Machine Vision–Based Deep Learning for Automated Crop Disease Classification in Precision Agriculture Article

Kaznah Alshammari
Northern Border University
Muhammad Sheraz Khan
Islamia College University
Kainat Nisa
Islamia College University
Irshad Ahmad
irshad@icp.edu.pk

Islamia College University

Article

Keywords: Precision Agriculture, Machine Learning, Deep Learning, Artificial Intelligence, Machine Vision

Posted Date: March 5th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8761764/v1

Additional Declarations: No competing interests reported.

Machine Vision–Based Deep Learning for Automated Crop Disease Classification in Precision Agriculture Article

Kaznah Alshammari[1], **Muhammad Sheraz Khan**[2], **Kainat Nisa**[2], **and Irshad Ahmad**[2]

[1]Department of Information Technology, Faculty of Computing and Information Technology, Northern Border University, 91911, Rafha, Saudi Arabia
[2]Department of Computer Science, Islamia College University, Peshawar, Khyber Pakhtunkhwa, 25000, Pakistan
Corresponding author: (Irshad Ahmad), irshad@icp.edu.pk

ABSTRACT
Cherry is widely cultivated but remains challenging to harvest due to economic and ecological constraints, especially in developing countries such as Pakistan. Climate change, limited use of technology, and foliar diseases worsened by pesticide use further reduce productivity, particularly during fruiting. Conventional disease assessment depends on expert observation and grower experience, making it subjective and time-consuming. A comprehensive evaluation was conducted on the PlantCity dataset, which contains 5,714 high-density, full-color RGB images collected under challenging conditions and categorized into 5 classes. We compared three approaches: deep learning pre-trained, transfer learning, and a machine learning pipeline. Models were evaluated by accuracy, precision, recall, F1-score, Cohen's Kappa, inference time, FLOPs, and throughput. Grad-CAM was used to improve interpretability. Transfer learning using DenseNet169 achieved the highest performance, with 99.80% accuracy, 99.80% precision, 99.80% recall, and a Cohen's Kappa of 99.74%. These results were significantly higher than those obtained by other deep learning architectures and handcrafted baselines. Grad-CAM heatmaps confirmed that the models focused their attention on pathological areas. The proposed transfer-learning-based framework, particularly DenseNet169, demonstrates state-of-the-art diagnostic accuracy and features a modular structure. This design enables deployment on both high-performance servers and resource-constrained embedded devices, thereby facilitating early disease detection in precision agriculture.

Academic Editor: Firstname Lastname Received: date Revised: date Accepted: date Published: date

Keywords: Precision Agriculture, Machine Learning, Deep Learning, Artificial Intelligence, Machine Vision

1. Introduction

Early diagnosis of plant diseases is essential for maintaining crop yields and overall agricultural productivity. If not detected promptly, plant diseases can rapidly spread, causing significant damage to individual plants and, in severe cases, entire fields or regions. This spread can result in substantial economic losses and threaten food security. Numerous species-specific diseases affect plants, originating from bacteria, fungi, virus-carrying insects, or harmful maggots that proliferate in damaged fruit fly eggs. Additionally, contamination from insecticides or contact with infected plants can exacerbate disease transmission. Without timely intervention, these infections may disseminate across extensive croplands, ultimately diminishing plant growth and harvest quality.

Traditionally, plant diseases have been diagnosed by agricultural experts through visual inspection of leaves, roots, stems, and seeds for observable symptoms. In addition to reducing crop yield, plant diseases can cause significant or partial damage to plants. Therefore, early and accurate detection is essential for effective intervention. Recently, computer-assisted systems have been developed to support disease evaluation and enhance detection rates compared to traditional expert-based diagnosis [1]. Powdery mildew, for example, is a prevalent disease affecting various plant species, including apricots, peaches, plums, roses, and cherries. Typical symptoms of mildew-infected leaves include a white powdery coating, leaf distortion, scrappiness, and premature defoliation [2]. Computer-aided agricultural operations have advanced considerably, aiming to improve the efficiency of agricultural processes through automated programs [3]. Currently, computer-aided technologies are widely implemented in activities such as seeding, irrigation, spraying, yield assessment, and harvesting. Among these applications, disease diagnosis has become a critical tool for sustaining plant productivity. Most of these systems use machine learning (ML) methods, often relying on expert-labeled datasets [4]. However, some advanced techniques, despite their speed or accuracy, have seen limited adoption due to high costs and the requirement for specialized expertise. Image-based machine learning methods offer high accuracy, real-time performance, and user-friendly interfaces for agricultural practitioners. In this context, plant disease diagnosis is treated as an image classification problem that involves feature extraction and classification [5]. ML algorithms have demonstrated high overall accuracy, yet identifying an optimal classification algorithm remains challenging due to performance variability across different input data types [6]. Although deep learning (DL), particularly Convolutional Neural Networks (CNNs), is increasingly applied to disease diagnosis and classification, the integration of Internet of Things (IoT) and Artificial Intelligence (AI) technologies further enables early and precise detection of plant diseases [7,8]. Through image analysis, these technologies offer an efficient, labor-saving alternative to manual monitoring, as diseases can be identified from recorded plant images [9]. Moreover, the combination of AI and computer vision enables early detection of plant diseases, enabling timely interventions to mitigate adverse effects [10].

Recent studies on plant disease detection have examined various plant parts, including roots, stems, blooms, and leaves, with leaf-based analysis being the most extensively investigated approach. A primary challenge in computer-aided detection systems is identifying plant part characteristics, such as roots, stems, flowers, and leaves that indicate disease symptoms [11]. Deep features learned directly by convolutional neural networks (CNNs) have demonstrated superior accuracy and efficiency compared to traditional methods that rely on manually extracted image features for classification. For example, a combination of GoogleNet, support vector machine (SVM), k-nearest neighbor (KNN), and backpropagation (BP) neural networks was employed to classify cherry leaf powdery mildew disorders, with the CNN-based method achieving the highest accuracy of 99.6% on dataset [12]. Additionally, an EfficientNet-based CNN was evaluated using transfer learning on the PlantVillage dataset, where the B4 and B5 models achieved accuracy rates of 99.97% (augmented data) and 99.91% (original data), respectively [13]. A novel deep CNN-based model that combines GoogLeNet Inception and Rainbow coupling was developed for the diagnosis of apple leaf disease. This model, trained on a dataset of 26,377 damaged apple leaves to recognize five common apple leaf diseases, achieved an accuracy of 78.80% [14]. Furthermore, a method for detecting grape diseases using principal component analysis and backpropagation networks was introduced, with the dataset including grape downy mildew and grape powdery mildew, achieving a prediction accuracy of 94.29% [15]. In another study, four classifiers were assessed after extracting color, size, shape, and texture features from images of fruit samples across 15 categories. These features enabled KNN to achieve an accuracy of 81.9%, outperforming both SVM and artificial neural networks for fruit classification and recognition [16]. Katarzyna R et al. [17] introduced a nine-layer deep neural network that classified six apple varieties, achieving an accuracy of 99.78%. In the context of cherry sorting, a hybrid pooling-based CNN achieved 99.41% accuracy, surpassing KNN, ANN, LBP, and HOG methods. This approach leverages deep learning to improve detection by capturing subtle distinctions and eliminating the need for manual feature selection [18]. For cherry detection, a deep learning model based on an enhanced YOLOv4 architecture with a DenseNet backbone and Leaky ReLU activation was developed to identify ripe, semi-ripe, and unripe cherries, resulting in a 94.7% F1-score [19]. Sethy PK et al. [20] demonstrated that pre-trained models, specifically AlexNet and VGG16, can classify tomato crop diseases using 13,262 images, achieving accuracy of 97.49%. Gehlot M et al. [21] reported that the AlexNet CNN pre-trained model achieved 95.75% accuracy in tomato leaf disease classification through transfer learning. Tomato leaf disease detection achieved high accuracies of 99.51% (five classes), 98.65% (seven classes), and 97.11% (ten classes) using a conditional generative adversarial network (C-GAN) and a pre-trained DenseNet121 model [22]. An ImageNet-pretrained VGG16 model was selected for the automatic diagnosis and detection of disease in rice leaf plants, achieving an accuracy of 92.00% [23]. Kayaalp et al. [24] developed a deep learning-based classification system for seven Turkish cherry varieties using a novel dataset (ISP7Cherry) containing 3,570 images. The DenseNet169 model achieved 99.57% accuracy on augmented data, and an ensemble of the two best models using Maximum Voting achieved perfect accuracy. Dönmez et al. [25] proposed a system for identifying bacterial disease in cherry leaves using deep features extracted from the DarkNet-19 CNN and a PlantVillage dataset comprising 1,906 images across three categories: healthy, less diseased, and very diseased. Using LDA, KNN, and SVM, the highest performance achieved was 88.1%. Murugesan et al. [26] introduced a hybrid ConvNet-ViT model for multiclass leaf disease classification in banana, cherry, and tomato plants, utilizing a public dataset of healthy and diseased leaves. This approach, which integrates

CNN-based local feature extraction with vision transformer (VT)-based global context modeling, achieved 99.29% test accuracy in five-fold cross-validation and outperformed state-of-the-art pre-trained models, including EfficientNetV2, ConvNeXt, Swin Transformer, and ViT.

A novel framework is introduced to enhance computational efficiency and accuracy in cherry leaf disease classification. This approach integrates pre-trained deep learning models, transfer learning, and hybrid LBP-SIFT feature extraction, combined with ensemble machine learning classifiers (KNN, RF, SVM), to provide scalable solutions for precision agriculture processing. The main contributions are as follows:

- This study utilizes a balanced cherry leaf disease dataset comprising five classes from PlantCity [27], which exceeds the class diversity found in previous works that included only one to three classes. This advancement facilitates more accurate disease differentiation and supports the implementation of targeted orchard treatments.
- DenseNet169, when applied with transfer learning, achieves the highest performance, with accuracy, precision, and recall all reaching 99.80%. The proposed model outperforms all baseline deep learning models, including MobileNet, VGG, InceptionV3, and Xception, as well as handcrafted feature extraction pipelines. These results suggest that dense connectivity and feature reuse are particularly advantageous for small- to medium-sized agricultural datasets.
- Another notable strength is the focus on deployment efficiency. The trade-offs among the deep learning models evaluated depend on the specific application. DenseNet169 provides higher accuracy but requires greater computational resources. In contrast, lighter models such as MobileNet enable faster inference and reduced memory usage, making them more suitable for edge devices with limited resources. The availability of various model sizes and weights allows for flexible deployment tailored to the requirements of mobile-based plant health monitoring systems.
- Benchmarking is conducted across multiple paradigms. This study systematically compares end-to-end DL, TL, and classical ML using texture descriptors (LBP and SHIFT). The above results demonstrate that TL with DenseNet169 is the best option in terms of both accuracy and reliability, achieving a Cohen's Kappa of 99.74%.

2. Materials and Methods

2.1 Model Architecture Overview:

The proposed hybrid plant leaf disease classification system offers several advantages over existing methods and is well-suited for field implementation in agricultural settings. Its modular architecture enables the integration of diverse techniques. Deep learning branches that utilize pre-trained DL facilitate end-to-end learning and effectively capture complex patterns, subtle texture changes, color gradient variations, and lesion boundaries in diseased leaves. These capabilities are essential for distinguishing between similar symptoms. In contrast, the traditional machine learning branch, which employs handcrafted features such as local binary patterns (LBP) and scale-invariant feature transform (SIFT), provide enhanced interpretability and can substantially reduce computational requirements. Therefore, this approach is appropriate for deployment of resource-constrained edge devices.

The data flow is standardized across all branches. Raw RGB image tensors, downsized to 224×224 pixels, are initially preprocessed with normalization and augmentation to enhance robustness against variations in lighting, orientation, and scale. The preprocessed images are subsequently input into the respective feature extractors, which generate either high-level semantic features in DL or local texture and shape descriptors in machine learning (ML). Extracted features are then either pooled using Global Average Pooling in DL or concatenated as histograms for Local Binary Patterns (LBP) and Scale-Invariant Feature Transform (SIFT) before being forwarded to classification heads. Final predictions are combined using ensemble voting, majority voting, or weighted voting across branches to improve overall reliability and reduce false positives in critical diagnostic applications.

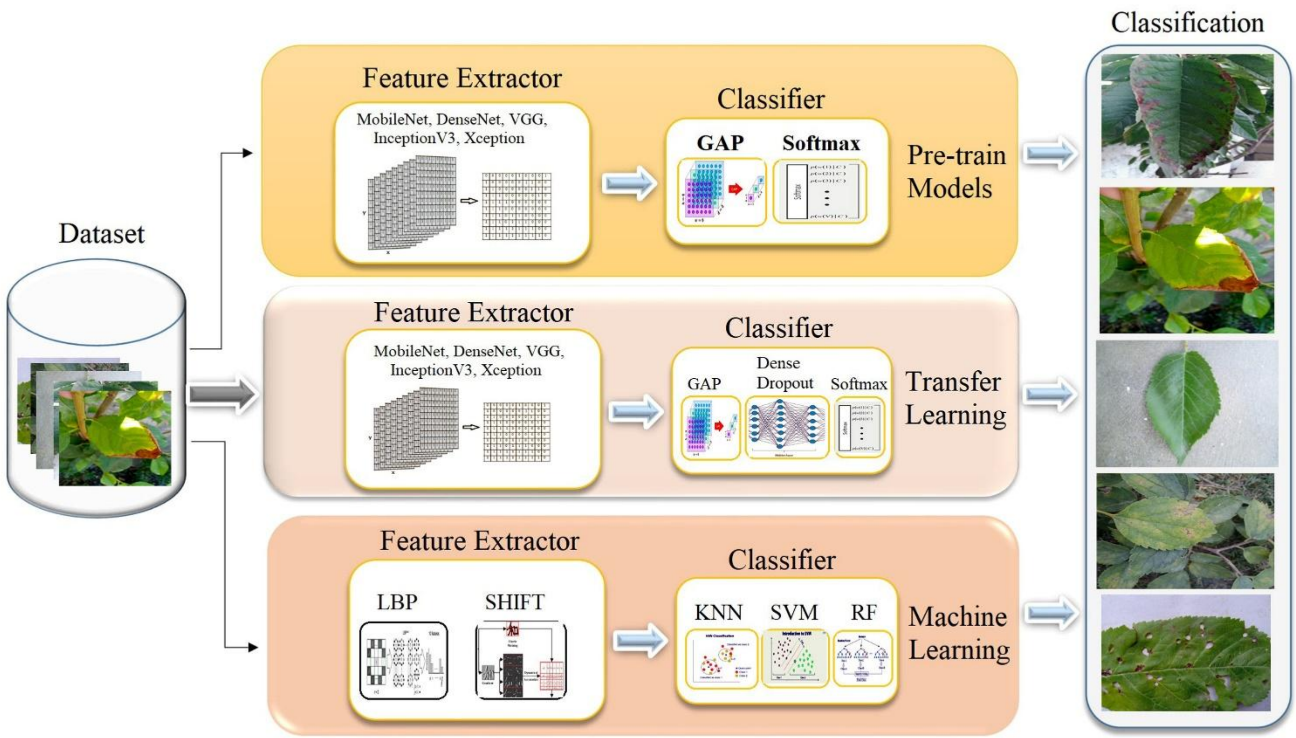

Figure 1: Overall pipeline of the proposed DL-based classification system for cherry leaf diseases.

2.2 Dataset

The PlantCity dataset [27], a publicly available benchmark comprising 5,714 high-resolution RGB images of cherry leaves, is employed in this study. The dataset contains five diagnostically relevant classes: Cherry brown spot, Cherry leaf scorch, Cherry normal leaf, Cherry purple leaf spot, and Cherry shot hole. To ensure balanced representation and robust model evaluation, a stratified sampling strategy divides the dataset into 3,496 training images, 500 validation images, and 999 test images. This approach reduces evaluation bias. The complete class-wise distribution, detailing the number of images per class in each partition, is provided in Table 1. For reproducibility, all input images are resized to 224 × 224 ×

3 (height × width × RGB channels) and normalized using the mean and standard deviation values from the ImageNet dataset for all experiments. Representative images illustrating each class are presented in Figure 2.

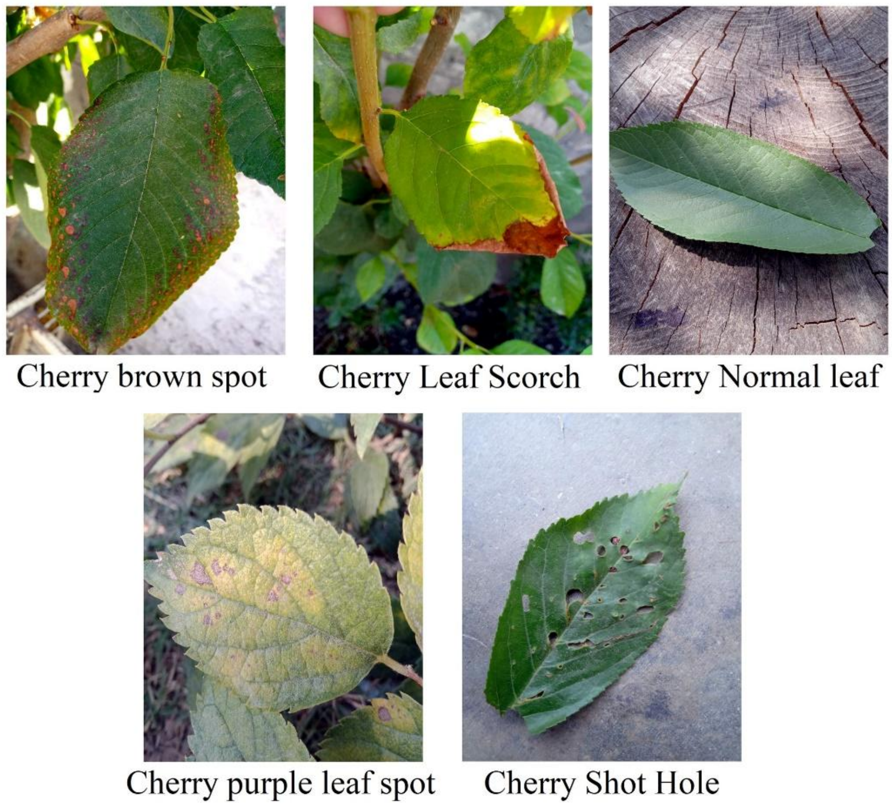

Figure 2: Sample images for each class of cherry leaves in PlantCity [27].

Table 1. The number of images per class is distributed across the test, train, and validation sets.

Class	**Training**	**Testing**	**Validation**	**Total**
Cherry brown Spot	847	242	121	1210
Cherry Leaf Scorch	960	282	141	1383
Cherry Normal leaf	437	125	63	625
Cherry purple leaf spot	867	240	120	1227
Cherry Shot Hole	385	110	55	550
Total	3496	999	500	4995

2.3. Preprocessing:

Preprocessing the dataset is fundamental for efficient model training. A 70%/20%/10 % training/testing/validation dataset split is used. All cherry leaf images are rescaled to 224×224×3 pixels. To increase data diversity and mitigate overfitting, the training set is augmented with rotation, shear, zoom, horizontal flip, rescaling, height shift, width shift, and vertical flip. These methods substantially improve the model's generalization and reduce the risk of overfitting, thereby enhancing classification accuracy and robustness in detecting cherry leaf diseases. The model thereby learns invariant patterns across different image transformations.

2.4. Deep Learning

With the popularity of deep learning, a slew of image recognition models has been proposed that can be used to solve crop leaf identification well. Popular deep convolutional neural network models are now applied in practice [28].

2.5. Dense Network

Dense convolutional network (DenseNet), each layer is connected to every other layer in a feed-forward fashion Figure 3. Traditional convolutional networks with L layers have L connections-a connection between each layer and its subsequent layer. Our network has direct connections. The feature maps of all previous layers are fed into the computation of each layer, and its feature maps are forwarded to all succeeding layers [29].

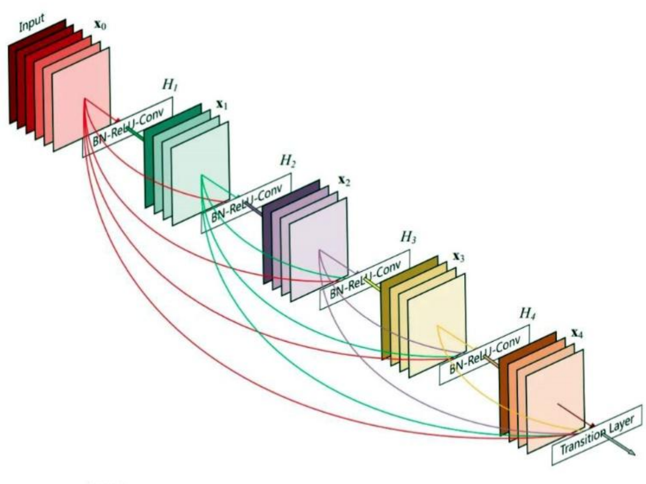

Figure 3: DenseNet Layer overview (Source from Google).

2.6. Transfer Learning

Transfer learning (TL) is a valuable technique for adapting networks and models originally designed for specific applications to new tasks [30]. In image classification, training DL models from scratch is both time-consuming and resource intensive. Furthermore, DL models require large datasets, which pose challenges when only small datasets are available [31]. To address these limitations, state-of-the-art pre-trained models were applied using TL. These pre-trained models are suitable for prediction, feature extraction, and fine-tuning [32]. The primary approach involves leveraging pre-trained models without their final classification layers for feature extraction, then using these features as input to shallow classifiers. A summary of the TL model and DenseNet169 used in this study is presented in Table 2.

Table 2. Summary of the proposed TL DenseNet169.

Layer	Output Shape	Activation
DenseNet121	(7,7,1024)	ReLU
GAP	(1024)	-
Dense	256	ReLU
Dropout	0.2	-
Dense	256	ReLU
Dropout	0.2	-
Output (Dense)	5	Softmax

2.6.1 Mathematical Formulation of Transfer Learning Pipelines

The overall classification function is defined as:

$$f = CNN_{\theta *}(\mathrm{x}) \in \mathbb{R}^d \tag{1}$$

Where $CNN_{\boldsymbol{\theta}*} \in$ {MobileNet, DenseNet, VGG, InceptionV3, Xception} with pre-trained weights 0^* from ImageNet (frozen).

Apply Global Average Pooling (GAP): is applied to the feature map to obtain a fixed-length vector:

$$z = GAP(f) = \frac{1}{H\prime W\prime} \sum_{i=1}^{H\prime} \sum_{j=1}^{W\prime} f_{i,j} \in \mathbb{R}^{d} \tag{2}$$

The last mile: The final convolutional feature map is space-averaged using equation 2, which pools it into a d-dimensional descriptor, irrespective of resolution.

Final prediction is made by a single linear layer followed by softmax in the frozen-backbone transfer learning:

$$\hat{\mathrm{y}} = \arg\max_k \quad Softmax(Wz + b)k \tag{3}$$

$$p(y = k|x) = \frac{\exp((wz+b)k)}{\sum_{j=1}^{c} \exp((wz+b)j)} \tag{4}$$

Equations (3) and (4) denote the class prediction and its associated probability distribution obtained by a lightweight classifier head on top of ImageNet features.

For full fine-tuning, the backbone parameters are updated as:

$$f = TL_{\theta}(x) \in \mathbb{R}^{H\prime \times W\prime \times d} \tag{5}$$

Where $\theta = \theta_{pre} \ + \Delta\theta(pre - trained \ + \ fine - tuned)$

Equation 5 is an end-to-end trainable feature extractor with learnable parameters θ.

The final fine-tuning classifier is expressed as

$$z = GAP(f) \in \mathbb{R}^d \tag{6}$$

$$h = ReLU(w_{1z+b1}) \in \mathbb{R}^m \tag{7}$$

$$h' = Droupout(h, p) \tag{8}$$

$$o = W2h' + b_{\ 2} \in \mathbb{R}^C \tag{9}$$

$$p(y = k|\ x) = Softmax(o)k \tag{10}$$

Equations (6)– (10) represent the sequence of operations in the custom head: global average pooling, followed by a fully-connected layer with ReLU activation, dropout for regularization, final linear projection, and softmax probability output. When applied to DenseNet169, this scheme yielded the best-performing model (99.80% accuracy) and was used in our experiments.

2.7. Handcrafted Features and Machine Learning

The integration of handcrafted feature extraction with conventional ML classifiers, as a complementary approach to DL frameworks, yields a computationally efficient, interpretable model for plant disease classification. This approach is particularly suitable for deployment on resource-constrained edge devices in remote agricultural settings. The process begins by generating domain-specific descriptors from leaf images that capture texture, color, and structural abnormalities indicative of disease symptoms. Local Binary Patterns (LBP) characterize local texture structures by comparing pixel intensities within neighborhoods, yielding rotation-invariant histograms that emphasize lesion edges and spotting patterns commonly observed in foliar diseases. Similarly, Scale-Invariant Feature Transform (SIFT) detectors identify scale- and rotation-invariant key points, and the corresponding descriptor vectors are clustered using a bag-of-visual-words (BoVW) model to achieve a comprehensive representation of disease motifs. These features are consolidated into a fixed-size vector, which serves as input for classifiers such as K-Nearest Neighbors (KNN), an instance-based method; Support Vector Machines (SVM), which identifies the maximum margin separating hyperplane in high-dimensional space; and Random Forests (RF), an ensemble of decision trees constructed with bootstrap samples to mitigate overfitting through bagging.

$$LBP\left(x_{c\ ,\ y_c}\right) = \sum_{p=0}^{p-1} s\left(g_{p-g_c}\right).2^p \ , s(t) = \begin{cases} 1 \ \ t \geq 0 \\ 0, \ t < 0 \end{cases} \tag{11}$$

Equation (11) gives the uniform rotation-invariant LBP code by thresholding each neighborhood pixel relative to the center pixel and obtaining a 256-bin histogram $hLBP \in \mathbb{R}256$, which serves as a texture descriptor.

SIFT: Detect key-points$\{k_{i,d_i}\}$, cluster $d_i \in \mathbb{R}^{128}$ →BoVW histogram $h_{SIFT} \in \mathbb{R}^K$

Combined Feature Vector:

$$v = [h_{LBP,} \; h_{SIFT}] \in \mathbb{R}^{256+k} \tag{12}$$

Equation (12) represents the combined texture + key-point descriptor fed to classical machine learning classifiers.

The evaluated classifiers are defined as follows:

$$KNN: \hat{y} = \{y_i : j \in \mathrm{N}_k(\mathrm{v})\} \tag{13}$$

Equation (13) is a majority voting among the k nearest training samples in feature space (k= 5).

$$SVM: \hat{y} = (w^t \; \phi \; (v) + b) \tag{14}$$

In (14), we have the linear SVM decision function trained on RBF kernel-mapped features.

$$RF: \hat{y} = \{T_t \; (v)\}_{t=1}^{T} \tag{15}$$

Equation 3 gives the ensemble prediction that resulted from majority voting over T = 500 decision trees, i.e., the handcrafted pipelines setting, which reached the highest accuracy.

3. Results

3.1. Experimental Setup

All experiments were conducted using the TensorFlow 2.12.0 framework on a workstation equipped with an Intel Core i7-3770 CPU, an NVIDIA RTX 2060 Super GPU with 8 GB GDDR6 memory, 20 GB DDR3 RAM, and CUDA 11.2 with cuDNN version 8.6 for GPU-based training. A consistent training protocol was applied to all deep learning models, utilizing a batch size of 20, 20 epochs, the Adam optimizer with an initial learning rate of 0.001, and categorical cross-entropy loss. Early stopping was implemented after 7 epochs without improvement in validation loss, and model checkpoints were saved based on validation accuracy.

3.2. Performance Metrics

To provide a holistic evaluation of the proposed hybrid framework for cherry leaf disease classification, we used a set of established performance metrics that trade off overall correctness against per-class reliability and agreement beyond change—critical value in practical applications such as agricultural diagnostics, where misclassification can result in excessive chemical treatments or undetected spread of disease.

Accuracy is a machine learning metric that measures the proportion of correct predictions. It is the proportion of predictions that were accurate out of all predictions.

$$Acuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{16}$$

Cohen's kappa is a statistic that considers agreement beyond what would be expected by chance. It is often used to test machine learning models or to serve as a human rater(s) in classification and labeling tasks, especially when multiple ratings per task are collected. The equation for Cohen's kappa is:

$$k = \frac{p_o - p_e}{1 - p_e} \tag{17}$$

Precision is the measure of how many positive predictions made by a model were actual positive instances. It is the answer to this question: "Out of all actual positive instances, how many were correctly picked as positive by the model?" The formula for precision is:

$$Pecision = \frac{TP}{TP+FP} \tag{18}$$

Recall is a measure of how well the model recognizes all positive examples. It answers the question: "Out of all actual positive examples, how many were predicted as positive"? The formula for recall is:

$$Recall = \frac{TP}{TP+FN} \tag{19}$$

Throughput (Samples per Second): The processing speed of the model. Formula (20) indicates the total number of real test cherry leaf images classified per second.

$$Throughput = \frac{N}{T}[samples/sec] \tag{20}$$

Batch throughput quantifies the speed at which a GPU processes multiple images simultaneously. Eq. (21) describes the effective throughput of a given batch size B, which is an important parameter for parallel inference in Agriculture Edge

$$Throughput_{\ batch} = \frac{B}{t_{\ batch}}[samples/sec] \tag{21}$$

Flops (Floating Point Operations) quantify the theoretical computational cost of the network. Equation (22) represents the total number of multiply-accumulate operations across all L layers and serves as a hardware-independent complexity measure.

$$FLOPS = \sum_{\ell=1}^{L} FLOPs\ell \tag{22}$$

For a convolutional layer (ℓ): Equation (23) represents the dominant computational cost in CNNs, where factor 2 accounts for multiplication and addition in each multiply-accumulate operation

$$FLOPS_{sconv} = 2 \times H_{out} \times W_{out} \times c_{out} \times \left(c_{in} \times k_{h \times k_w} \right) \quad (23)$$

For a fully connected layer:

$$FLOPs_{sfc} = 2 \times I \times O \quad (24)$$

The cost of dense layers is generally negligible compared to that of convolutional layers in recent backbones, as indicated by Equation (24).

Average Inference Time Per Image

$$t_{avg} = \frac{T}{N} = \frac{1}{Throughput}\, t\ [seconds/image] \quad (25)$$

Equation (25) represents the practical latency experienced when classifying a single cherry leaf image in the field, directly impacting real-time monitoring applications.

3.3. Experimental Results

The experimental results demonstrate the robustness and effectiveness of the hybrid approach in classifying cherry leaf diseases, achieving high performance on the PlantCity dataset [27]. This dataset comprises five diagnostically challenging classes: Cherry brown spot, Cherry leaf scorch, Cherry regular leaf, Cherry purple leaf spot, and Cherry shot hole. The dataset was stratified into 70% training, 10% validation, and 20% testing, ensuring adequate representation of all classes and mitigating the biases commonly observed in imbalanced agricultural datasets, which often overrepresent healthy leaves.

The results from transfer learning, pre-training, and classical ML approaches are comprehensively detailed in Tables 3, 4, and 5, respectively. Table 3 presents the evaluation of models including DenseNet121, DenseNet169, DenseNet201, MobileNet, MobileNetV2, VGG16, VGG19, NASNetMobile, Xception, and InceptionV3 across multiple performance metrics: parameter counts, cold start load time, throughput (samples per second), FLOPs, average inference time per image, precision, recall, F1 accuracy, and Cohen's Kappa. The analysis reveals trade-offs between computational efficiency and classification effectiveness. Notably, lighter models such as MobileNetV2 demonstrate shorter inference latency and higher throughput, although this may come at the expense of reduced accuracy compared to deeper models like DenseNet201.

Table 3 shows that DenseNet169 outperformed the other nine transfer learning architectures evaluated, achieving 99.80% accuracy, 99.80% precision, 99.80% recall, and a Cohen's Kappa of 99.74%. This superior performance is attributed to DenseNet's dense connectivity pattern, which facilitates the reuse of features from lower layers in deeper layers and enables the detection of subtle disease symptoms, such as slightly irregular lesion boundaries and subtle red-to-green color gradients in cherry leaves. Although

DenseNet169 has a relatively large parameter count of 13.1 million and a model size of 49.8 MB, it achieves a strong balance between computational efficiency and predictive efficacy, with an average inference time of 110.92 ms and a throughput of 9.02 samples per second. In comparison, MobileNetV2 (2.65M parameters, 77 ms inference time) and other lightweight models provide faster inference but at a modest reduction in accuracy, decreasing to 98.90%. This result underscores the common trade-off between efficiency and accuracy in deployable systems. Conversely, more complex models such as InceptionV3 and Xception achieve over 98% accuracy but incur significantly higher FLOPs and execution times, making them less suitable for real-time edge applications.

Table 3. The comparison results for all models using the transfer learning approach.

Method	Param	Model Weight Size	Throughput samples/sec	Flops	Avg Inference time/image	Precision	Recall	Accuracy	Cohen's Kappa
DenseNet121	7,366,981	28.0	9.47s	5.67	105.64ms	98.70%	98.70%	98.70%	98.32%
DenseNet169	**13,136,197**	**49.8**	**9.02s**	**6.72**	**110.92ms**	**99.80%**	**99.80%**	**99.80%**	**99.74%**
DenseNet201	18,880,837	77.4	6.97s	8.58	143.53ms	99.70%	99.70%	99.70%	99.60%
MobileNet	3,558,341	16.2	14.52s	1.14	68.85ms	99.70%	99.70%	99.70%	99.61%
MobileNetV2	2,652,997	13.4	12.87s	0.60	77ms	99.00%	98.80%	98.90%	98.58%
VGG16	14,913,093	58.4	10.69s	30.71	93.58ms	94.98%	90.99%	93.09%	91.06%
VGG19	20,222,78	78.7	12.85s	39.04	77.81ms	94.02%	89.69%	91.79%	89.44%
NASNetMobile	4,607,385	21.3	8.29s	1.14	120.64ms	97.66%	96.20%	96.70%	95.73%
Xception	21,453,101	86.6	10.88s	9.11	91.94ms	99.20%	98.90%	99.10%	98.84%
InceptionV3	22,394,405	90.4	10.72s	5.69	93.29ms	98.90%	98.90%	98.90%	98.58%

Table 4 presents the complete results for full pre-training, including fine-tuning of all layers on the PlantCity [27] dataset. The outcomes are mixed when compared to transfer learning. DenseNet201 achieved the highest test accuracy of 97.50% and a Kappa score of 96.77%, attributed to domain-specific weight optimization for generalization. In contrast, framework-level statistics were lower than those observed with transfer learning for most models; for example, DenseNet121 accuracy decreased from 98.70% to 94.79%, likely due to overfitting on the limited 3,978-image training set. Lightweight models, such as MobileNet, benefited most from pre-training, achieving 97.10% accuracy. Inference time was slightly reduced for DenseNet201 (112.88 ms), while larger models, such as VGG19, remained computationally expensive.

Table 4. The comparison results for all models using the pre-training approach.

Method	Param	Model Weight Size	Throughput samples/sec	Flops	Avg Inference time/image	Precision	Recall	Accuracy	Cohen's Kappa
DenseNet121	7,042,629	28.0	10.13s	5.67	98.74ms	96.83%	91.79%	94.79%	93.28%
DenseNet169	12,651,205	49.8	8.58s	6.72	116.51ms	97.43%	94.79%	96.40%	95.35%
DenseNet201	**18,331,589**	**71.8**	**8.86s**	**8.58**	**112.88ms**	**98.46%**	**96.00%**	**97.50%**	**96.77%**

MobileNet	3,233,989	12.6	14.87s	1.14	67.26ms	97.65%	95.80%	97.10%	96.26%
MobileNetV2	2,264,389	9.14	11.94s	0.60	83.72ms	97.10%	93.79%	95.70%	94.45%
VGG16	14,717,253	56.2	13.11s	30.71	76.28ms	96.95%	95.68%	95.92%	94.83%
VGG19	20,026,949	76.5	12.69s	39.04	78.78ms	95.32%	95.12%	94.73%	94.18%
NASNetMobile	4,275,001	18.3	8.50s	1.14	117.64ms	96.44%	95.51%	94.31%	94.43%
Xception	20,871,725	80.0	11.88s	9.11	84.14ms	96.55%	89.69%	93.59%	91.69%
InceptionV3	21,813,029	84.1	11.09s	5.69	90.20ms	95.50%	91.39%	93.89%	92.10%

Table 5 presents the performance of traditional machine learning pipelines. The combination of LBP with RF achieved 98.00% accuracy and a Kappa score of 97.42%, with a model size of 9.08 MB and an inference time of 6.55 ms. This represents a significant speedup over deep learning models, achieving a throughput of 152.78 samples per second. The effectiveness of this approach is attributed to LBP's local texture encoding, which adequately represents spotting and edge patterns without requiring deep hierarchical structures. In contrast, SIFT-based pipelines, such as SIFT with KNN, achieved lower performance (95.40%), as SIFT's key-point emphasis is less effective for diffuse foliar symptoms. SVM variants also underperformed, likely due to the sensitivity of high-dimensional histograms and suboptimal kernel selection.

Table 5. Comparison results for all models using a machine-learning approach.

Feature Extractor	**Classifier**	**Model Weight Size**	**Avg Inference time/image**	**Throughput samples/sec**	**Accuracy**	**Cohen's Kappa**
LBP	KNN	674 kb	2.88ms	347.07s	86.69%	82.81%
	SVM	765 kb	0.59ms	1699.93s	57.76%	43.90%
	RF	**9.08 MB**	**6.55ms**	**152.78s**	**98.00%**	**97.42%**
SHIFT	KNN	1.58 MB	1.55ms	644.39s	95.40%	94.06%
	SVM	2.16 MB	0.58ms	1730.33s	77.48%	70.57%
	RF	7.11 MB	6.71ms	149.06s	88.89%	85.58%

Figure 4 shows the training and validation curves for accuracy, precision, recall, and loss across epochs using transfer learning with DenseNet169. The close alignment between validation and training statistics indicates rapid convergence without overfitting. The low final loss, combined with high precision and recall, underscores the effectiveness of feature adaptation for target classification.

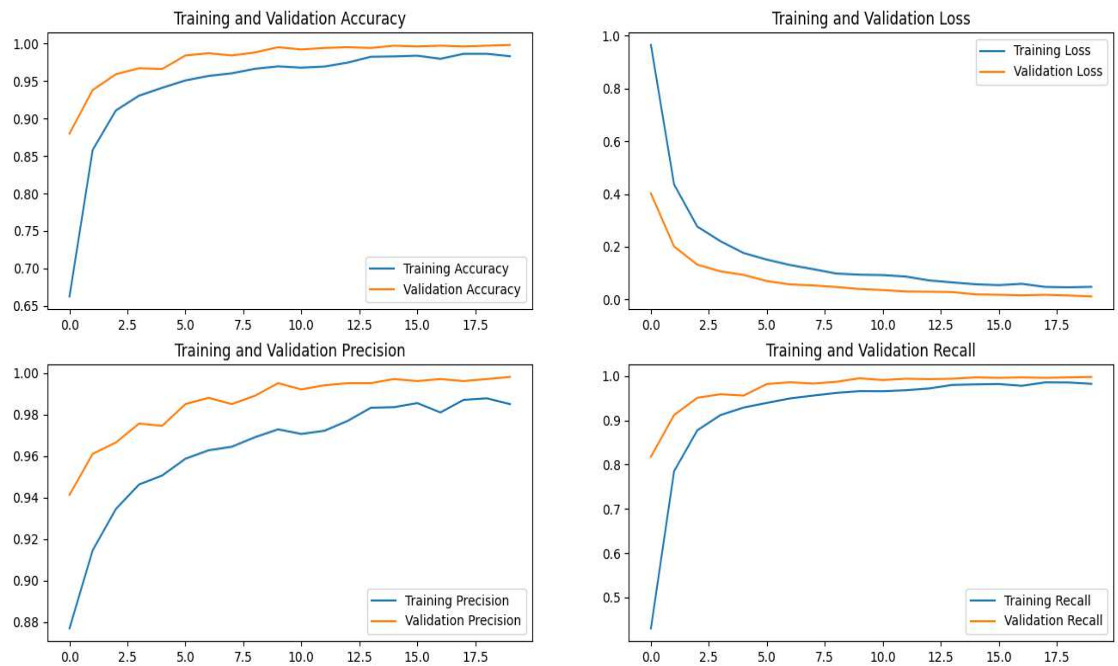

Figure 4: Training and validation curves of accuracy, precision, recall, and loss versus epochs for DenseNet169 using a transfer learning approach.

Figure 5 presents the confusion matrix for DenseNet169 utilizing the transfer learning approach, illustrating class-wise prediction performance. The predominance of diagonal elements indicates that the model achieves high accuracy and strong overall performance, whereas the presence of two smaller off-diagonal elements highlights specific class confusions. When considered alongside quantitative performance metrics, this matrix elucidates both the strengths and weaknesses of the model and system, as well as the conditions that lead to misclassification.

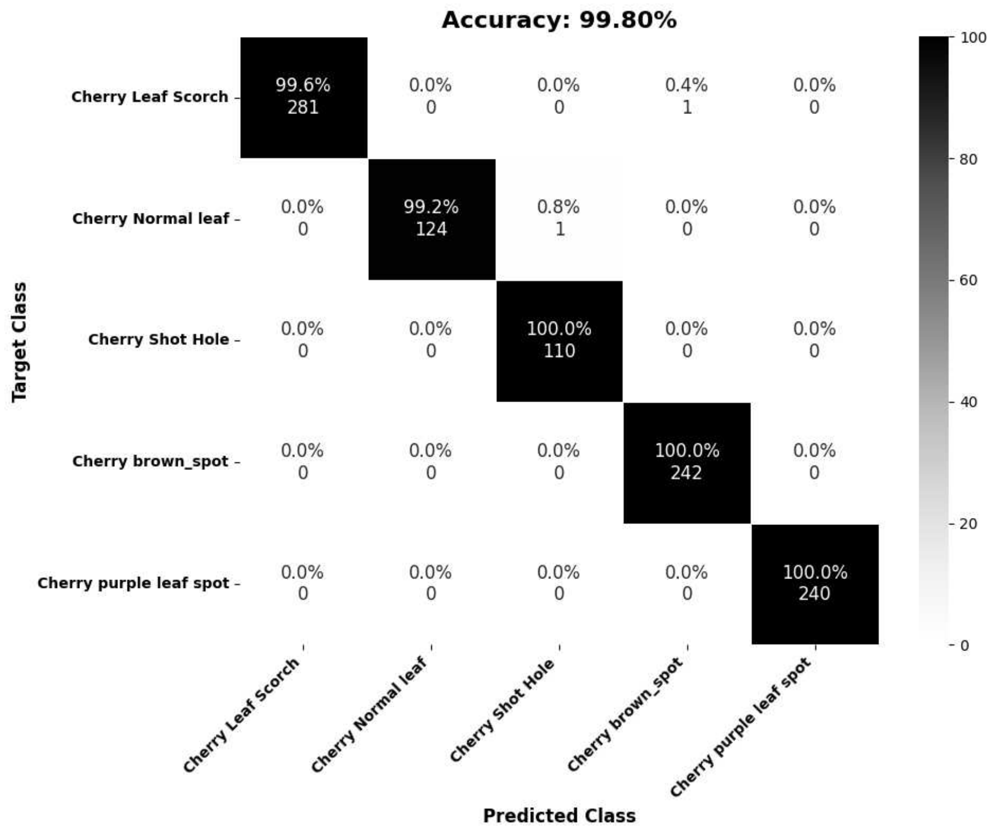

Figure 5: Confusion matrix of the DenseNet169 using the transfer learning approach.

Figure 6 presents the accuracy, precision, recall, and loss curves for the complete pre-trained DenseNet201 model across training epochs. The results indicate smoother convergence and a higher upper bound on peak performance compared to transfer learning, thereby demonstrating the effectiveness of end-to-end training. Additionally, the figure illustrates enhanced generalization resulting from pre-training on the target data.

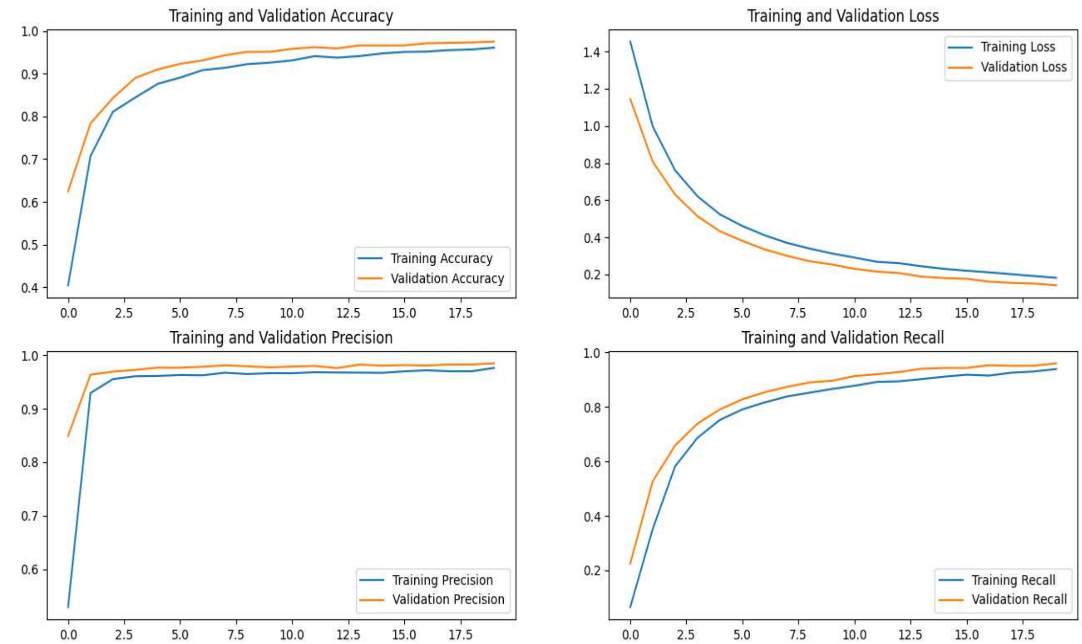

Figure 6: Training and validation curves of accuracy, precision, recall, and loss versus epochs for the fully pre-trained DenseNet201 model.

Figure 7 presents the confusion matrix for the DenseNet201 model, which closely aligns with an ideal diagonal for most classes. The observed reduction in misclassifications, compared to the transfer learning solution in Figure 5, demonstrates the advantages of full pre-training. This analysis offers insight into the remaining classification challenges.

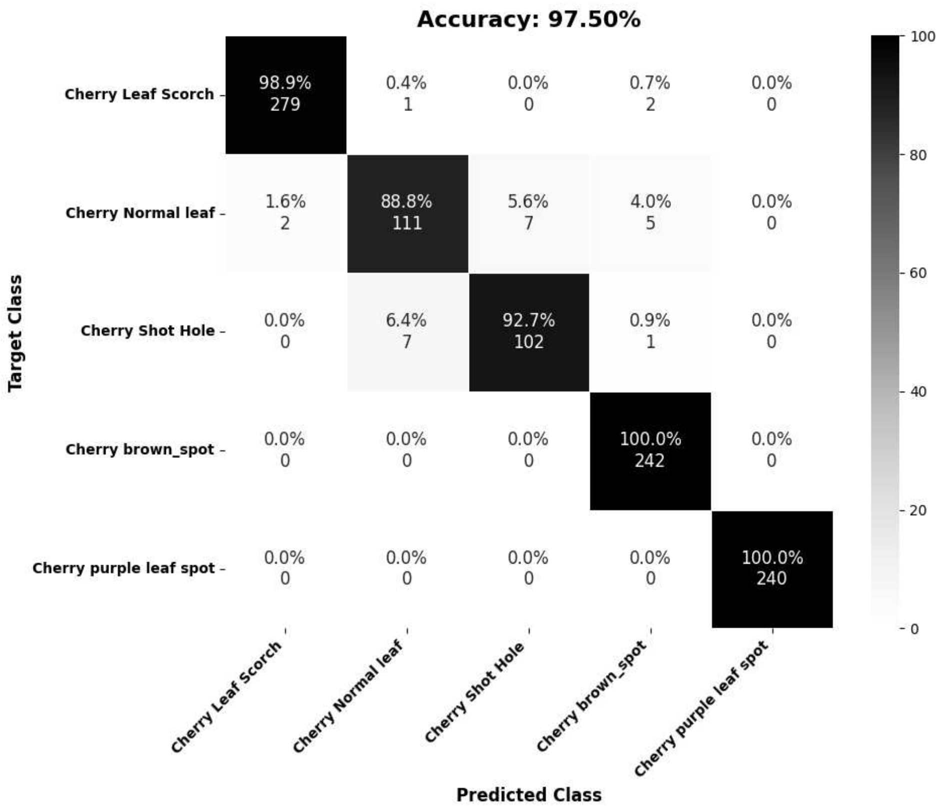

Figure 7: Confusion matrix of the fully pre-trained DenseNet201 model.

Figure 8 presents the confusion matrix for the ML pipeline using RF and LBP features. The matrix exhibits a strong diagonal concentration, indicating robust performance despite the pipeline's relatively simple design. Although off-diagonal errors are slightly higher compared to deep learning models, they remain minimal. This visualization effectively demonstrates the value of handcrafted features for efficient and interpretable classification.

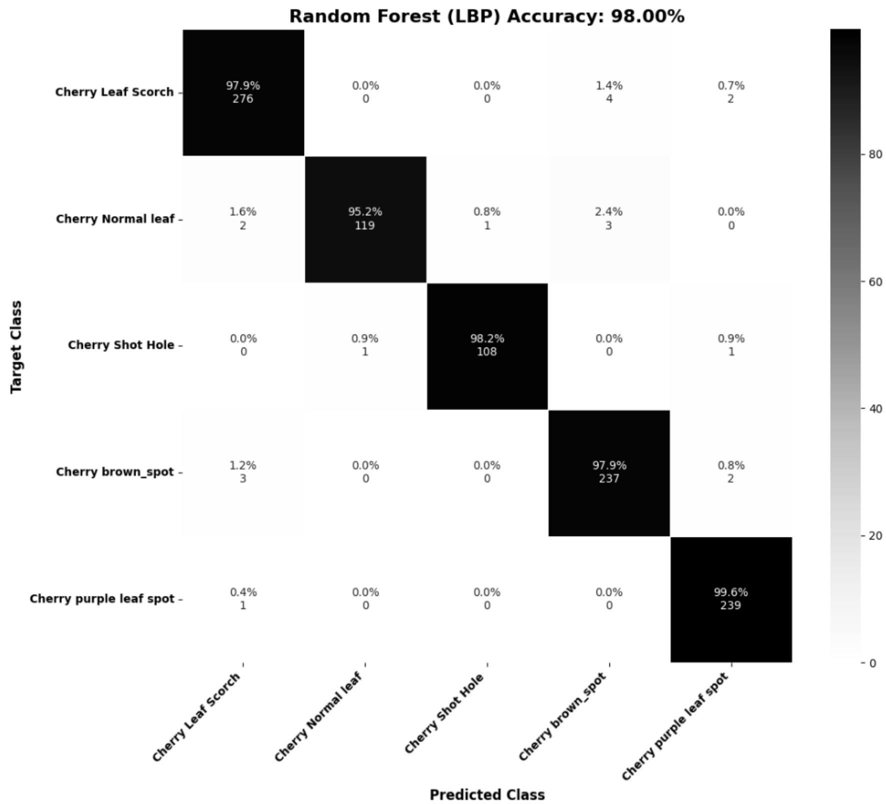

Figure 8: Confusion matrix of the Random Forest classifier with LBP features.

3.3. Explainable Artificial Intelligence (XAI)

To enhance the reliability and applicability of deep learning models for plant leaf disease classification, Explainable Artificial Intelligence (XAI) methods were implemented, including Gradient-weighted Class Activation Mapping (Grad-CAM). This post-hoc visualization technique generates heat maps that identify the most influential image regions for model predictions, thereby improving interpretability regarding the opaque nature of convolutional neural networks (CNNs) in transfer learning (TL) and pre-training approaches. Grad-CAM was applied to the top-performing models: DenseNet169 for the transfer learning scenario and DenseNet201 for the pre-trained scenario, which achieved accuracies of 99.80% and 97.50%, respectively. The process involved computing gradients of the target-class score with respect to the final convolutional feature maps, followed by global average pooling and a weighted sum to produce class-specific activation maps (Figure 9). The resulting heatmaps, overlaid on the original leaf images, demonstrated that the models consistently focused on pathological features such as necrotic lesions, chlorotic spots, and fungal structures, rather than non-pathological background elements. This finding supports the conclusion that the models engage in targeted learning of disease-relevant symptoms. The strong alignment with domain expert expectations, combined with high quantitative metrics (precision, recall, and Kappa >98% for top models), provides validation and reduces concerns about overfitting by demonstrating biologically realistic decision-making. In contrast to classical machine learning pipelines such as LBP+RF, which inherently provide more interpretable explanations of feature importance, integrating XAI bridges the gap between explanation and transparency in deep learning architectures. This advancement increases confidence in field applications and enables local farmers and plant pathologists to implement targeted control measures.

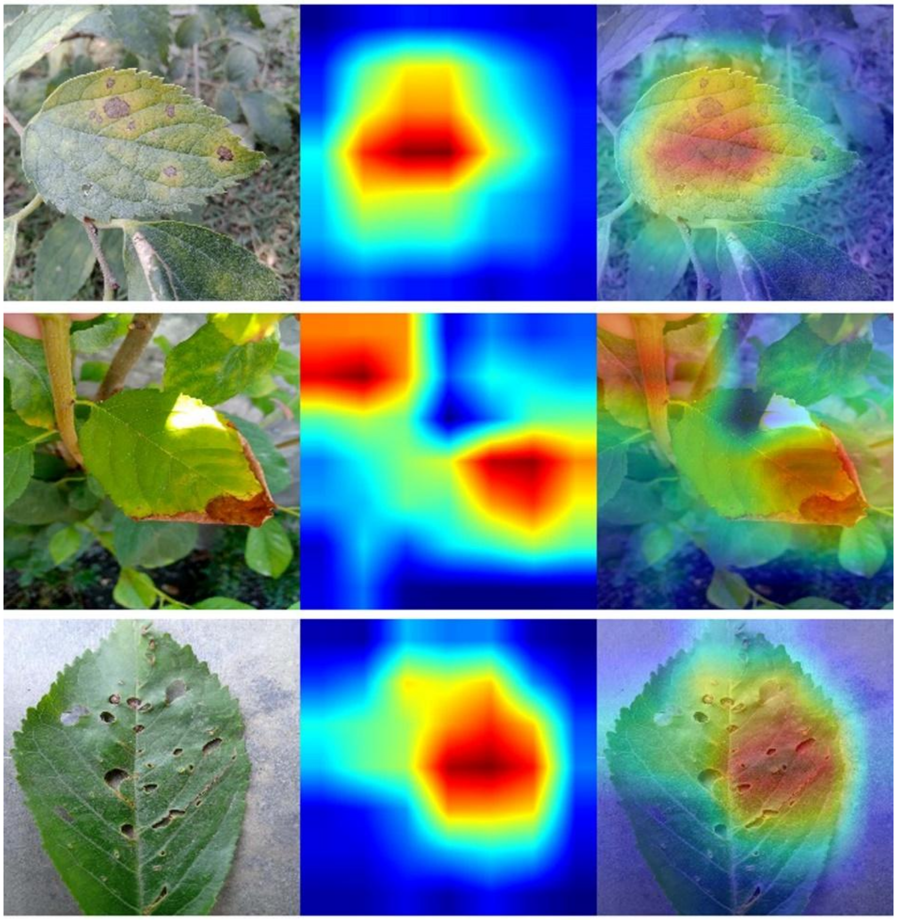

Figure 9: Demonstrates the XAI on TL DenseNet169 on some sample images.

4. Discussion

The experimental results demonstrate that deep learning models, especially those based on TL, outperform both fully pre-trained architecture and traditional machine learning pipelines for cherry leaf disease classification on the PlantCity dataset. Among the ten TL models evaluated, DenseNet169 achieved the highest performance, with an accuracy of 99.80%, precision and recall of 99.80%, and Cohen's Kappa statistic of 0.9974. DenseNet201 and Xception also performed well, each exceeding 99% accuracy. While MobileNet and MobileNetV2 exhibited slightly lower accuracy (99.70% and 98.90%, respectively, compared to Inception-V3), they required substantially less computational power (0.6–1.14 GFLOPs), making them suitable for deployment on resource-limited devices such as drones or edge nodes in orchards. In contrast, larger architectures such as VGG16 and VGG19 were considerably slower and achieved only 91–93% accuracy, indicating that increasing network depth without sufficient feature connectivity or parameterization may not enhance performance.

Global pre-training on the target dataset, while keeping ImageNet weights unfrozen, yielded fewer improvements and generally lower performance than the transfer learning approach. For example, when fully pre-trained, DenseNet201 achieved 97.50% accuracy. Although this outcome is well above chance, it remains significantly below the accuracy range observed when training networks with frozen pre-trained weights and fine-tuned classifiers (99.10–99.80%). These findings support the established advantage of transfer learning for relatively small, domain-specific agricultural datasets. ImageNet pre-training offers robust, general-purpose feature extraction that requires only minimal fine-tuning in the target domain. In contrast, initializing from scratch with limited data increases the risk of overfitting and reduces generalization performance.

The simple ML baseline, while significantly smaller (in terms of memory and latency) than the deep models, did not perform as well in raw classification. The optimal ML combination, LBP features with Random Forest, can only achieve 98.00% accuracy and a Kappa of 0.9742, which is high given that these are handcrafted features and still a very much operating point for ultra-low-power devices, where even MobileNetV2 would be overkill. Nevertheless, these contributions, which create a 1.8-percentage-point gap compared to DenseNet169, and the lack of end-to-end differentiability or built-in interpretability tools make classical pipelines less well-suited when optimal diagnostic reliability is a concern.

For pragmatic deployment, the balance between accuracy and efficiency is essential in realistic agricultural scenarios. DenseNet169 still takes about 110 ms per image on an RTX 2060S, but it can process around 9 images per second, which is enough to perform post-hoc analysis of images collected by drones or mounted cameras. For real-time, MobileNetV2 with 77 ms inference time and <1 GFLOPs model size seems the most balanced solution, while not compromising diagnostic reliability at <99% accuracy. Table 6 compares the proposed work with state-of-the-art.

Table 6. Illustrated the comparison between the state-of-the-art.

Authors	Model	Accuracy %
18	Hybrid CNN	99.41
19	YOLOv4 + DenseNet	94.7
20	CNN based	97.49
21	Modified AlexNet	95.75
22	DenseNet121	97.11
23	VGG16	92.00
24	Denset169	99.57
25	DarkNet	88.1
Proposed	TL DenseNet169	99.80

5. Conclusions

This study introduces a novel hybrid approach for diagnosing cherry leaf disease using the PlantCity database. The methodology evaluates transfer learning, full pre-training, and hand-crafted machine learning workflows. The transfer learning model DenseNet169 demonstrated the highest performance, achieving an accuracy of 99.80% and Cohen's Kappa of 99.74%. This model surpassed deeper architectures in the efficiency-accuracy trade-off. In contrast, full pre-training on DenseNet201 provided minimal improvement in generalization performance. Consequently, the LBP, RF, and machine learning branches achieved accuracy comparable to deep learning (98.00%), while offering enhanced speed and deployability, which are advantageous for resource-limited agricultural settings. Quantitative evaluation and visual analysis, including confusion curves, corroborated these findings; observed overfitting was minimal, and class-wise performance was elucidated. The modular structure of the hybrid design enhances adaptability, interpretability, and robustness, thereby advancing precision agriculture by facilitating early disease detection, minimizing crop loss, and reducing pesticide usage.

Future extensions include the exploration of multi-crop datasets for greater generalization, the use of multimodal data composed, for instance, of hyperspectral imaging and environmental sensors, and the development of advanced ensembles, such as dynamic branch weighting based on input complexity. To enhance practical application, the hybrid model could be deployed on edge devices using TensorFlow Lite and federated learning for farmer-contributed data. Ablation studies on augmentation methods and attention mechanisms to further improve performance could be an extension of this work.

The limitations are that our dataset's structure focuses only on cherry leaves, limiting cross-crop generalization, and that it lacks real-world factors. Besides, those computerized measurements have been tested on standard hardware and need to be verified for edge devices.

Acknowledgments:

The authors extend their appreciation to the Deanship of Scientific Research at Northern Border University, Arar, Kingdom of Saudi Arabia, for funding this research work through project number "NBU-FFR-2026-2467-03".

Funding:

This work is supported by the Deanship of the Scientific Research, Northern Border University, Arar, Saudi Arabia, under Grant number "NBU-FFR-2026-2467-03".

Data Availability Statement:

https://data.mendeley.com/datasets/w8kh2xkspx/3 or
https://www.kaggle.com/datasets/codewithsk/cherry-leaf-disease-on-plantcity-2025

Author Contributions:

Conceptualization, I.A, M.S.K. and K.N.; methodology, I.A, M.S.K.; software, K.A.; validation, K.A. and I.A.; formal analysis, K.A.; investigation, I.A.; data curation, M.S.K. and K.N.; writing—original draft preparation, I.A, M.S.K., and K.N.; writing—review and editing, K.A. and I.A.; visualization, M.S.K.; supervision, I.A.; funding acquisition, K.A.

References

[1] Jain, A., Sarsaiya, S., Wu, Q., Lu, Y., & Shi, J. (2019). A review of plant leaf fungal diseases and their environmental speciation. *Bioengineered*, *10*(1), 409-424.

[2] Shetty, R., Jensen, B., Shetty, N. P., Hansen, M., Hansen, C. W., Starkey, K. R., & Jørgensen, H. J. L. (2012). Silicon induced resistance against powdery mildew of roses caused by Podosphaera pannosa. *Plant pathology*, *61*(1), 120-131.

[3] Dönmez, E. (2020, October). Discrimination of haploid and diploid maize seeds based on deep features. In *2020 28th Signal Processing and Communications Applications Conference (SIU)* (pp. 1-4). IEEE.

[4] Jackulin, C., & Murugavalli, S. J. M. S. (2022). A comprehensive review on detection of plant disease using machine learning and deep learning approaches. *Measurement: Sensors*, *24*, 100441.

[5] Yan, Z., Zhan, Y., Peng, Z., Liao, S., Shinagawa, Y., Zhang, S. & Zhou, X. S. (2016). Multi-instance deep learning: Discover discriminative local anatomies for bodypart recognition. *IEEE transactions on medical imaging*, *35*(5), 1332-1343.

[6] Annabel, L. S. P., Annapoorani, T., & Deepalakshmi, P. (2019, April). Machine learning for plant leaf disease detection and classification–a review. In *2019 international conference on communication and signal processing (ICCSP)* (pp. 0538-0542). IEEE.

[7] Sharma, R., Singh, A., Jhanjhi, N. Z., Masud, M., Jaha, E. S., & Verma, S. (2022). Plant Disease Diagnosis and Image Classification Using Deep Learning. *Computers, Materials & Continua*, *71*(2).

[8] Rauf, H. T., Saleem, B. A., Lali, M. I. U., Khan, M. A., Sharif, M., & Bukhari, S. A. C. (2019). A citrus fruits and leaves dataset for detection and classification of citrus diseases through machine learning. *Data in brief*, *26*, 104340.

[9] Chowdhury, M. E., Rahman, T., Khandakar, A., Ayari, M. A., Khan, A. U., Khan, M. S., ... & Ali, S. H. M. (2021). Automatic and reliable leaf disease detection using deep learning techniques. *AgriEngineering*, *3*(2), 294-312.

[10] Ayikpa, K. J., Mamadou, D., Gouton, P., & Adou, K. J. (2022). Experimental evaluation of coffee leaf disease classification and recognition based on machine learning and deep learning algorithms. *Journal of Computer Science*, *18*(12), 1201-1212.

[11] Hassan, S. M., Amitab, K., Jasinski, M., Leonowicz, Z., Jasinska, E., Novak, T., & Maji, A. K. (2022). A survey on different plant diseases detection using machine learning techniques. *Electronics*, *11*(17), 2641.

[12] Zhang, K., Zhang, L., & Wu, Q. (2019). Identification of cherry leaf disease infected by Podosphaera pannosa via convolutional neural network. *International Journal of Agricultural and Environmental Information Systems (IJAEIS)*, *10*(2), 98-110.

[13] Atila, Ü., Uçar, M., Akyol, K., & Uçar, E. (2021). Plant leaf disease classification using EfficientNet deep learning model. *Ecological Informatics*, *61*, 101182.

[14] Jiang, P., Chen, Y., Liu, B., He, D., & Liang, C. (2019). Real-time detection of apple leaf diseases using deep learning approach based on improved convolutional neural networks. *Ieee Access*, *7*, 59069-59080.

[15] Wang, H., Li, G., Ma, Z., and Li, X. (2012). "Image recognition of plant diseases based on principal component analysis and neural networks," in Proceedings of the 8th International Conference on Natural Computation, Okinawa Prefecture, 246–251.

[16] De Goma, J. C., Quilas, C. A. M., Valerio, M. A. B., Young, J. J. P., & Sauli, Z. (2018). Fruit recognition using surface and geometric information. *Journal of Telecommunication, Electronic and Computer Engineering (JTEC)*, *10*(1-15), 39-42.

[17] Katarzyna, R., & Paweł, M. (2019). A vision-based method utilizing deep convolutional neural networks for fruit variety classification in uncertainty conditions of retail sales. *Applied Sciences*, *9*(19), 3971.

[18] Momeny, M., Jahanbakhshi, A., Jafarnezhad, K., & Zhang, Y. D. (2020). Accurate classification of cherry fruit using deep CNN based on hybrid pooling approach. *Postharvest Biology and Technology*, *166*, 111204.

[19] Gai, R., Chen, N., & Yuan, H. (2023). A detection algorithm for cherry fruits based on the improved YOLO-v4 model. *Neural Computing and Applications*, *35*(19), 13895-13906.

[20] Sethy, P. K., Barpanda, N. K., Rath, A. K., & Behera, S. K. (2020). Nitrogen deficiency prediction of rice crop based on convolutional neural network. *Journal of Ambient Intelligence and Humanized Computing*, *11*(11), 5703-5711.

[21] Gehlot, M., & Saini, M. L. (2020, December). Analysis of different CNN architectures for tomato leaf disease classification. In *2020 5th IEEE international conference on recent advances and innovations in engineering (ICRAIE)* (pp. 1-6). IEEE.

[22] Abbas, A., Jain, S., Gour, M., & Vankudothu, S. (2021). Tomato plant disease detection using transfer learning with C-GAN synthetic images. *Computers and electronics in agriculture*, *187*, 106279.

[23] Chen, J., Zhang, D., Nanehkaran, Y. A., & Li, D. (2020). Detection of rice plant diseases based on deep transfer learning. *Journal of the Science of Food and Agriculture*, *100*(7), 3246-3256.

[24] Kayaalp, K. (2024). A deep ensemble learning method for cherry classification. *European Food Research and Technology*, *250*(5), 1513-1528.

[25] Dönmez, E., Ünal, Y., & Kayhan, H. (2024). Bacterial Disease Detection of Cherry Plant Using Deep Features. *Sakarya University Journal of Computer and Information Sciences*, *7*(1), 1-10.

[26] Murugesan, S., Chinnadurai, J., Srinivasan, S., Mathivanan, S. K., Chandan, R. R., & Moorthy, U. (2025). Robust multiclass classification of crop leaf diseases using hybrid deep learning and Grad-CAM interpretability. *Scientific Reports*, *15*(1), 29955.

[27] Khan, M. S., Nisa, K., Ahmad, I., Zubair, M., & Alshammari, K. (2025). PlantCity: A Comprehensive Image Based on Multi-Crop Leaves in Pakistan. *Data in Brief*, 112130.

[28] Zhou, C., Zhou, S., Xing, J., & Song, J. (2021). Tomato leaf disease identification by a restructured deep residual dense network. *IEEE Access*, *9*, 28822-28831.

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

[30] Pan, S. J. (2020). Transfer learning. *Learning*, *21*, 1-2.

[31] Xu, W., Yu, G., Zare, A., Zurweller, B., Rowland, D. L., Reyes-Cabrera, J., ... & Juenger, T. E. (2020). Overcoming small minirhizotron datasets using transfer learning. *Computers and Electronics in Agriculture*, *175*, 105466.

[32] Kaya, A., Keceli, A. S., Catal, C., Yalic, H. Y., Temucin, H., & Tekinerdogan, B. (2019). Analysis of transfer learning for deep neural network based plant classification models. *Computers and electronics in agriculture*, *158*, 20-29.