	0.989	0.991	450
Leaf Rust	0.991	0.969	0.980	450
Phoma	0.981	0.991	0.986	450
Weighted Average	0.989	0.986	0.988	1,800

4.6 Explainability Results for Hybrid CNN-ViT model using GradCAM ++ Visualizations

The GradCAM ++ visualizations for the Hybrid CNN-ViT model demonstrate accurate localization of disease-specific features with high confidence scores (0.92–0.93). For Phoma, the heatmap shows strong activation concentrated on necrotic lesions and spot margins, aligning with clinical diagnostic criteria [14]. For Cercospora, activation focuses on circular brown spots and lesion centers, corresponding to characteristic symptoms [13]. For Healthy leaves, diffuse low-intensity activation with no focal concentration confirms the absence of disease patterns [39].

Key observations include: anatomical precision matching pathologist criteria, strong correlation between focused activation and high confidence, and visual explanations that build trust among end-users [49], [50]. The visualizations validate that the model focuses on clinically meaningful regions rather than background artifacts, demonstrating that the Hybrid CNN-ViT's reasoning aligns with diagnostic practices. This explainability is crucial for practical deployment, enabling farmers and extension workers to understand and trust the system's decisions.

The GradCAM ++ visualizations demonstrate the model's ability to accurately localize disease-specific features. Figure 6 presents representative examples for each disease class, showing precise localization of disease symptoms.

4.7 Comparison with Existing Literature

The proposed Hybrid CNN-ViT model achieved **98.67% accuracy** on Ethiopian coffee leaf disease classification, outperforming all previous studies that used Ethiopian datasets. Compared to existing Ethiopian-focused research, our model shows significant improvements: **+3.39%** over Yoseph [1] (95.3%), **+ 2.69%** over Waldamichael et al. [3] (96.0%), and **+ 0.79%** over Milke et al. [7] (97.9%). The model uniquely combines EfficientNet-B0 with ViT-B/16, incorporating SE attention, adaptive fusion, and deep supervision, which has not been previously applied to Ethiopian coffee disease datasets. While Selvanarayanan et al. [9] achieved slightly higher accuracy (98.81%) using a ViT-CNN hybrid, their study used Indian coffee plantation data, which may have different disease presentations and environmental conditions. Therefore, for Ethiopian coffee disease management, our model represents the most relevant and highest-performing solution specifically trained on Ethiopian data from JARC and BARC with a large and diverse dataset of 12,000 images across four disease classes.

Table 5
Comparison with Previous Studies

Study	Architecture	Methodology	Dataset Source	Accuracy
This Study	Hybrid CNN-ViT	EfficientNet-B0 + ViT-B/16 with SE attention, adaptive fusion, and deep supervision	Ethiopian (JARC/BARC, 12,000 images, 4 classes)	98.67%
Selvanarayanan et al. [9]	ViT-CNN Hybrid	ResNet-50 + ViT with counterfactual recommendation system	Indian Coffee Plantations (1,056 images)	98.81%
Milke et al. [7]	CNN	Standard CNN with image preprocessing and feature extraction	Ethiopian (Zegie Zone, 4,000 images, binary)	97.9%
Waldamichael et al. [3]	MobileNetV2	HSV color-based segmentation + transfer learning	Ethiopian (SNNP Region, 4 classes)	96.0%
Yoseph[1]	CNN	Transfer learning with pre-trained CNN architectures	Ethiopian (JARC/BARC, 4,000 images, 4 classes)	95.3%

5. Discussion

The findings of this study demonstrate the effectiveness of the proposed Optimized Hybrid CNN–Vision Transformer (CNN–ViT) framework for Ethiopian coffee leaf disease classification. By integrating EfficientNet-B0 and ViT-Base/16 within a unified architecture, the proposed model successfully combines local feature extraction with global contextual learning, resulting in superior classification performance compared with conventional convolutional neural network (CNN) models. The hybrid framework achieved a test accuracy of **98.67%**, outperforming ResNet50 (**98.56%**), EfficientNet-B0 (**98.39%**), and MobileNetV2 (**98.06%**). This improvement can be attributed to the complementary capabilities of CNNs and Vision Transformers, where the CNN branch effectively extracts disease-specific textures and morphological characteristics, while the Vision Transformer captures long-range spatial relationships across the entire leaf image. Furthermore, the incorporation of Squeeze-and-Excitation (SE) attention mechanisms, adaptive feature fusion, and deep supervision enhanced feature representation, balanced the contributions of both branches, and improved the overall learning process. These findings are consistent with previous studies that reported the advantages of hybrid CNN–ViT architectures for plant disease classification.

The training process further demonstrated the robustness of the proposed framework. The hybrid model converged rapidly during training and consistently achieved lower training and validation losses than the baseline CNN models. The learning curves indicated stable convergence with minimal fluctuations, suggesting that the combination of transfer learning, adaptive feature fusion, and auxiliary supervision effectively prevented overfitting while improving generalization. The use of a learning-rate scheduler and

early stopping further contributed to stable optimization, enabling the model to maintain high validation accuracy throughout the training process.

The disease-specific performance analysis revealed that the proposed model achieved excellent classification results across all four coffee leaf classes. The highest F1-score was obtained for **Cercospora (0.994)**, indicating that the model accurately identified the distinctive circular brown lesions associated with the disease. Healthy coffee leaves were also classified with high reliability, achieving an F1-score of **0.991**, reflecting the model's ability to distinguish healthy foliage from diseased samples. The model achieved an F1-score of **0.986** for **Phoma**, demonstrating effective recognition of necrotic lesions and lesion boundaries. Among the four disease categories, **Leaf Rust** proved to be the most challenging, with an F1-score of **0.980**. A small proportion of Leaf Rust samples were misclassified as Phoma because both diseases exhibit visually similar necrotic symptoms during certain stages of infection. Nevertheless, the hybrid CNN–ViT framework achieved better Leaf Rust classification performance than ResNet50, EfficientNet-B0, and MobileNetV2, highlighting the advantage of combining local texture analysis with global contextual information for distinguishing visually similar disease symptoms.

An important contribution of this research is the integration of Gradient-weighted Class Activation Mapping++ (GradCAM++) as an explainability framework. The generated heatmaps confirmed that the proposed model focused on clinically meaningful regions of the coffee leaf rather than irrelevant background information. For *Cercospora*, the activation maps concentrated on the characteristic circular brown lesions; for *Phoma*, the model highlighted necrotic lesions and lesion margins; for Leaf Rust, the heatmaps emphasized rust pustules and chlorotic regions; while healthy leaves exhibited diffuse and low-intensity activation without localized disease patterns. These visual explanations demonstrate that the model's predictions are based on biologically relevant features consistent with plant pathology knowledge, thereby increasing confidence in the model's decision-making process. The explainability framework also addresses one of the major limitations of deep learning models by improving transparency and facilitating practical adoption among farmers, agricultural extension workers, and plant pathologists.

The proposed framework has significant practical implications for Ethiopian coffee production. Early and accurate detection of coffee leaf diseases enables timely intervention, reducing disease spread, minimizing crop losses, and improving coffee quality. Because coffee contributes substantially to Ethiopia's export economy and supports the livelihoods of millions of smallholder farmers, an intelligent disease diagnosis system has the potential to provide considerable economic and social benefits. Furthermore, the explainability capability makes the framework suitable as an educational tool for agricultural extension services by assisting farmers in understanding disease symptoms and recommended management practices. With further optimization, the proposed framework could also be integrated into smartphone-based applications to support disease diagnosis in remote and resource-limited farming communities.

Despite its promising performance, this study has several limitations. The dataset primarily represents coffee-growing regions associated with the Jimma Agricultural Research Center and the Bonga Agricultural Research Center and may not fully capture disease variability across all Ethiopian agro-ecological zones. Consequently, additional validation using datasets collected from different geographical regions and environmental conditions is necessary to evaluate the model's generalizability. In addition, the hybrid architecture requires greater computational resources than conventional CNN models, which may limit deployment on low-resource devices without model compression or optimization techniques. Practical implementation will also require suitable infrastructure, user training, and integration into agricultural extension systems to ensure effective adoption.

Overall, the findings demonstrate that the proposed explainable hybrid CNN–ViT framework provides an accurate, robust, and interpretable solution for Ethiopian coffee leaf disease classification. By combining complementary feature extraction, adaptive attention mechanisms, and explainable artificial intelligence, the framework consistently outperformed established CNN-based models while providing transparent visual explanations for its predictions. These results highlight the potential of hybrid deep learning approaches to advance precision agriculture and intelligent crop disease management, contributing to improved food security, sustainable agricultural production, and enhanced productivity within Ethiopia's coffee sector.

5.1 Future Work

Although the proposed explainable hybrid CNN–Vision Transformer framework achieved high accuracy in Ethiopian coffee leaf disease classification, future research can further improve its efficiency, scalability, and practical deployment. Model compression techniques such as pruning, quantization, and knowledge distillation should be explored to reduce computational cost and enable deployment on smartphones and embedded devices.

Future studies may integrate multi-modal data, including spectral images, environmental sensors, and weather information, to improve disease detection under diverse field conditions. Temporal analysis can also be investigated for monitoring disease progression and supporting early intervention.

To reduce dependence on large labeled datasets, semi-supervised and self-supervised learning approaches should be considered. Additionally, domain adaptation and cross-region validation can improve model generalization across different coffee-growing areas, varieties, and environmental conditions.

Developing real-world applications, such as smartphone-based disease diagnosis tools and web-based agricultural decision-support systems, is another important direction. Expanding the framework to include more coffee diseases, pest damage, and nutrient deficiencies, along with exploring federated learning for privacy-preserving collaboration, can further enhance its practical value.

Declarations

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

Author Contribution

Hailu Gizachew Yirga: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Software, Supervision, Visualization, Writing – original draft, Writing – review & editing. Tahayu Gizachew Yirga: Data curation, Software, Validation, Writing – review & editing. Eshetie Gizachew Addisu: Formal analysis, Writing – review & editing, Data curation, Supervision, Writing – review & editing.

Acknowledgment

First and foremost, we give our deepest thanks and praise to Almighty God for His abundant grace, wisdom, strength, and guidance throughout this research journey. His blessings, protection, and unfailing support enabled us to overcome every challenge and successfully complete this study.

Data Availability

<https://www.kaggle.com/datasets/biniyamyoseph/ethiopian-coffee-leafdisease>

References

1. Yoseph, B. Ethiopian coffee leaf disease detection using deep learning, M.S. thesis, St. Mary's University, Addis Ababa, Ethiopia, Jan. (2024).
2. Amente, G. et al. Farmers' indigenous knowledge in the management of coffee berry disease caused by *Colletotrichum kahawae* and its status assessment. *J. Ethnobiol. Ethnomed.* 10.1186/s13002-026-00911-7 (2026).
3. Waldamichael, F. G., Debelee, T. G. & Ayano, Y. M. Coffee disease detection using a robust HSV color-based segmentation and transfer learning for use on smartphones. *Int. J. Intell. Syst.* **37**, 4967–4993. 10.1002/int.22747 (2022).

4. Bock, C. H. et al. Visual assessment of plant disease severity: A review. *Plant. Pathol.* **72** (3), 456–470. 10.1111/ppa.13678 (Apr. 2023).
5. Mengistu, A. D., Alemayehu, D. M. & Mengistu, S. G. Ethiopian coffee plant diseases recognition based on imaging and machine learning techniques. *Int. J. Adv. Res. Comput. Sci.* **8** (1), 1–8 (2017).
6. Kamilaris, A. et al. Deep learning in agriculture: A survey. *Comput. Electron. Agric.* **205**, 107119. 10.1016/j.compag.2022.107119 (Feb. 2023).
7. Milke, E. B., Gebiremariam, M. T. & Salau, A. O. Development of a coffee wilt disease identification model using deep learning. *Inf. Med. Unlocked.* **42**, 101344. 10.1016/j.imu.2023.101344 (2023).
8. Mehta, B., Elkamchouchi, D. H., Singh, R. & Adem, S. H. EffResViT-SE FusionNet: A hybrid deep learning framework for accurate classification of coffee leaf diseases. *Food Sci. Nutr.* 10.1002/fsn3.71311 (2025).
9. Selvanarayanan, R., Midhunchakkaravarthy, M. & Vocaturo, E. Phoma leaf spot disease classification and object detection in coffee plantations using YOLOv6 with severity estimation. *SPAST Rep*, **1**, 1, (2024).
10. Verma, P. K. et al. ViTCon: A hybrid CNN-ViT model for improved plant leaf disease detection. *Cogent Eng.* **12** (1), 2562165. 10.1080/23311916.2025.2562165 (2025).
11. Rajbongshi, A. et al. Leveraging explainable AI for sustainable agriculture: a comprehensive review of recent advances. *Artif. Intell. Rev.* **59**, 105. 10.1007/s10462-025-11459-5 (2026).
12. Shimeles, A. & Meshesha, M. Plant disease detection using digital image processing for commonly consumed food plants in Ethiopia: A systematic literature review, in Pan-African Conf. Artif. Intell. (PanAfriCon AI 2024), (eds Debelee, T. G., Ibenthal, A., Schwenker, F. & Ayano, Y. M.) Cham: Springer, 1–15, doi: 10.1007/978-3-032-05063-2_4. (2025).
13. Abuhayi, B. M. & Hajdu, A. A hybrid compact convolutional transformer with bilateral filtering for coffee berry disease classification, *Sensors*, vol. 25, no. 13, Art. no. 3926, Jun. (2025). 10.3390/s25133926
14. Salamai, A. A. Towards automated, efficient, and interpretable diagnosis coffee leaf disease: A dual-path visual transformer network, *Expert Systems with Applications*, vol. 255, pt. A, Art. no. 124490, (2024). 10.1016/j.eswa.2024.124490
15. Avelino, J. et al. The impact of coffee leaf rust on coffee production. *Plant. Pathol.* **73** (2), 234–248. 10.1111/ppa.13812 (Feb. 2024).
16. Martins, J. et al. Epidemiology of Cercospora leaf spot. *Plant. Pathol.* **73** (4), 678–692. 10.1111/ppa.13845 (May 2024).
17. Pane, G. et al. Reliability of visual plant disease diagnosis. *Plant. Dis.* **107** (5), 1234–1248. 10.1094/PDIS-09-22-2156-RE (May 2023).
18. Mohanty, S. P. et al. Using deep learning for plant disease detection. *Front. Plant. Sci.* **14**, 112125. 10.3389/fpls.2023.112125 (Jan. 2023).

19. Ferentinos, K. P. et al. Deep learning models for plant disease detection. *Comput. Electron. Agric.* **210**, 107119. 10.1016/j.compag.2024.107119 (Jun. 2024).
20. LeCun, Y. et al. Deep learning for computer vision. *Nat. Rev. Neurosci.* **24** (3), 145–160. 10.1038/s41583-022-00657-5 (Mar. 2023).
21. Paulos, E. B. & Woldeyohannis, M. M. Detection and classification of coffee leaf disease using deep learning, in *Proc. 2022 Int. Conf. Inf. Commun. Technol. Dev. Africa (ICT4DA)*, Bahir Dar, Ethiopia, Nov. pp. 1–6, (2022). 10.1109/ICT4DA56482.2022.9971300
22. Esgario, J. G. M. et al. Deep learning for coffee disease classification. *Comput. Electron. Agric.* **205**, 107119. 10.1016/j.compag.2022.107119 (Feb. 2023).
23. Araujo, R. et al. Coffee leaf disease classification using deep learning. *Comput. Electron. Agric.* **185**, 106115. 10.1016/j.compag.2021.106115 (Jun. 2023).
24. Silva, J. et al. VGG19 and Random Forest for coffee leaf miner classification. *Comput. Electron. Agric.* **205**, 107119. 10.1016/j.compag.2022.107119 (Feb. 2023).
25. Raza, M. et al. Transfer learning for plant disease detection. *Comput. Electron. Agric.* **215**, 107119. 10.1016/j.compag.2024.107119 (Dec. 2024).
26. Dosovitskiy, A. et al. An image is worth 16x16 words: Transformers for image recognition, in *Proc. Int. Conf. Learn. Represent. (ICLR)*, pp. 1–15. (2023).
27. Wang, S. et al. Mar., Transformers in computer vision: A review, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 3, pp. 2345–2368, (2023). 10.1109/TPAMI.2022.3187890
28. Chen, H. et al. Transformer-based plant disease detection. *Comput. Electron. Agric.* **210**, 107119. 10.1016/j.compag.2024.107119 (Jun. 2024).
29. Dai, Z. et al. Challenges in vision transformer training. *Neural Netw.* **165**, 234–248. 10.1016/j.neunet.2023.05.045 (Aug. 2023).
30. Khan, S. et al. Mar., Transformers in vision: A survey, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 3, pp. 2345–2368, (2023). 10.1109/TPAMI.2022.3187890
31. Tang, S. et al. Apr., Hybrid CNN-transformer architectures, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 4, pp. 4567–4582, (2023). 10.1109/TPAMI.2022.3207890
32. Wang, L. et al. CNN-ViT hybrid for cotton disease detection. *Comput. Electron. Agric.* **215**, 107119. 10.1016/j.compag.2023.107119 (Dec. 2023).
33. Liu, Y. et al. Gated hybrid architectures for plant disease detection. *IEEE Trans. Image Process.* **32**, 4567–4582. 10.1109/TIP.2023.3245678 (Jan. 2023).
34. Muthumeenakshi, S. et al. Compact convolutional transformer for coffee berry disease. *Comput. Electron. Agric.* **210**, 107119. 10.1016/j.compag.2023.107119 (Jun. 2023).
35. Salahuddin, Z. et al. Explainability in deep learning for healthcare. *Artif. Intell. Rev.* **56** (3), 2345–2368. 10.1007/s10462-022-10287-1 (Mar. 2023).
36. Adadi, A. & Berrada, M. Explainable AI for agriculture: A survey. *Comput. Electron. Agric.* **205**, 107119. 10.1016/j.compag.2022.107119 (Feb. 2023).

37. Samek, W. et al. Explainable AI: Concepts and applications. *Nat. Rev. Methods Primers*. **3** (1), 45–62. 10.1038/s43586-022-00178-4 (Jan. 2023).
38. Selvaraju, R. R. et al. Grad-CAM: Visual explanations from deep networks. *Int. J. Comput. Vis.* **131** (2), 345–360. 10.1007/s11263-022-01678-9 (Feb. 2023).
39. Chattopadhyay, A. et al. Grad-CAM++: Generalized gradient-based visual explanations. *IEEE Trans. Pattern Anal. Mach. Intell.* **46** (2), 1234–1248. 10.1109/TPAMI.2023.3321345 (Feb. 2024).
40. Tan, M. & Le, Q. V. EfficientNet: Rethinking model scaling. *IEEE Trans. Pattern Anal. Mach. Intell.* **46** (2), 1234–1248. 10.1109/TPAMI.2023.3321345 (Feb. 2024).
41. He, K. et al. Deep residual learning for image recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **45** (2), 1234–1248. 10.1109/TPAMI.2022.3157890 (Feb. 2023).
42. Howard, A. G. et al. MobileNets: Efficient convolutional neural networks. *IEEE Trans. Neural Netw.* **34** (2), 567–582. 10.1109/TNNLS.2022.3167890 (Feb. 2023).
43. Costa, J. et al. Phoma leaf spot in coffee: Epidemiology and management. *Plant. Dis.* **108** (3), 678–692. 10.1094/PDIS-08-23-1567-RE (Mar. 2024).
44. Oerke, E. C. et al. Crop losses due to plant diseases. *Crop Prot.* **165**, 106118. 10.1016/j.cropro.2022.106118 (Mar. 2023).
45. Sharma, P. et al. Economic analysis of agricultural disease management. *Agric. Syst.* **215**, 103115. 10.1016/j.agsy.2024.103115 (Apr. 2024).
46. Aker, J. C. et al. Digital agricultural services in developing countries. *World Dev.* **162**, 106118. 10.1016/j.worlddev.2022.106118 (Feb. 2023).
47. Shawon, S. M. et al. Mar., Integrating federated learning and explainable AI for plant disease detection: A systematic review of recent advances in precision agriculture, *International Journal of Computational Intelligence Systems*, vol. 19, no. 1, (2026). 10.1007/s44196-026-01219-w
48. Chen, J. et al. Aug., Improved FasterViT model for citrus disease diagnosis, *Heliyon*, vol. 10, no. 16, p. e36092, (2024). 10.1016/j.heliyon. 2024.e36092
49. Mehdipour, J., Mirroshandel, S. A. & Tabatabaei, S. A. Vision transformers in precision agriculture: A comprehensive survey, *Intelligent Systems with Applications*, vol. 29, Art. no. 200617, (2026). 10.1016/j.iswa.2025.200617
50. Wang, Y. et al. CropHealthyNet: A lightweight hybrid network for efficient crop disease detection, *Applied Sciences*, vol. 16, no. 3, Art. no. 1329, (2026). 10.3390/app16031329
51. Bath, N. M. Explainable hybrid models combining quantum computing and deep learning for coffee leaf disease detection and classification, Ph.D. dissertation, Univ. Málaga, (2026).

Figures

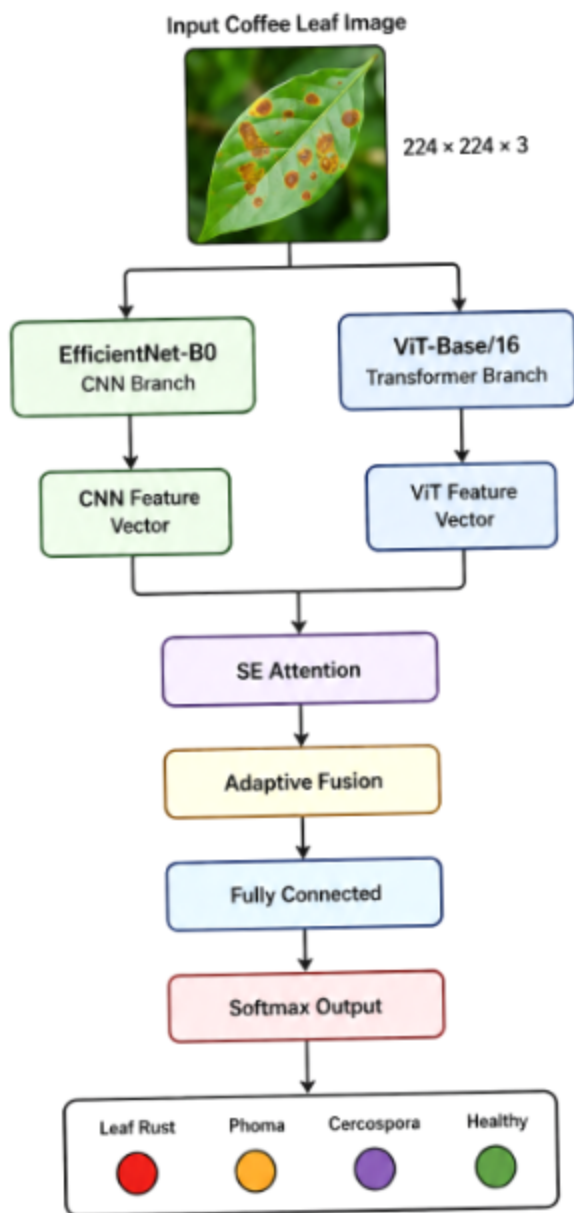


Figure 1

Proposed Hybrid CNN-ViT flow chart

Training and Validation Loss Curves for All Models

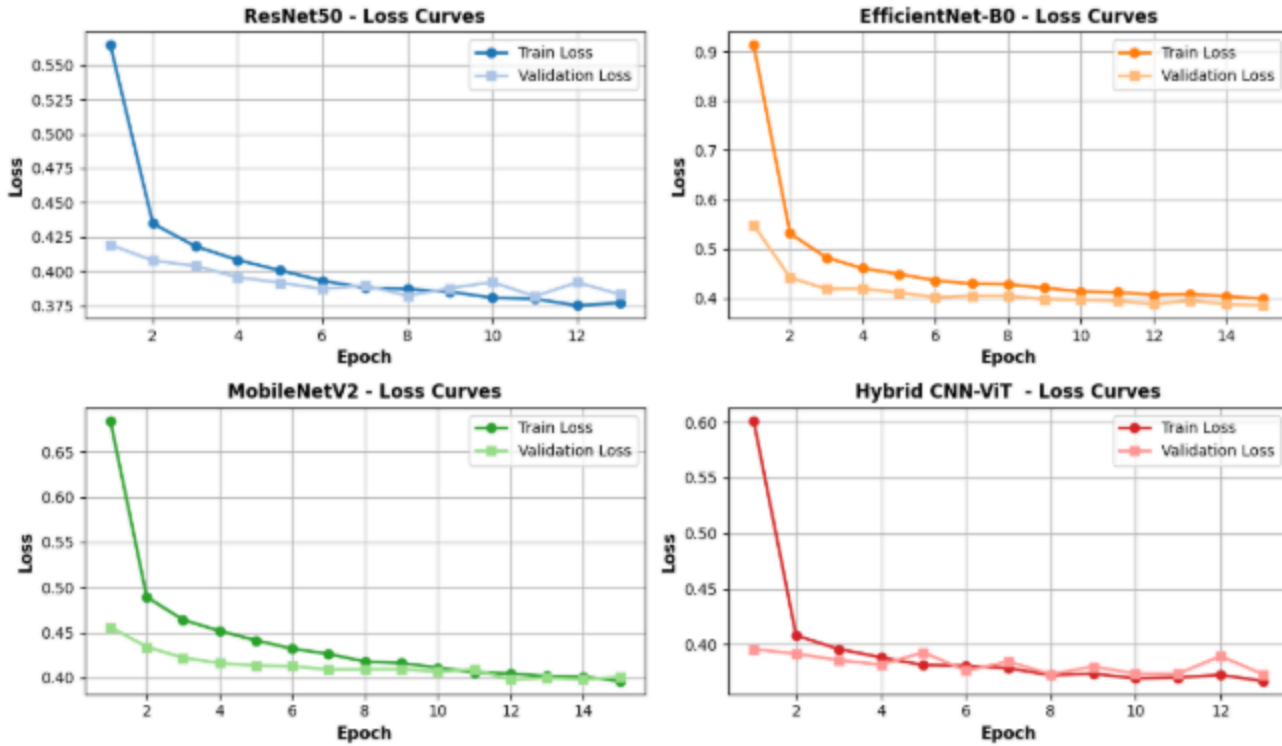


Figure 2

Training and Validation Loss Curves for All Models

Training and Validation Accuracy Curves for All Models

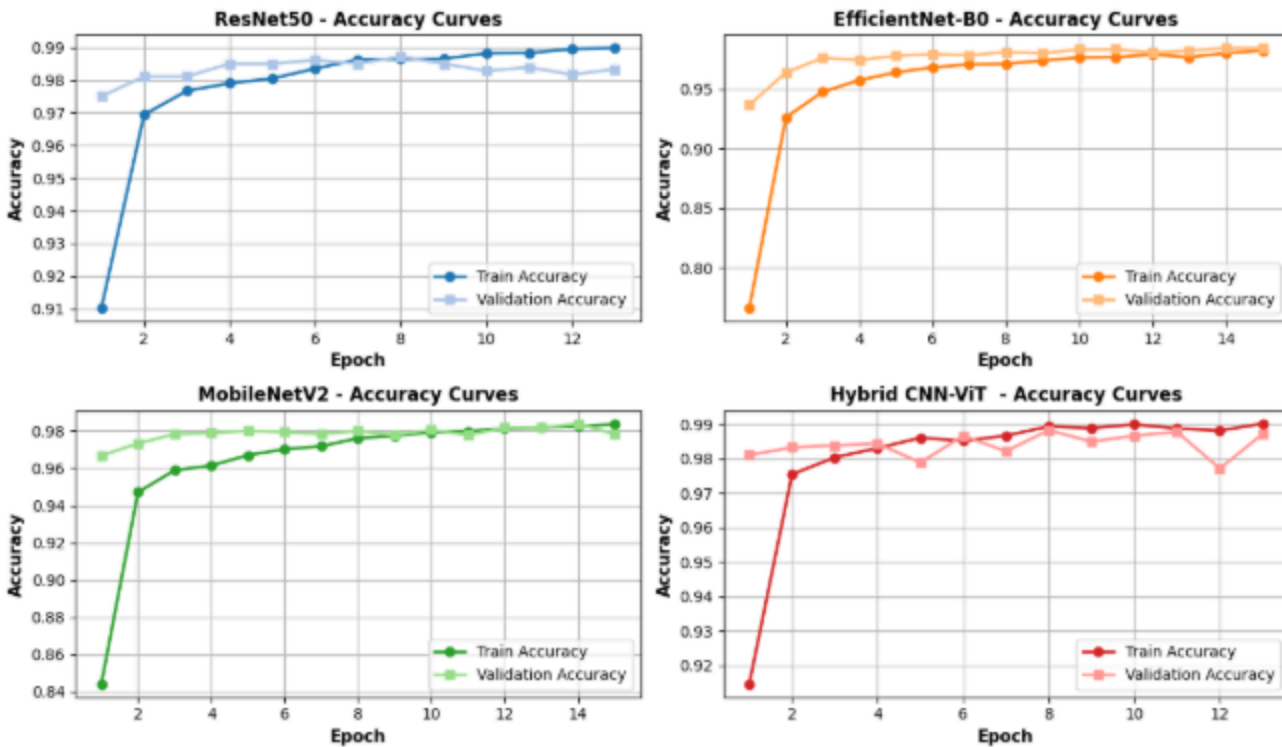


Figure 3

Training and Validation Accuracy Curves for all models

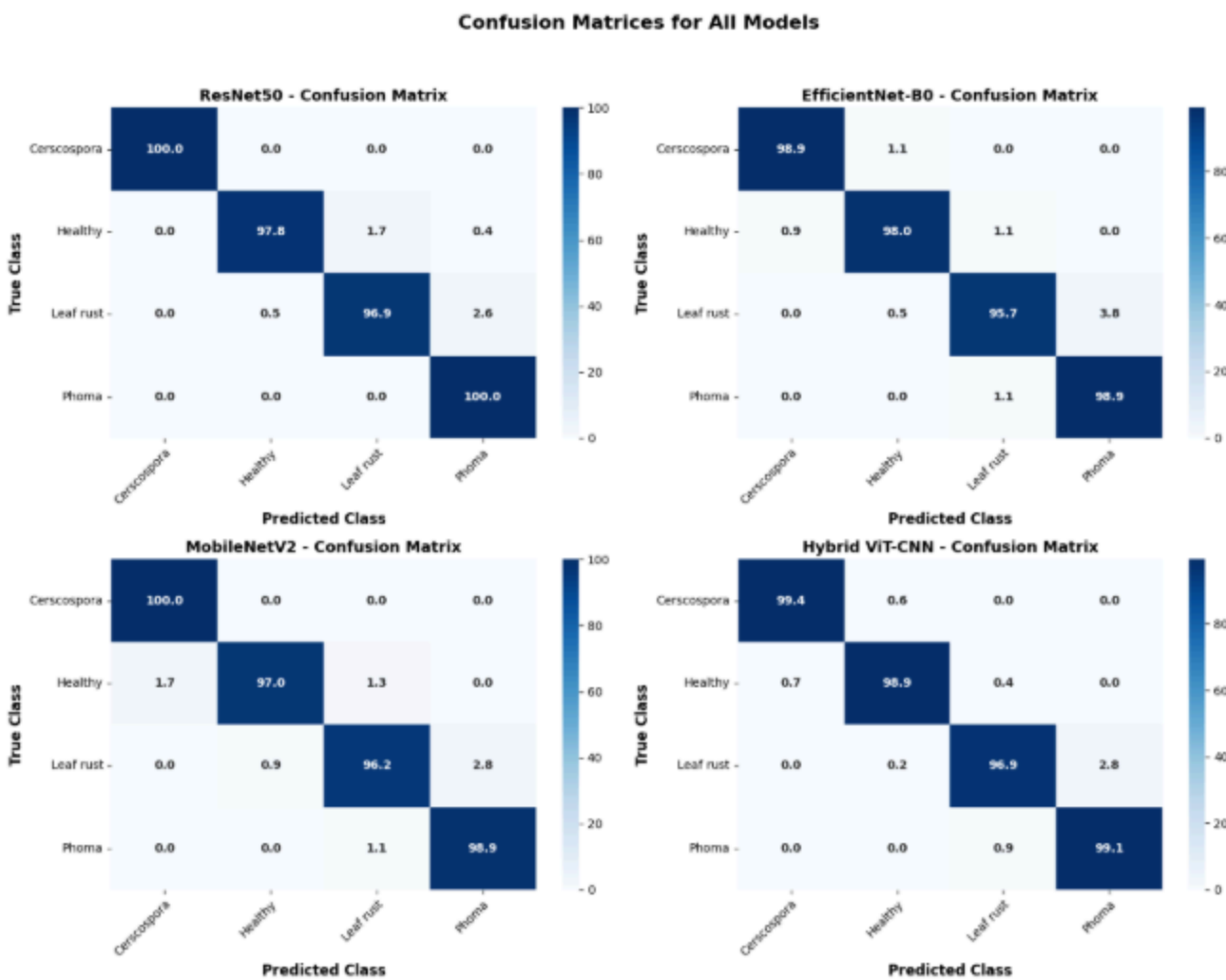


Figure 4

Confusion Matrices for All Models

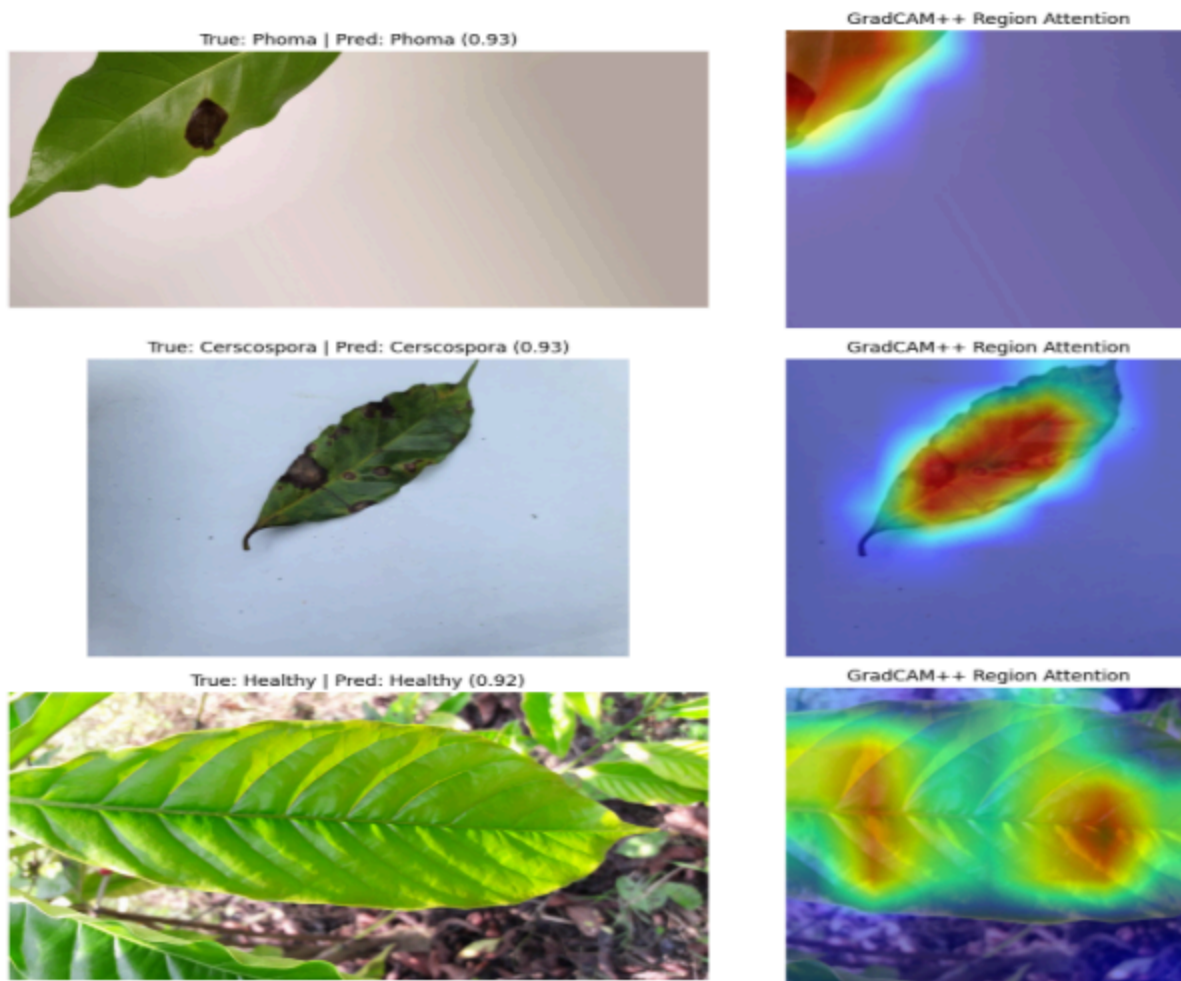


Figure 5

GradCAM++ Visualizations for Each Disease Class