



A Segmentation-driven Dual-branch Deep Learning Framework with Attention Fusion for Plant Disease Phenotyping

Syama Krishna S^{1*} A Rajesh¹

¹Department of Computer Science and Engineering, Vels Institute of Science,
Technology and Advanced Studies, Chennai, India

* Corresponding author's Email: syamaks545@gmail.com

Abstract: Plant phenotyping plays a critical role in understanding plant health and improving agricultural productivity by enabling quantitative analysis of disease-related physiological characteristics. Among these, leaf diseases significantly impact crop yield and quality, necessitating accurate and automated phenotyping approaches. Traditional phenotyping methods rely on manual inspection or handcrafted feature extraction, which are time-consuming, prone to human error, and lack scalability under diverse environmental conditions. This study introduces a deep learning (DL)-based approach for image-based plant phenotyping, focusing on the classification of disease-affected traits. The developed method integrates segmentation-based region extraction, dual-branch feature learning, and attention-based feature fusion. Initially, input images are preprocessed and passed through a TransUNet-based segmentation module to isolate phenotypically relevant leaf regions while suppressing background interference. Both the original image and the segmented region are then processed using a RegNet-based feature extraction network to capture global structural information and localized disease-specific characteristics. The extracted features are fused using an attention-based mechanism, followed by fully connected layers for multiclass classification. Experimental results obtained on the controlled PlantVillage grape leaf dataset, which serves as a standardized benchmark for plant disease classification, demonstrate an overall classification accuracy of 97.8%, with precision, recall, and F1-score values of 97.7%, 97.9%, and 97.8%, respectively. In addition, the segmentation module achieves an Intersection over Union (IoU) of 94.1% and a Dice score of 96.8%, confirming its effectiveness in isolating relevant phenotypic regions.

Keywords: Plant phenotyping, Leaf disease classification, TransUNet, Image segmentation, Dual branch feature learning, Region of interest, RegNet.

1. Introduction

Global food security and economic stability rely heavily on agriculture, particularly in developing nations where farming serves as the primary source of income for a significant portion of the population [1, 2]. With the growing global population, the demand for enhanced agricultural productivity and crop quality has increased, necessitating efficient crop monitoring systems. In this context, plant phenotyping, which involves the quantitative assessment of plant characteristics, has emerged as a critical approach for evaluating plant health and performance [3, 4]. Among various phenotypic attributes, physiological characteristics, particularly

disease symptoms, play a crucial role as they directly impact crop yield and quality if not identified at an early stage [5, 6].

Traditionally, disease-related phenotyping has relied on manual inspection by agricultural experts, which is labor-intensive, subjective, and prone to inconsistencies due to variations in expertise [7, 8]. Furthermore, continuous large-scale phenotypic monitoring is impractical under field conditions. Early studies explored image-based techniques that utilize visual features such as texture, shape, and color to detect disease-related traits [9]. However, these approaches are highly dependent on manual feature design and predefined rules, limiting their ability to handle complex phenotypic variations and diverse environmental conditions [10].

One approach to enhancing automated plant phenotyping is the integration of machine learning techniques that enable data-driven classification of disease-associated characteristics [11]. Although these methods improve accuracy compared to traditional approaches, they rely on manual feature engineering and are limited in capturing complex nonlinear patterns in plant leaf images [12]. Recent advances in deep learning (DL) have further improved image-based plant phenotyping by enabling end-to-end learning of hierarchical feature representations directly from raw images [13]. Despite these advancements, challenges remain in achieving accurate phenotypic analysis due to complex backgrounds, variations in illumination, and differences in disease expression across plant species [14].

Another important challenge is the integration of both global contextual information and localized disease-specific phenotypic characteristics. Global features capture overall structural information, whereas local features are essential for identifying fine-grained disease patterns. Relying on a single representation may result in incomplete characterization and reduced classification performance. In real-world scenarios, it is often difficult to isolate relevant phenotypic regions due to background noise, including soil, shadows, and surrounding vegetation. Although segmentation techniques have been explored, their effective integration with classification frameworks requires careful design to achieve meaningful improvements.

To address these challenges, this study introduces a deep learning-based plant phenotyping framework that integrates segmentation-driven region extraction, dual-branch feature learning, and attention-based feature fusion within a unified architecture. Unlike existing approaches that primarily rely on either whole-image analysis or segmented regions independently, the developed framework simultaneously exploits global contextual information from original leaf images and localized disease-specific characteristics from segmented regions. Furthermore, an adaptive attention-based fusion mechanism is incorporated to effectively combine complementary feature representations and enhance the discrimination of subtle disease symptoms. This integrated design improves the characterization of disease-related phenotypic traits while reducing the influence of irrelevant background information. Compared with conventional single-branch and whole-image-based classification approaches, the framework enables more effective utilization of both global and localized disease information, leading to enhanced feature

representation, improved robustness to phenotypic variability, and more reliable disease classification.

The primary contribution of this study lies in the unified integration of segmentation-guided region localization, parallel representation learning, and adaptive feature aggregation within a single phenotyping framework. While existing plant disease classification methods commonly employ segmentation, attention mechanisms, transformer architectures, or dual-branch learning as separate enhancements, the proposed framework establishes a coordinated learning strategy in which disease-relevant regions are first localized and subsequently analyzed alongside the corresponding full-image representations. This enables the model to preserve contextual information while simultaneously emphasizing disease-discriminative characteristics. The resulting architecture provides a distinct formulation for phenotypic feature learning and offers a clearer interaction between region-focused analysis and global structural understanding than conventional single-stream, segmentation-assisted, or attention-guided classification frameworks. The key objectives of this study are:

- To develop an efficient plant phenotyping framework for disease-related physiological trait classification using controlled grape leaf imagery, while providing a foundation for future adaptation to more complex agricultural environments.
- To enhance feature representation by integrating segmentation-based region extraction with dual branch learning of global and localized phenotypic characteristics.
- To implement an adaptive attention-based fusion mechanism for effective feature integration while minimizing background noise influence.
- To evaluate the proposed framework using standard performance metrics.

The remainder of this paper is organized as follows. Section 2 reviews the existing literature on plant disease detection and highlights research gaps. Section 3 presents the proposed methodology, including preprocessing, segmentation, feature extraction, and classification. Section 4 presents the results and comparative analysis. Finally, Section 5 concludes the paper and outlines future research directions.

2. Literature review

Using the Turkey Plant Pests and Diseases dataset comprising 4,447 images across 15 classes, Shafik et

al. [15] developed a ResNet9-based model for multiclass classification of plant pests and diseases. The images were captured using a Nikon 7200D camera and resized to 300×300 pixels. To address class imbalance, extensive data augmentation and hyperparameter optimization were employed. The model achieved 96.4% precision, 97.4% accuracy, 95.7% F1-score, and 97.09% recall, with an AUROC of 97.1%, demonstrating statistically significant improvement over EfficientNet ($t = 23.701$, $p < 0.00001$). SHAP-based saliency maps provided pixel-level interpretability. Misclassifications occurred among visually similar classes, such as Plum Aphis spp. and Walnut Gnomonia leptostyla. Limitations included class imbalance, sensitivity to environmental variations, and reliance on single-modality image data.

Evwiekpaefe et al. [16] conducted a comparative study on binary and multiclass classification of potato leaf diseases using a VGG19-based CNN. Images of healthy, virus-infected, early blight, and late blight leaves were collected from farms in Kaduna using a Sony W610 camera at a distance of 0.3 m. In the binary classification setup, the model achieved training, validation, and testing accuracies of 93.25%, 90.00%, and 91.67%, respectively. For multiclass classification, the corresponding accuracies were 91.28%, 87.50%, and 91.67%. Loss values were higher in the multiclass setting during both training and testing phases. A paired t-test yielded $t = 7.4054$ and $p = 4.08 \times 10^{-5}$. The study highlighted limitations related to dataset size and increased class complexity, which adversely affected generalization performance.

Shafik et al. [17] developed PDDNet models using early average fusion (PDDNet-AE) and lead voting ensemble (PDDNet-LVE) by integrating nine pre-trained CNNs for plant leaf disease classification. Deep features were extracted from 15 PlantVillage classes comprising over 54,000 images, followed by classification using logistic regression with ten-fold cross-validation. Fine-tuned DenseNet201 achieved an accuracy of 93.48%, which improved to 94.86% after modifying the final layers. The ensemble approaches achieved accuracies of 96.94% for PDDNet-AE and 97.79% for PDDNet-LVE, with corresponding F1-scores of 95.02% and 97.07%, respectively. Despite improved generalization, reliance on controlled PlantVillage data and high computational cost limit real-world applicability and robustness.

Using 11,000 images from the PlantVillage dataset, Najim et al. [18] developed a lightweight CNN for early recognition of tomato leaf diseases across ten classes. The images were resized from 256×256 to 128×128 pixels to reduce computational

cost, and no noise removal was applied due to the clean nature of the dataset. The four-layer CNN was trained for 40 epochs, achieving 96% training accuracy with a loss of 0.1. The test accuracy reached 92%, while precision, F1-score, and recall were 91%, 92%, and 92%, respectively. Compared to LeNet (94%), ResNet50 (96%), and other CNN models, the architecture was compact (23 MB), resulting in improved computational efficiency. However, reliance on noise-free images limits robustness in real-world scenarios.

Bouacida et al. [19] proposed a generalized approach for plant disease detection using a lightweight Inception-based CNN trained on segmented PlantVillage images comprising 54,305 samples across healthy and diseased categories. The method subdivides images into smaller patches to focus on infected regions, rather than classifying entire leaves, enabling more detailed disease severity assessment. The model achieved a training accuracy of 94.66% after 300 epochs and a testing accuracy of 94.04%. On the test set, recall, precision, and F1-score were 93.95%, 95.05%, and 94.50%, respectively. Evaluation on the PDDb dataset, containing unseen crops and diseases, yielded an accuracy of 97.22%. Misclassifications were primarily attributed to the similarity between healthy light-green regions and mild disease symptoms, difficulty in detecting very small lesions, and dataset imbalance.

Naveenkumar and Nandagopal [20] investigated automated plant leaf disease classification to improve early detection and reduce crop losses. The study enhances accuracy and efficiency using AI-driven techniques on the PlantifyDr dataset. Noise is removed using median filtering, and segmentation is performed using the adaptive hybrid K-means with fuzzy C-means (AHKM-FCM) algorithm, optimized via the IRV-WSA algorithm. For classification, a transfer learning-based network combining EfficientNet, ResNet, and DenseNet with optimized weights is employed. The IRV-WSA-ETLNet model achieves 94.853% accuracy, 94.750% sensitivity, 96.068% F1-score, and 94.888% specificity, with a computation time of 24.378 ms. The results demonstrate improved segmentation and classification performance compared to existing approaches.

To improve accuracy and reduce manual effort, Kaur and Devendran [21] developed a semi-automated framework for grape leaf disease diagnosis using the PlantVillage dataset. Image segmentation was performed using K-means clustering optimized with Grey Wolf Optimization (GWO).

Feature extraction combined Local Binary Pattern (LBP), Law's mask, Gabor features, and Gray-Level Co-occurrence Matrix (GLCM) to generate a hybrid feature representation. Classification was carried out using an ensemble of multiple machine learning classifiers. The system achieved an accuracy of 95.69%. However, reliance on handcrafted feature extraction and limited generalization hinders its effectiveness in adapting to diverse disease conditions and varying environmental factors.

To improve decision-making and reduce crop losses, Khan et al. [22] proposed an intelligent system for grape disease diagnosis. The study focused on identifying powdery mildew, blotches, and healthy leaves using image data captured from strategically positioned farm cameras. The data were preprocessed using augmentation techniques and Histogram of Oriented Gradients (HOG) for feature extraction. The CNN model achieved satisfactory performance, while transfer learning and fine-tuning further improved classification accuracy. However, the approach exhibited limitations, including sensitivity to lighting conditions, weather variations, and image quality, as well as challenges in scalability and high implementation and maintenance costs for small-scale farmers.

To address challenges associated with limited datasets, Ishengoma and Lyimo [23] developed an ensemble model for grape leaf disease detection. The study utilized the PlantVillage grape leaf dataset containing approximately 4,040 images across four classes: Black Rot, Healthy, Esca, and Leaf Blight. Preprocessing involved background removal, intensity averaging, bilateral filtering, and data augmentation. Feature extraction was performed using multiple CNN architectures, including VGG16, InceptionV3, Xception, and ResNet50, while classification was carried out using a Random Forest classifier in a parallel ensemble framework. The VGG16–Xception model achieved an accuracy of 95.34%, with performance improvements of up to 5.6% on modified images. However, the model exhibited class confusion among visually similar disease categories, which negatively affected classification performance.

Yücalar and Yıldırım [24] investigated grape leaf disease detection to provide rapid and reliable diagnosis for improved agricultural productivity. The study utilized a hybrid dataset comprising publicly available images and manually collected field images captured using a smartphone under natural conditions. Transfer learning-based classification was performed using multiple CNN architectures. Among these, Inception-ResNet-V2 achieved training and testing

accuracies of 97.45% and 93.00%, respectively. However, the approach exhibited limitations, including class imbalance, variations in image quality, and high computational requirements, which increased training time and affected classification consistency across certain disease categories.

Li et al. [25] developed a Transformer-based multimodal fusion framework for grape disease classification by integrating RGB images, hyperspectral data, and environmental sensor information collected under diverse vineyard conditions. The framework employed Transformer-based feature learning, temporal enhancement, adaptive attention, and multi-source data fusion to capture spatial, spectral, and environmental dependencies associated with grape diseases. Experimental evaluation was conducted on a dataset containing 3,676 RGB images and 6,297 hyperspectral samples representing five grape disease categories. The model achieved the highest performance among the compared object detection approaches, attaining a precision of 0.93, recall of 0.95, accuracy of 0.94, mAP@50 of 0.94, mAP@75 of 0.93, and an F1-score of 0.94. The study reported sensitivity to variations in acquisition devices and environmental conditions, along with challenges in identifying subtle symptoms during the early stages of disease development.

Unal [26] investigated automated grapevine leaf disease diagnosis to improve the identification of grape leaf blister mite, downy mildew, and healthy leaves using deep learning techniques. The study initially evaluated several pre-trained convolutional neural networks, including VGG16, VGG19, InceptionV3, Xception, and DenseNet121. Attention mechanisms based on Convolutional Block Attention Module (CBAM) and Squeeze-and-Excitation Networks (SENet) were subsequently integrated with DenseNet121 to enhance feature extraction and classification performance. Experimental results showed that DenseNet121 achieved a validation accuracy of 92.73%, which increased to 96.36% after integrating CBAM and SENet. The enhanced model also attained a precision of 0.96, recall of 0.96, and F1-score of 0.96. Limitations included a region-specific dataset, class imbalance, limited sample size, and variations in vineyard lighting conditions.

Despite recent advancements in plant disease detection, existing systems exhibit several limitations. Many approaches employ single-branch CNN architectures that do not explicitly isolate diseased regions and lack sensitivity to subtle phenotypic variations [19, 23]. Segmentation-based methods often rely on clustering or handcrafted features, limiting adaptability under varying environmental

conditions [20, 21]. Furthermore, most models fail to effectively integrate both global and localized features, which negatively impacts classification performance [15, 18]. Although ensemble and transfer learning approaches improve accuracy, they introduce increased computational complexity and scalability challenges [17, 22]. Therefore, there is a need for a robust framework that integrates effective segmentation, dual-branch feature learning, and adaptive feature fusion.

To further position the proposed framework relative to existing plant disease classification architectures, Table 1 summarizes the key architectural characteristics of representative model categories. The comparison highlights the extent to which existing approaches incorporate segmentation, dual-branch learning, attention mechanisms, transformer architectures, global-local feature learning, and adaptive feature fusion. Most existing approaches employ only a subset of these capabilities, whereas the proposed framework integrates all components within a unified phenotyping architecture.

It can be observed that existing CNN, transfer learning, segmentation-guided, attention-assisted, and transformer-based approaches generally focus on specific aspects of feature learning. In contrast, the proposed framework combines segmentation-guided ROI extraction, dual-branch representation learning, transformer-enhanced feature modeling, and adaptive attention-based fusion within a single architecture. This integration enables simultaneous exploitation of global contextual information and localized disease-specific characteristics, thereby providing a more comprehensive phenotypic feature learning strategy.

3. Materials and methods

The framework illustrated in Fig. 1 presents a systematic pipeline for plant disease phenotyping

aimed at accurate classification of physiological traits. Initially, input images are preprocessed to enhance reliability and generalization. The processed images are then passed through a segmentation module, which generates a binary mask to isolate the phenotypic region of interest (infected leaf areas) while suppressing background noise. Both the original image and the segmented region are subsequently fed as parallel inputs into a dual-branch feature learning model. One branch captures global contextual information of the plant phenotype, whereas the other focuses on localized disease-specific characteristics of the segmented region. The features extracted from both branches are integrated using an adaptive attention-based fusion mechanism, which effectively balances their contributions and enhances discriminative capability. Finally, the fused features are passed through a classification head to predict the corresponding disease category, enabling robust and accurate plant disease analysis.

3.1 Dataset description

The data used in this study are derived from the PlantVillage dataset, which is widely utilized for image-based plant phenotyping and disease analysis [27]. The complete dataset comprises 54,304 RGB images across 38 classes representing 14 crop species. In this work, a subset of grape leaf images is selected to enable a controlled and detailed analysis. The selected subset contains 4,063 images belonging to four classes: Esca (Black Measles), Black Rot, Leaf Blight (Isariopsis Leaf Spot), and Healthy. These disease categories are primarily associated with fungal infections and exhibit diverse visual characteristics, including necrotic lesions, discoloration, and irregular spatial patterns. Table 2 presents the class-wise distribution of the grape leaf dataset along with the corresponding disease types, causal pathogens, and sample counts.

Table 1. Architectural comparison of representative plant disease classification frameworks

Model Category	Segmentation	Dual Branch	Attention	Transformer	Global + Local Learning	Adaptive Fusion
CNN Models [18]	No	No	No	No	Partial	No
Transfer Learning Models [24]	No	No	No	No	Partial	No
Segmentation-Based Models [20,21]	Yes	No	No	No	Limited	No
Attention-Based Models [15]	No	No	Yes	No	Partial	Limited
Transformer Models	No	No	Self-Attention	Yes	Partial	No
Existing Dual-Branch Models	Partial	Yes	Limited	No	Yes	Limited
Proposed Framework	Yes	Yes	Yes	Yes	Yes	Yes

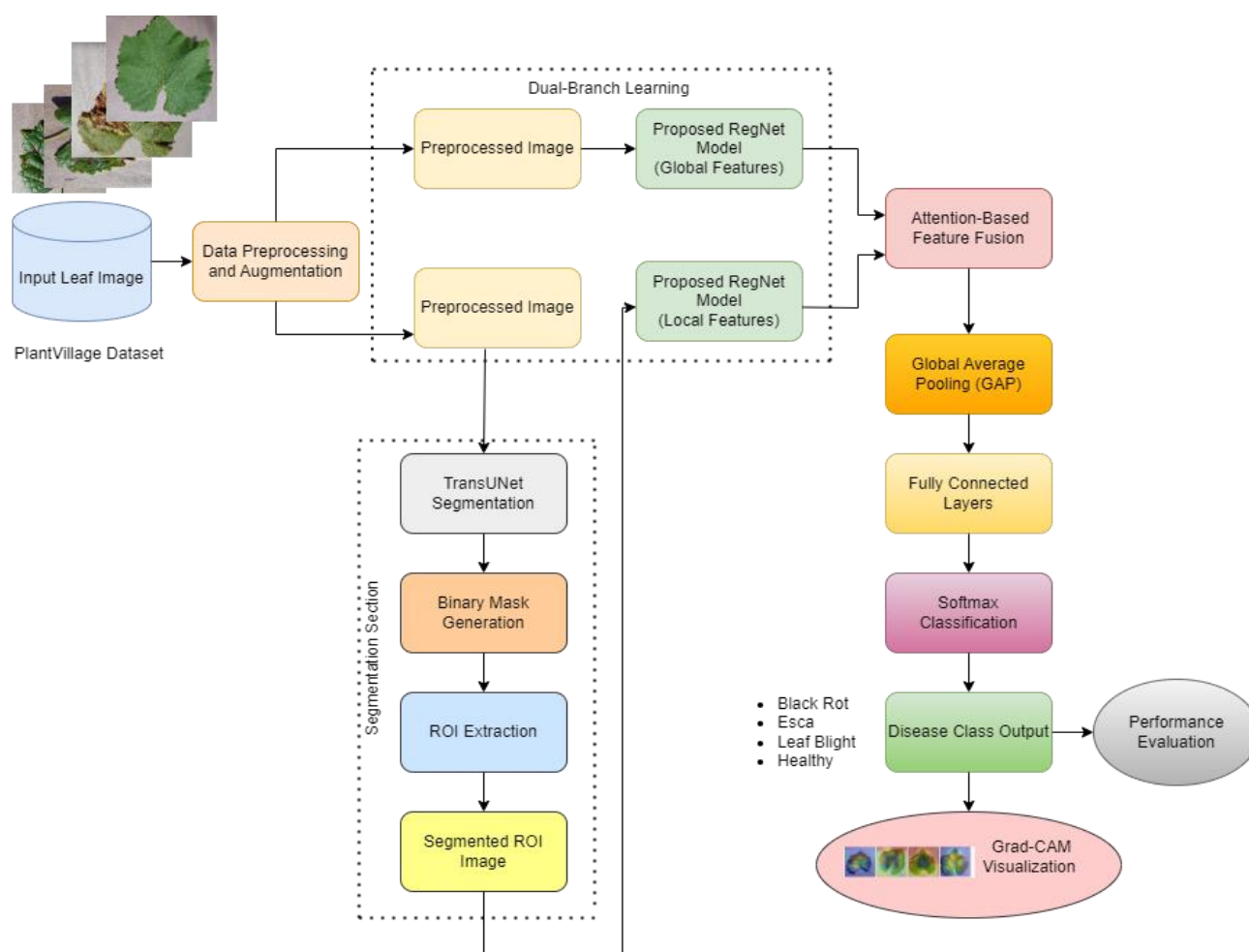


Figure. 1 Block diagram of the plant phenotyping framework for disease related trait classification

Table 2. Description of grape leaf dataset used for phenotyping analysis

Class Name	Cause of Disease	Pathogen	Number of Samples
Black Rot	Fungus	Guignardia bidwellii	1180
Esca	Fungus	Phaeomoniella spp.	1384
Leaf Blight	Fungus	Pseudocercospora vitis	1076
Healthy	-	-	423
Total	-	-	4063

The corresponding causal organisms for each disease include *Guignardia bidwellii* (Black Rot), *Phaeomoniella* spp. (Esca), and *Pseudocercospora vitis* (Leaf Blight). These diseases were selected due to their high intra-class variability and visually distinct yet partially overlapping symptom patterns, making them suitable for evaluating segmentation-based and multi-scale phenotyping approaches. All images were originally collected under controlled experimental conditions at agricultural research stations, including Penn State University and Cornell University. Leaves were manually detached and photographed on uniform backgrounds (gray or

black) under natural lighting conditions. Multiple images (typically 4–7 per sample) were captured at different angles to account for variations in phenotypic expression. The images were subsequently preprocessed to remove background noise and standardize leaf positioning.

Each image is annotated with its corresponding disease or healthy class label, verified by expert plant pathologists to ensure labeling accuracy.

The dataset is accessed through TensorFlow Datasets (`tlds.datasets.plant_village`) and consists of uniformly formatted RGB images suitable for deep learning models. Fig. 2 illustrates representative samples from the dataset.

3.2 Data preprocessing and augmentation

Image-based plant phenotyping relies on data preprocessing to ensure input consistency while preserving disease-related physiological characteristics essential for accurate classification. The preprocessing pipeline includes image resizing and intensity normalization to standardize inputs and

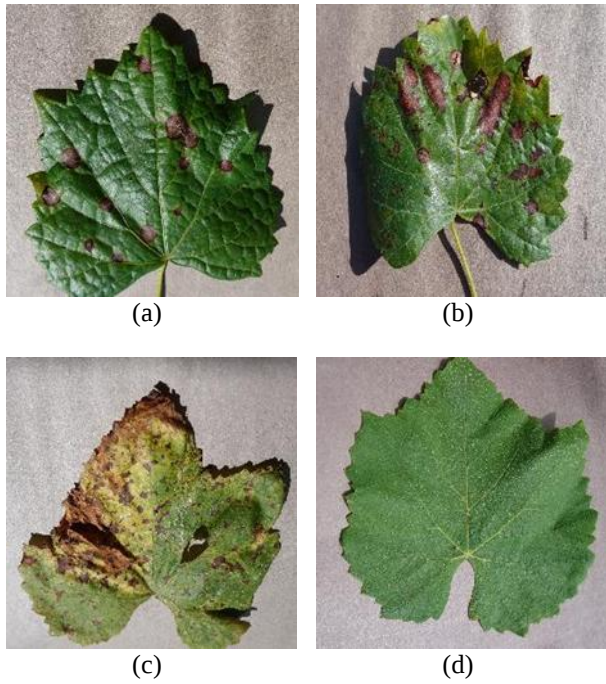


Figure. 2 Representative grape leaf images illustrating disease related phenotypic variations: (a) Black Rot, (b) Esca, (c) Leaf Blight and (d) Healthy

enhance robustness to variations in scale, orientation, and illumination. An input image is represented as $I \in \mathbb{R}^{H \times W \times 3}$. The resizing operation transforms it into a standardized image $I' \in \mathbb{R}^{h \times w \times 3}$, as given in Eq. (1).

$$I' = R(I) \quad (1)$$

where $R(\cdot)$ denotes the resizing function and h and w symbolize the target height and width. After scaling, the pixel intensity normalization is utilised to bring the image values to a constant range to stabilize gradient updates and enhance convergence during training. Eq. (2) calculates the normalized image.

$$I_n = \frac{I'}{255} \quad (2)$$

To further normalize the data distribution, standard score normalization is applied, as in Eq. (3).

$$I_n = \frac{I' - \mu}{\sigma} \quad (3)$$

where μ and σ denote the mean and standard deviation of pixel intensities across the dataset. The generalization is made more effective with the help of data augmentation strategies and the overfitting is minimized. These changes mimic the changes that are usually found in the phenotypes of plants such as

differences in orientation, scale and viewpoint. The augmented image is given by Eq. (4).

$$I_a = T(I_n) \quad (4)$$

where $T(\cdot)$ represents a composition of augmentation operations. The transformations include rotation, flipping and scaling, provided as Eq. (5).

$$I_r = R_\theta(I_n), I_f = F(I_n), I_s = S(I_n) \quad (5)$$

where θ denotes the rotation angle and F and S represent flipping and scaling operations, respectively. 80% of the images correspond to training while the remaining 20% (813 images) were reserved for testing, and data augmentation was applied exclusively to the training set while the test set was kept unchanged to ensure unbiased evaluation. To maintain a rigorous experimental protocol, the train-test partition was established prior to model development and consistently preserved throughout all subsequent segmentation and classification stages. Images assigned to the test partition were excluded from all training procedures and were used exclusively for final performance evaluation. The same partitioning strategy was employed for both segmentation and classification tasks, thereby preventing information leakage between training and testing samples and ensuring an unbiased assessment of model generalization capability.

The preprocessing stage preserves key phenotypic traits, including lesion structures, discoloration patterns, and texture variations, which are essential for disease analysis. The images are normalized and augmented to facilitate effective feature learning. To address class imbalance, augmentation is applied more extensively to underrepresented classes, particularly the Healthy category. The augmentation operations increase data diversity by generating transformed image variants that simulate variations in orientation, scale, and viewpoint, thereby improving model robustness and generalization capability. The augmentation is performed dynamically during training (on-the-fly augmentation), and therefore the original dataset size and class counts remain unchanged. Consequently, no fixed post-augmentation class counts are generated, as transformed image variants are created only during model training. No undersampling was performed during dataset preparation, and all original images were retained throughout the training process. The augmentation strategy was primarily intended to reduce the effect of class imbalance and improve feature learning rather than enforce an exactly equal

class distribution. The preprocessing and augmentation results are illustrated in Fig. 3.

3.3 Image segmentation using transunet

A DL-based segmentation approach utilizing the TransUNet architecture is employed to accurately separate leaf regions from the background and enhance the extraction of phenotypic features. The segmentation module was employed to generate disease-relevant regions of interest (ROIs) from the input leaf images prior to feature extraction and

classification. ROI generation was performed independently within the predefined training and testing partitions established during dataset preparation. Consequently, test images were not utilized during segmentation model training, parameter optimization, or feature learning. The generated segmentation masks were subsequently used to isolate disease-relevant leaf regions while suppressing background information before being forwarded to the dual-branch feature extraction stage.

Fig. 4 illustrates the TransUNet-based segmentation framework.

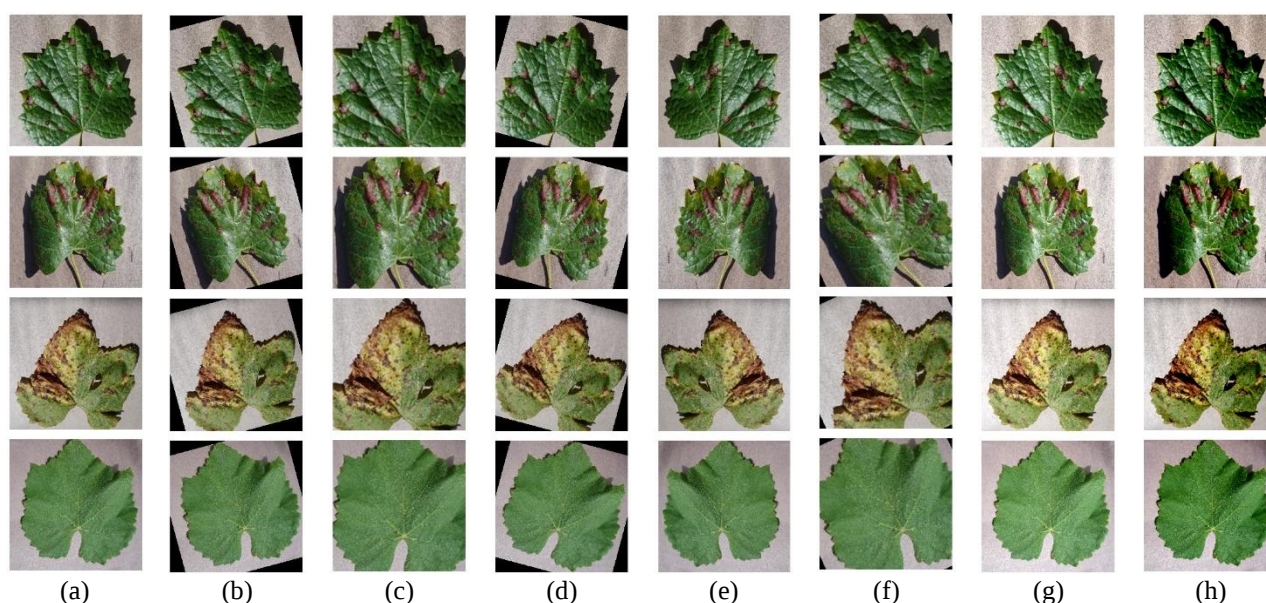


Figure. 3 Data preprocessing and augmentation results: (a) original image, (b) rotation R_θ , (c) scaling S , (d) inverse rotation, (e) flipping F , (f) rotation with cropping, (g) intensity variation, and (h) contrast normalization

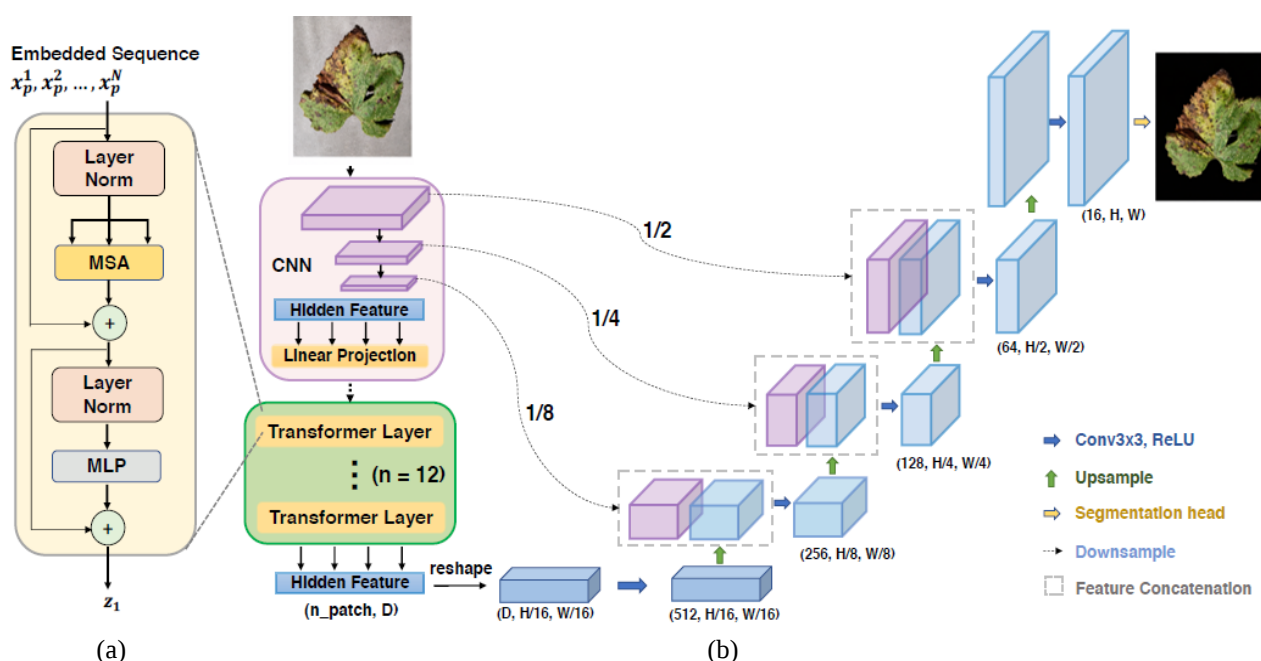


Figure. 4 Architecture of the TransUNet segmentation framework

The model integrates convolutional neural networks (CNNs) with transformer-based self-attention mechanisms to capture both local spatial features and long-range contextual dependencies, enabling precise segmentation of plant leaves [28].

The input image is represented as $I \in \mathbb{R}^{H \times W \times C}$, where H , W and C denote the height, width and number of channels respectively. All input images are resized to $224 \times 224 \times 3$ prior to being fed into the segmentation network. The encoder consists of convolutional layers based on a ResNet50 backbone, which extracts low-level and intermediate features such as textures, edges, and disease-related lesion patterns. The resulting feature maps are partitioned into 16×16 non-overlapping patches and linearly projected into a sequence of tokens. These tokens are processed by a transformer encoder comprising 12 layers, each with 12 attention heads and an embedding dimension of 768. The self-attention mechanism is defined in Eq. (6).

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V \quad (6)$$

where Q , K and V denote the query, key and value matrices and $d = 64$ denotes the dimensionality of each attention head.

The decoder progressively restores the spatial resolution of the feature maps through consecutive upsampling using bilinear interpolation. To preserve fine-grained spatial information and improve boundary localization, skip connections are incorporated between corresponding encoder and decoder layers. The output of the segmentation network is a pixel-wise probability map, $S \in \mathbb{R}^{224 \times 224}$. A binary segmentation mask is obtained by applying a threshold value $\tau = 0.5$, as given in Eq. (7).

$$M(x, y) = \begin{cases} 1, & \text{if } S(x, y) \geq 0.5 \\ 0, & \text{otherwise} \end{cases} \quad (7)$$

The element-wise multiplication method is used to extract the region of interest (ROI) which corresponds to the leaf area, as illustrated in Eq. (8).

$$I_{\text{ROI}} = I \odot M \quad (8)$$

The segmentation network is trained using a composite loss function that combines Dice loss and binary cross-entropy (BCE) loss, as defined in Eq. (9). The segmentation results for grape leaf images are illustrated in Fig. 5.

To assess the reliability of the pseudo-generated segmentation masks, a subset of 300 images was

manually annotated by domain experts. These annotations were used as reference ground truth for training and evaluation. The manually annotated images were selected from the training partition and subjected to visual verification to ensure annotation consistency and segmentation quality.

The 300 annotated masks were used exclusively for segmentation model development and were divided into training and validation subsets. Segmentation performance was subsequently evaluated using the held-out test partition established during dataset preparation. No test images were used during model training, validation, parameter optimization, or feature learning. Consequently, the reported IoU and Dice scores were obtained from images that were entirely independent of the segmentation training process, thereby preventing data leakage and ensuring an unbiased evaluation of segmentation performance.

$$L = 0.5 L_{\text{BCE}} + 0.5 L_{\text{Dice}} \quad (9)$$

The model is optimized using the Adam optimizer with a batch size of 16 and a learning rate of 1×10^{-4} . To preserve evaluation integrity, segmentation outputs generated from test samples were used solely during inference and were not incorporated into any training, validation, or optimization stage of the framework. The same train-test partition established during dataset preparation was consistently maintained throughout both segmentation and classification stages, thereby preventing information leakage between training and testing samples. The segmentation approach effectively isolates phenotypically relevant regions by suppressing background noise while preserving disease-specific characteristics such as lesion distribution, discoloration, and texture variations. Although the dataset consists of images captured under controlled conditions, segmentation is employed to emphasize relevant features and simulate region-based learning, thereby improving generalization to more complex real-world environments.

3.4 Model development

3.4.1. RegNet architecture

RegNet is a convolutional neural network (CNN) architecture designed using a structured network design space that systematically defines network width and depth through parameterized functions [29].

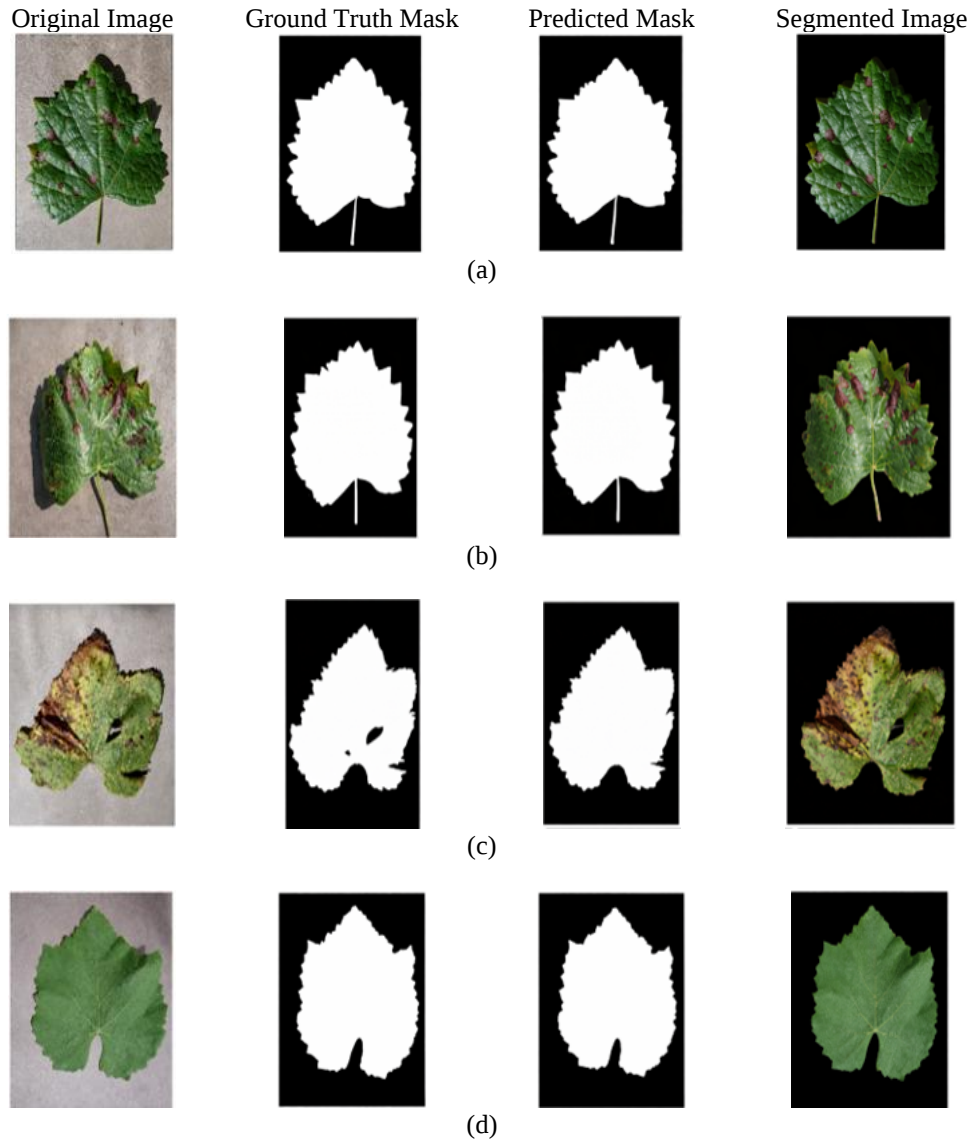


Figure. 5 Segmentation results of grape leaf images: (a) Black Rot, (b) Esca, (c) Leaf Blight and (d) Healthy leaf

The architecture follows a stage-wise configuration, as illustrated in Fig. 6, enabling scalable model design while maintaining computational efficiency. It employs a hierarchical feature representation mechanism to effectively capture multi-level features from input images. In this framework, RegNet is utilized for feature extraction, where the channel width is progressively increased across network blocks in accordance with Eq. (10).

$$w_i = w_0 + k \cdot i \quad (10)$$

where w_i denotes the width of the i^{th} block, w_0 represents the initial width and k controls the rate of width increment. The computed widths are quantized to the nearest multiple of a predefined factor q , ensuring hardware-efficient implementation and a regularized network structure. The input image is represented as $I \in \mathbb{R}^{224 \times 224 \times 3}$. RegNet consists of

multiple sequential stages, each comprising residual bottleneck blocks that perform hierarchical feature transformation. Each block includes a series of convolutional operations with 1×1 , 3×3 , and 1×1 kernels, followed by nonlinear activation and batch normalization (BN). The transformation within a block is defined in Eq. (11).

$$F_{\text{out}} = \sigma(\text{BN}(\text{Conv}_{1 \times 1}(\text{BN}(\text{Conv}_{3 \times 3}(\text{BN}(\text{Conv}_{1 \times 1}(F_{\text{in}})))))) \quad (11)$$

Residual connections combine the input and output feature maps as in Eq. (12).

$$F_{\text{res}} = F_{\text{out}} + F_{\text{in}} \quad (12)$$

This formulation ensures stable gradient propagation and conserves the feature information through the layers.

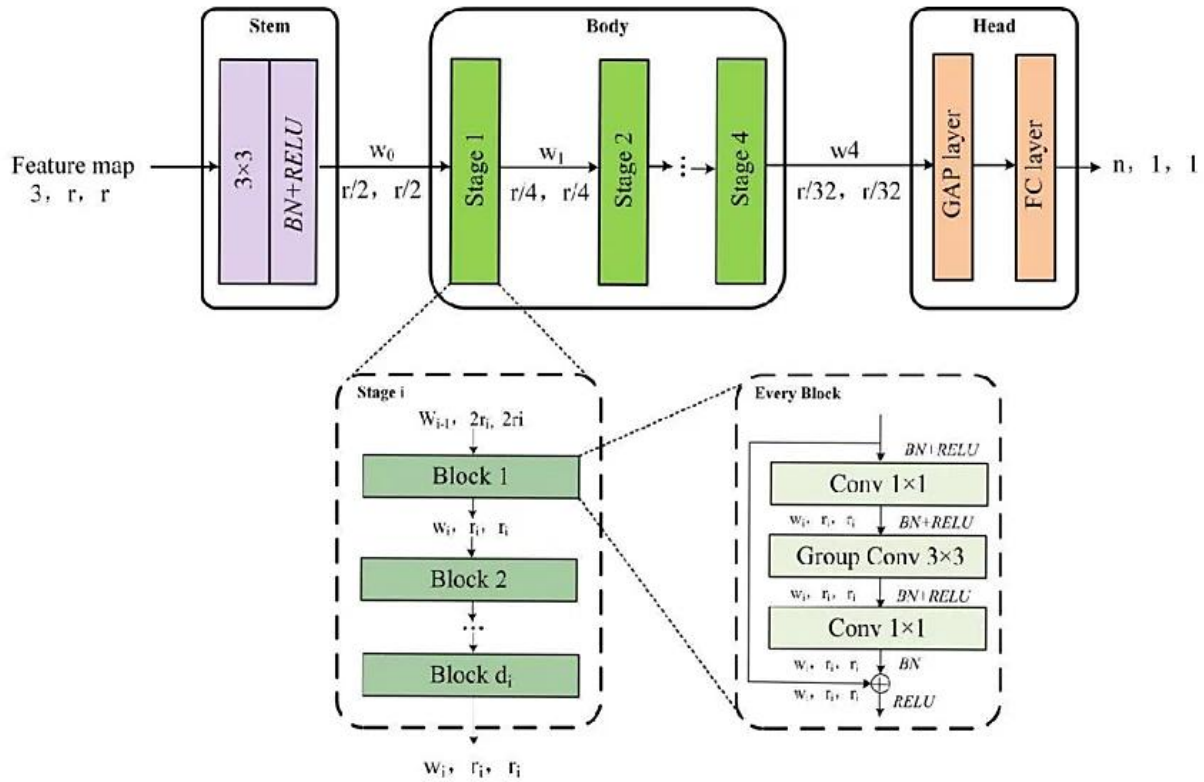


Figure. 6 RegNet architecture

The spatial resolution of feature maps is increasingly reduced through strided convolution operations, transitioning from 224×224 to 112×112 , 56×56 , 28×28 , 14×14 and finally 7×7 . The overall feature extraction process is represented as Eq. (13).

$$F = f_{\theta}(I), \quad F \in \mathbb{R}^{7 \times 7 \times d} \quad (13)$$

where $f_{\theta}(\cdot)$ denotes the RegNet mapping function parameterized by θ and d represents the number of output channels in the final stage. Spatial information is aggregated through global average pooling (GAP) to generate a compact feature representation, as in Eq. (14).

$$z = \text{GAP}(F) \quad (14)$$

The high-level feature vector generated in this stage captures the structural and disease-specific characteristics of the input image. These features encode variations in texture, lesion patterns, and overall leaf structure, which are critical for accurate disease analysis and subsequent classification.

3.4.2. Proposed deep learning framework

The proposed framework integrates segmentation, dual-branch feature extraction, attention-driven fusion, and classification into a unified architecture

for plant disease phenotyping. The design aims to simultaneously capture global structural information and localized disease-specific phenotypic characteristics. The model takes RGB plant leaf images as input, which are resized to a fixed resolution and normalized to ensure consistent representation. Data augmentation strategies are employed to improve generalization and enhance robustness to variations in orientation and scale.

The preprocessed images are first passed through a segmentation module based on the TransUNet architecture to identify phenotypically relevant regions. The segmentation network comprises a convolutional encoder with transformer-based self-attention layers and a decoder with skip connections. The encoder captures hierarchical spatial features, while the transformer layers model long-range dependencies within the image. The decoder reconstructs the spatial resolution to generate a pixel-wise probability map. A threshold value of $\tau = 0.5$ is applied to obtain a binary mask, which is used to extract the region of interest (ROI). This process effectively suppresses background elements such as soil, shadows, and surrounding vegetation while preserving disease-specific phenotypic features.

Following segmentation, a dual-branch feature learning strategy is employed using two parallel inputs: the original image and the segmented ROI.

Each input is processed independently using a RegNet backbone with structured width scaling to extract hierarchical feature representations. The initial layers capture low-level features such as edges, textures, and color variations, while deeper layers encode high-level semantic representations of disease characteristics. The outputs from both branches are feature maps with reduced spatial resolution and increased channel depth.

An attention-based fusion mechanism is then applied to integrate these complementary representations. The fusion module adaptively assigns weights to the feature maps from each branch, enabling the model to emphasize global contextual information from the original image and localized disease-specific features from the segmented region. This adaptive integration enhances the discriminative capability of the model by effectively combining structural and lesion-specific information.

The fused feature map is subsequently processed using a global average pooling (GAP) layer to obtain a compact feature vector.

This vector is passed through fully connected (FC) layers with 512 and 128 neurons, respectively, followed by ReLU activation.

A dropout layer with a rate of 0.5 is introduced between the FC layers to mitigate overfitting. The final classification layer employs a softmax activation function to generate class probabilities, and the class with the highest probability is selected as the predicted output.

The proposed framework effectively integrates global and localized feature representations while minimizing background interference. By explicitly modeling disease-related phenotypic characteristics through segmentation and dual-branch learning, the framework improves robustness and classification performance under varying conditions. Algorithm 1 outlines the complete model.

3.5 Experimental Setup

A high-performance computing workstation was used to conduct all experiments, configured to support DL-based image analysis. The system was equipped with an NVIDIA GPU with 8 GB of dedicated memory, an Intel Core i7 processor, and 32 GB of RAM. GPU acceleration significantly improved computational efficiency during training, particularly for segmentation and feature extraction stages involving extensive convolutional and attention-based operations.

The proposed framework was implemented using the Python programming language. Deep learning components, including the segmentation model and

feature extraction network, were developed using TensorFlow and Keras, which provide efficient tools for building and training complex neural architectures. Image processing tasks such as resizing, normalization, augmentation, and region extraction were performed using OpenCV, while NumPy was used for numerical computations and data manipulation.

The hyperparameters used to regulate the learning process are summarized in Table 3. The TransUNet-based segmentation module and the RegNet-based feature extraction module were trained using the Adam optimizer with mini-batch gradient descent to ensure stable convergence. Multiclass classification was performed using categorical cross-entropy loss. All experiments were conducted in a 64-bit environment, with reproducibility ensured through a fixed dataset split and controlled random initialization.

Algorithm 1: Segmentation-Driven Dual-Branch Deep Learning Framework with Attention Fusion for Plant Disease Phenotyping

Input: Plant leaf image dataset (RGB images)

Output: Predicted Plant Disease Class

Begin

Step 1: Data Preprocessing

- Resize input image I to fixed size $224 \times 224 \times 3$
- Normalize pixel intensities using Min-Max scaling:

$$I' = \frac{I - I_{\min}}{I_{\max} - I_{\min}}$$

- Apply standard score normalization:

$$I'' = \frac{I' - \mu}{\sigma}$$

- Perform data augmentation:
 - Rotation, flipping, scaling, cropping
- Split dataset into training (80%) and testing (20%)

Step 2: Image Segmentation using TransUNet (Region Extraction)

- Input image:

$$I'' \in \mathbb{R}^{H \times W \times C}$$

- Extract feature maps using CNN encoder
- Convert feature maps into patches and apply Transformer encoder
- Self-attention computation:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

- Decoder reconstructs segmentation map:

$$P = f_{\text{TransUNet}}(I'')$$

- Generate binary mask:

$$M(x, y) = \begin{cases} 1, & P(x, y) \geq T \\ 0, & \text{otherwise} \end{cases}$$

- Extract Region of Interest (ROI):

$$I_{roi} = I'' \odot M$$

Step 3: Dual-Branch Feature Extraction (RegNet)

Branch 1: Global Feature Learning

- Input: Original image I''
- Extract features using RegNet: $H_{global} = f_{RegNet}(I'')$

Branch 2: Local Feature Learning

- Input: Segmented ROI I_{roi}
- Extract features: $H_{local} = f_{RegNet}(I_{roi})$
- Residual learning inside RegNet:
 $H(x) = F(x) + x$

Step 4: Attention-Based Feature Fusion

- Combine global and local features using attention mechanism:

$$H_{fuse} = \alpha H_{global} + \beta H_{local}$$

Step 5: Feature Aggregation and Classification

- Apply Global Average Pooling:
 $H_{gap} = \text{GAP}(H_{fuse})$
- Pass through Fully Connected layers:
 $Z_1 = f(W_1 H_{gap} + b_1)$
 $Z_2 = f(W_2 Z_1 + b_2)$
- Apply Dropout (0.5) for regularization
- Final classification using Softmax:
 $y = \text{softmax}(Z_2)$

Step 6: Model Training and Evaluation

- Compile model using: Optimizer: Adam, Loss Function: Categorical Cross-Entropy
- Train with: Batch size: 16, Epochs: 100
- Evaluate the model

Step 7: Model Output

- Predict disease class label:
 $y = \arg \max (\text{softmax}(Z_2))$

Save the trained model

End

4. Results and discussion

4.1 Training and validation accuracy and loss analysis

The accuracy and loss curves illustrate the learning behavior of the model over the training process. Fig. 7(a) shows that the accuracy increases rapidly during the initial training phase, as the model learns fundamental visual patterns such as leaf structure, color variations, and disease-related textures. The training accuracy improves from approximately 58–60% in the early epochs to around 90% within the first 20 epochs, indicating effective initial feature learning. As training progresses, the improvement becomes more gradual, and both

training and validation accuracies stabilize at approximately 97%–98% after 40 epochs, with the final validation accuracy reaching about 98.0%.

The close alignment between the training and validation curves indicates strong generalization capability with minimal overfitting.

The loss curve in Fig. 7(b) further supports this observation, showing a consistent reduction in prediction error. The training loss starts at approximately 0.65–0.70 and decreases rapidly to below 0.10 within the first 25 epochs, demonstrating effective optimization. Beyond this stage, both training and validation losses gradually decline and stabilize in the range of 0.01–0.03, indicating convergence of the model.

4.2 Performance evaluation and model interpretation

A comprehensive experimental evaluation was conducted to assess the effectiveness of the developed DL model for plant leaf disease classification. Standard classification metrics were employed to ensure reliable and balanced performance assessment. Accuracy measures the overall proportion of correctly classified samples, while precision and recall evaluate the model's ability to correctly identify disease classes and minimize misclassification. The evaluation metrics are mathematically defined in Eqs. (15) to (18).

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (15)$$

$$\text{Precision} = \frac{TP}{TP+FP} \quad (16)$$

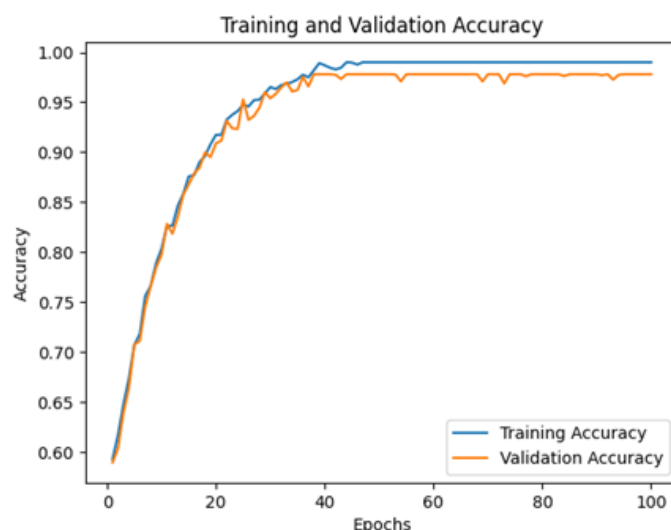
$$\text{Recall} = \frac{TP}{TP+FN} \quad (17)$$

$$\text{F1-score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (18)$$

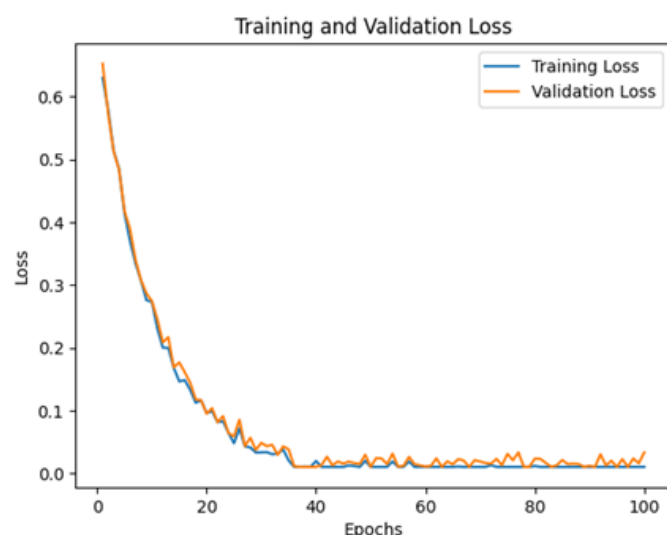
where TP , FP , TN and FN denote true positives, true negatives, false positives and false negatives respectively.

Table 3. Key hyperparameters of the proposed framework

Hyperparameter	Value
Input Image Size	$224 \times 224 \times 3$
Batch Size	16
Activation Function	ReLU
Loss Function	Categorical Cross-Entropy
Number of Epochs	100
Optimizer	Adam
Dropout Rate	0.5
Output Layer	Softmax
Learning Rate	1×10^{-4}



(a)



(b)

Figure. 7 Training performance of the proposed model: (a) accuracy curves and (b) loss curves

All classification results reported in this section, including the class-wise metrics, aggregate performance measures, confusion matrix, ROC analysis, and classification report, were derived from the same held-out test partition consisting of 813 images established during dataset preparation. This ensured that all reported performance indicators were computed using an identical evaluation dataset and were directly comparable.

Table 4 presents the class-wise classification performance of the developed model. The Black Rot class achieves a precision of 98.7%, recall of 97.9%, and F1-score of 98.3%, indicating highly accurate identification with minimal confusion among related disease categories. Similarly, the Esca class demonstrates strong classification performance, achieving a precision of 98.2%, recall of 97.8%, and

F1-score of 98.0%. The Leaf Blight class records a precision of 97.2%, recall of 97.7%, and F1-score of 97.4%, reflecting effective recognition of disease-specific phenotypic characteristics.

Table 4. Class-wise classification report of the developed model

Class	Precision (%)	Recall (%)	F1-score (%)
Black Rot	98.7	97.9	98.3
Esca	98.2	97.8	98.0
Leaf Blight	97.2	97.7	97.4
Healthy	95.4	97.6	96.5
Accuracy	97.8		
Macro Avg	97.4	97.8	97.6
Weighted Avg	97.7	97.8	97.8

The Healthy class achieves a recall of 97.6% and an F1-score of 96.5%, with a slightly lower precision of 95.4% due to a small number of healthy samples being classified as diseased leaves. The close agreement between the macro-average and weighted-average metrics indicates balanced classification performance across all categories without significant class-specific bias.

Fig. 8 depicts the overall performance of the developed model. The model achieves an accuracy of 97.8%, with precision, recall, and F1-score values approximately in the range of 97.7%–97.9%, indicating consistent and balanced classification performance.

The confusion matrix compares predicted and ground-truth class labels to evaluate model performance. Fig. 9 illustrates the confusion matrix of the developed model. A total of 795 out of 813 samples are correctly classified, resulting in an overall accuracy of approximately 97.8%.

For the Black Rot class, 231 out of 236 samples are correctly identified, with minor misclassifications into Esca (2), Leaf Blight (2), and Healthy (1). The Esca class records 271 correct predictions out of 277 samples, with limited confusion toward Black Rot (2), Leaf Blight (3), and Healthy (1). The Leaf Blight class achieves 210 correct predictions out of 215 samples, with minor misclassifications into Black Rot (1), Esca (2), and Healthy (2). The Healthy class shows 83 correct predictions out of 85 samples, with only two samples misclassified as diseased classes,

including Esca (1) and Leaf Blight (1). The confusion matrix demonstrates that the developed framework maintains strong discriminative capability across all disease categories while exhibiting only a limited number of misclassifications between visually similar disease classes.

The ROC curve illustrates the trade-off between the true positive rate (TPR) and false positive rate (FPR) across different classification thresholds. As shown in Fig. 10, the model achieves an area under the curve (AUC) of 0.98, indicating strong discriminative capability across disease classes. The combination of high TPR and low FPR demonstrates that the model provides accurate and reliable classification with minimal misclassification.

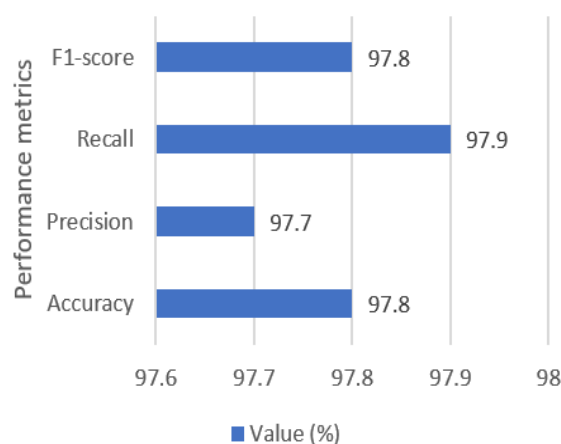


Figure. 8 Overall performance of the developed model

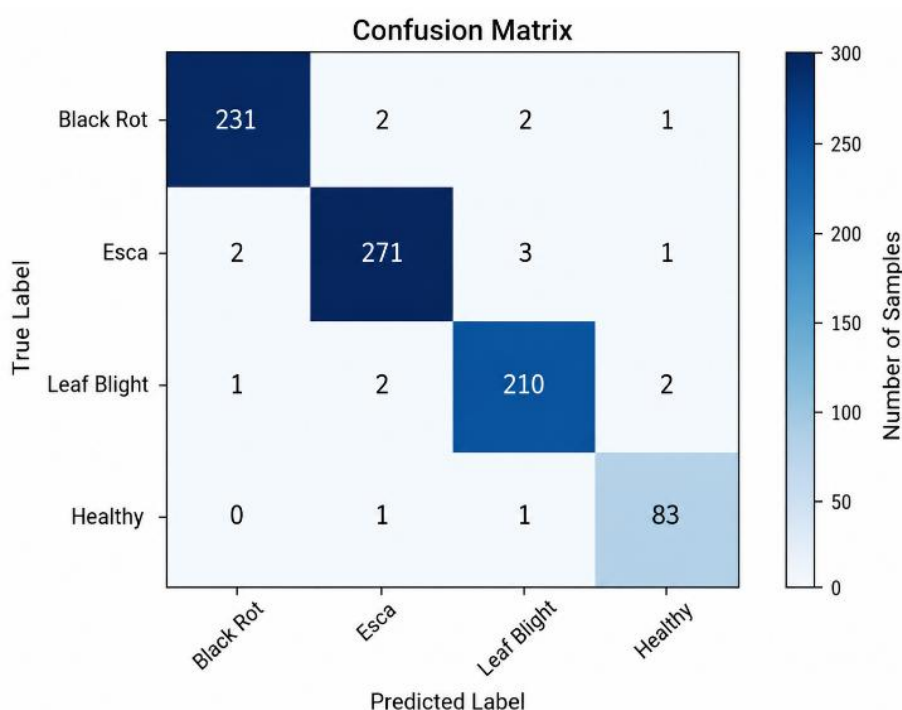


Figure. 9 Confusion matrix of the developed model

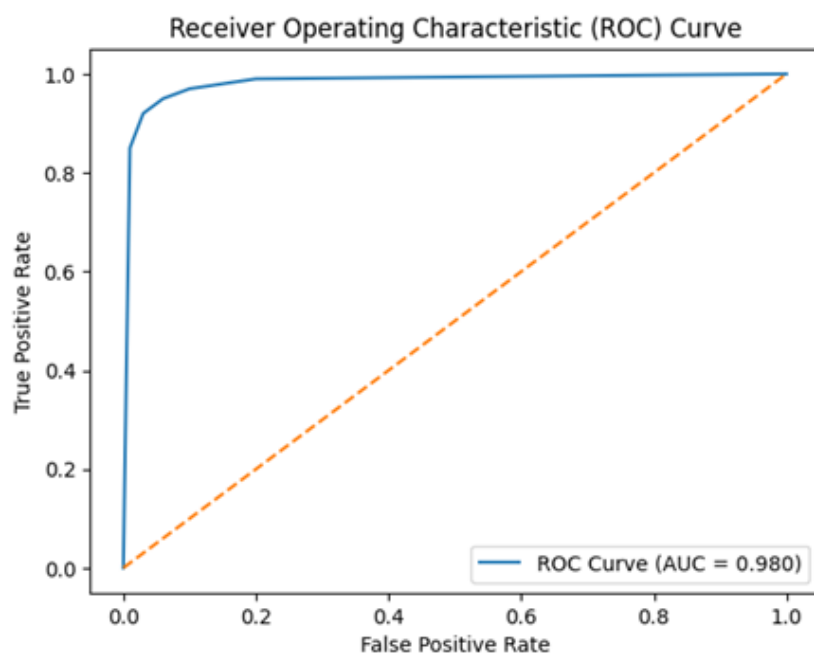


Figure. 10 ROC curve of the proposed model (AUC = 0.98)

Grad-CAM (Gradient-weighted Class Activation Mapping) is a visualization technique that highlights the regions of an image contributing most to a model's prediction by utilizing gradients of the target class flowing into the final convolutional layer. As shown in Fig. 11, the model effectively focuses on disease-specific regions such as lesions, discoloration patterns, and infected areas for Esca, Black Rot, and Leaf Blight samples, whereas attention is more uniformly distributed across the leaf surface in healthy samples.

These results indicate that the framework captures meaningful and relevant features, thereby enhancing both classification performance and model interpretability.

4.3 Segmentation performance

The performance of the segmentation module is evaluated to assess the effectiveness of the TransUNet-based approach in isolating phenotypically relevant leaf regions from the background.

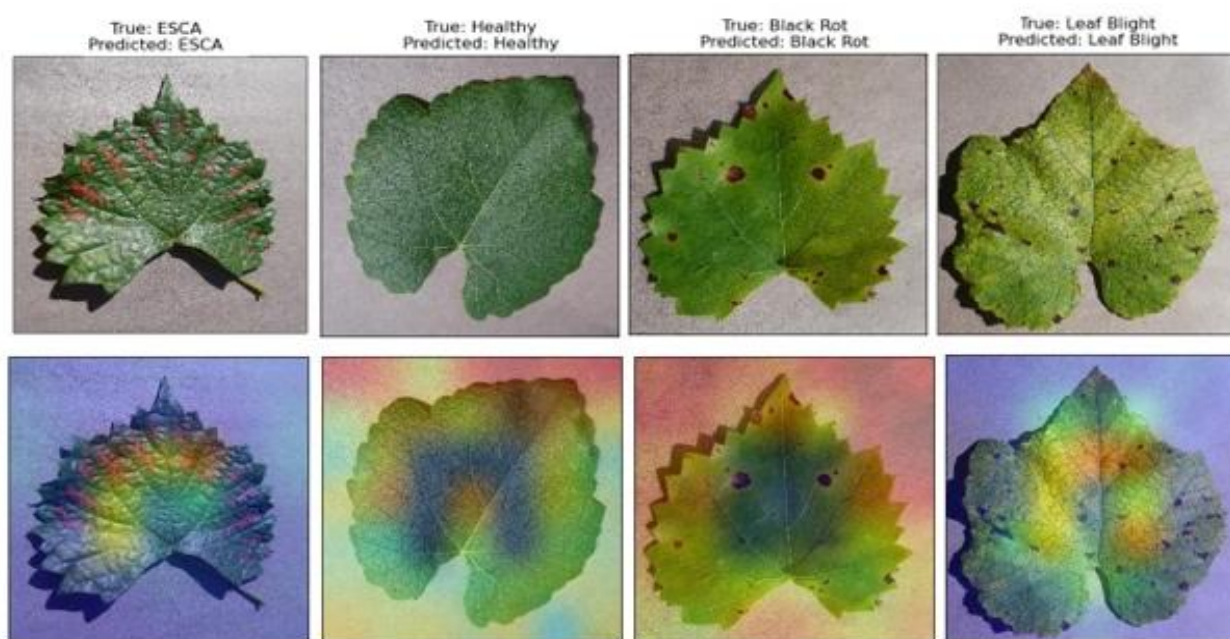


Figure. 11 Grad-CAM visualization results for grape leaf disease classification

Accurate segmentation enhances feature extraction by reducing the influence of irrelevant background information while preserving disease-specific characteristics such as lesion patterns and discoloration.

The model performance is quantified using the Dice coefficient and Intersection over Union (IoU), as reported in Table 5. The Dice coefficient measures the overlap between the predicted segmentation mask and the ground truth, whereas IoU evaluates the ratio of the intersection to the union of the predicted and actual regions. The Dice coefficient and IoU are calculated as follows:

$$Dice = \frac{2|P \cap G|}{|P| + |G|} \quad (19)$$

$$IoU = \frac{|P \cap G|}{|P \cup G|} \quad (20)$$

where P denotes the predicted segmentation mask and G represents the corresponding ground-truth mask. The Dice and IoU values are first computed independently for each test image and subsequently averaged across the test set to obtain the reported mean segmentation performance metrics.

The segmentation results indicate that the model achieves an IoU of 94.1% and a Dice score of 96.8%, demonstrating accurate localization of leaf regions. The reported mean IoU and Dice values are obtained by computing the respective metrics independently for each test image and subsequently averaging them across the test set. Consequently, these mean values do not necessarily satisfy the exact analytical relationship between IoU and Dice derived from a single aggregated overlap measure. The high overlap between predicted and ground-truth masks indicates that the segmentation module effectively captures leaf boundaries while minimizing background inclusion. This improves feature representation in subsequent stages, particularly for localized disease-related feature extraction. Consequently, the segmentation module contributes to overall classification performance by providing refined inputs for dual-branch feature learning. The reported segmentation metrics were obtained using the held-out test partition established during dataset preparation. Consequently, the Dice and IoU scores represent segmentation performance on previously unseen images, and the test images used for evaluation did not overlap with any images employed during segmentation model training or validation. Furthermore, ROI extraction and subsequent classification evaluation were performed using the same test samples without incorporating any test-set information during model training or optimization,

thereby ensuring a consistent and leakage-free evaluation protocol across both segmentation and classification stages.

4.4 Comparison with baseline models

A comparative evaluation was conducted to assess the performance of the developed DL framework against several baseline models under consistent experimental conditions. All models were trained and tested using the same dataset split to ensure fairness and comparability. For reproducibility and fair comparison, all baseline models were evaluated using the same training and testing partitions, preprocessing pipeline, and augmentation strategy adopted for the proposed framework. The traditional machine learning models were trained using feature vectors extracted from the preprocessed leaf images. KNN was implemented with $k = 5$ using the Euclidean distance metric. SVM employed a radial basis function (RBF) kernel with a regularization parameter of $C = 1.0$. RF was configured with 100 decision trees and an automatically determined maximum tree depth. For deep learning baselines, the CNN model consisted of convolutional, max-pooling, and fully connected layers trained using the Adam optimizer with a learning rate of 0.0001, batch size of 16, and 100 training epochs. VGG16 and ResNet50 were initialized with ImageNet-pretrained weights and subsequently fine-tuned for grape leaf disease classification. The final classification layers were replaced with task-specific fully connected layers corresponding to the four target classes. All deep learning models employed identical preprocessing, augmentation, optimization settings, and evaluation protocols to ensure a fair and consistent comparative assessment. A fixed random seed value of 42 was used throughout data partitioning, model initialization, and training procedures to enhance experimental reproducibility. Model selection was based on the highest validation accuracy achieved during training.

The results, presented in Table 6 and Fig. 12, include both traditional machine learning (ML) and deep learning (DL) approaches. Traditional methods such as Support Vector Machine (SVM), Random Forest (RF), and K-Nearest Neighbors (KNN) rely on handcrafted feature representations and exhibit lower

Table 5. Segmentation performance metrics

Metric	Value (%)
Dice Score	96.8
IoU	94.1

Baseline performance analysis

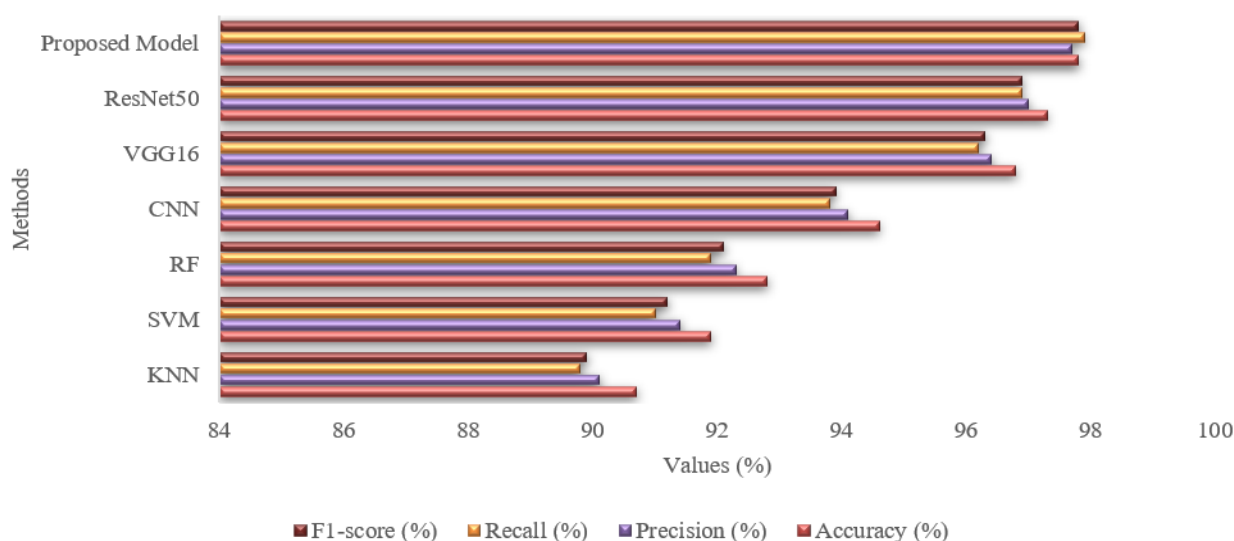


Figure. 12 Baseline performance analysis comparing the proposed model with conventional ML and DL approaches

Table 6. Performance comparison with baseline models

Model	Accuracy (%)	Recall (%)	Precision (%)	F1-score (%)
KNN	90.7	89.8	90.1	89.9
SVM	91.9	91.0	91.4	91.2
RF	92.8	91.9	92.3	92.1
CNN	94.6	93.8	94.1	93.9
VGG16	96.8	96.2	96.4	96.3
ResNet50	97.3	96.9	97.0	96.9
Proposed Model	97.8	97.9	97.7	97.8

performance, with accuracies of 91.9%, 92.8%, and 90.7%, respectively. These approaches are limited in their ability to capture complex and subtle disease patterns in leaf images.

In contrast, deep learning models demonstrate superior performance due to their ability to learn hierarchical feature representations directly from data.

The conventional CNN model achieves an accuracy of 94.6%, while VGG16 and ResNet50 attain 96.8% and 97.3%, respectively. The developed framework achieves an accuracy of 97.8%, demonstrating the effectiveness of integrating segmentation-based region extraction, dual-branch feature learning, and attention-based feature fusion for grape leaf disease classification.

The results indicate a consistent performance improvement when transitioning from traditional machine learning approaches to deep learning frameworks. Among baseline models, RF outperforms SVM and KNN, while deeper architectures such as VGG16 and ResNet50 provide further gains due to their enhanced representation

capability. Compared with conventional single-stream architectures such as ResNet50, the developed framework combines global contextual information from original images with localized disease-specific characteristics extracted from segmented regions, resulting in richer and more discriminative feature representations. The developed model achieves strong classification performance along with closely aligned precision and recall values, indicating balanced and robust recognition across all disease categories.

It should be noted that the PlantVillage dataset represents a controlled experimental setting with relatively uniform imaging conditions. Although the proposed segmentation and classification framework demonstrates strong performance under these conditions, further evaluation on field-acquired datasets containing diverse environmental variations would provide additional insight into the model's robustness and practical deployment potential.

4.5 Computational complexity and deployment analysis

A computational complexity analysis was performed to evaluate the balance between classification performance and computational efficiency for practical plant phenotyping applications. The assessment considered parameter count, floating-point operations (FLOPs), inference time, and deployment cost, providing insights into the computational demands and real-time applicability of different architectures. To ensure a fair comparison, all models were trained and tested using the same

dataset partition, preprocessing pipeline, augmentation strategy, and input image size of $224 \times 224 \times 3$. VGG16, ResNet50, and EfficientNet-B3 were initialized with ImageNet-pretrained weights and fine-tuned for grape leaf disease classification. ViT-B16 employed pretrained transformer weights with 16×16 image patches, while the Dual-Branch CNN baseline used parallel convolutional streams followed by feature fusion and classification layers. The developed framework integrated TransUNet-based segmentation, dual-branch RegNet feature extraction, and attention-guided fusion. All models were trained using the Adam optimizer with a learning rate of 0.0001, batch size of 16, up to 100 epochs, and a fixed random seed of 42 to ensure reproducibility.

The computational complexity comparison presented in Table 7 demonstrates notable differences among the evaluated architectures in terms of predictive performance and resource requirements. VGG16 achieves 96.8% accuracy but requires the highest parameter count (138.4 M) and substantial computational resources. ResNet50 improves accuracy to 97.3% while significantly reducing parameters, FLOPs, and inference time. EfficientNet-B3 is the most computationally efficient model, with only 12.0 M parameters and 1.8 GFLOPs, though its accuracy decreases to 96.9%. ViT-B16 attains 97.4% accuracy but exhibits the largest computational overhead in terms of FLOPs and inference latency. The Dual-Branch CNN demonstrates moderate complexity with 96.7% accuracy. The developed framework achieves the highest accuracy of 97.8%

while maintaining a moderate computational footprint of 34.8 M parameters, 7.2 GFLOPs, and 22.6 ms inference time. Although slightly more demanding than ResNet50 and EfficientNet-B3, it remains substantially more efficient than VGG16 and ViT-B16. These findings demonstrate an effective balance between predictive accuracy, computational efficiency, and practical deployment suitability for automated plant phenotyping systems.

4.6 Stability analysis

To evaluate the robustness and reproducibility of the proposed framework, the classification model was assessed across five independent training runs using identical preprocessing, augmentation, and training configurations. The repeated-run analysis was performed to assess the consistency of the learning process and the stability of the obtained performance metrics across multiple experiments. The objective was to determine whether the reported performance gains were consistently achievable and not attributable to favorable optimization behavior during a single training run. Table 8 presents the performance variability of the proposed framework across five experimental runs together with the corresponding mean, standard deviation, and 95% confidence interval estimates. The confidence intervals were computed using the Student's *t*-distribution, which is appropriate for estimating interval statistics from a limited number of experimental runs.

Table 7. Computational complexity and deployment analysis

Model	Accuracy (%)	Parameters (M)	FLOPs (G)	Inference Time (ms)	Deployment Cost
VGG16	96.8	138.4	15.5	34.2	High
ResNet50	97.3	25.6	4.1	18.7	Medium
EfficientNet-B3	96.9	12.0	1.8	16.4	Low
ViT-B16	97.4	86.0	17.6	41.5	High
Dual-Branch CNN	96.7	31.5	6.8	21.8	Medium
Proposed Framework	97.8	34.8	7.2	22.6	Medium

Table 8. Stability analysis of the proposed model across multiple experimental runs

Run	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
Run 1	97.72	97.63	97.80	97.71
Run 2	97.85	97.76	97.91	97.82
Run 3	97.78	97.70	97.84	97.75
Run 4	97.69	97.58	97.72	97.65
Run 5	97.88	97.83	97.92	97.86
Mean $\pm$ Std	97.78 $\pm$ 0.08	97.70 $\pm$ 0.10	97.84 $\pm$ 0.08	97.76 $\pm$ 0.09
95% CI	97.68–97.88	97.58–97.82	97.74–97.94	97.65–97.87

The stability analysis confirms the robustness and reproducibility of the developed framework across repeated experimental runs. As presented in Table 8, the model achieves a mean accuracy of $97.78\% \pm 0.08$, with a 95% confidence interval of 97.68–97.88%, indicating highly consistent performance and stable convergence across repeated experiments. Precision, recall, and F1-score also exhibit strong results, with mean values of $97.70\% \pm 0.10$, $97.84\% \pm 0.08$, and $97.76\% \pm 0.09$, respectively, while the corresponding confidence intervals demonstrate limited performance variation between runs. The low standard deviations across all metrics demonstrate minimal sensitivity to random initialization and training variability.

These findings suggest that the integration of segmentation-based region extraction, dual-branch feature learning, and attention-guided feature fusion consistently contributes to reliable disease classification. Overall, the limited variation among repeated runs confirms the framework's stability, generalizability, and reproducibility, indicating that the reported performance gains are robust and not dependent on isolated experimental outcomes.

4.7 Ablation study

An ablation study was conducted to analyze the contribution of key components of the developed framework, including segmentation, dual-branch feature learning, and attention-based fusion, as summarized in Table 9. All configurations were evaluated under identical experimental settings to ensure reliability and comparability.

The baseline configuration, consisting of a single-branch RegNet model trained on raw images, achieved an accuracy of 93.8%, indicating that global features alone capture overall structural patterns but are insufficient for modeling fine-grained disease characteristics. When segmentation was incorporated

and only the region of interest (ROI) was used as input, the accuracy improved to 95.1%, demonstrating the importance of background removal in enhancing feature relevance.

The dual-branch configuration, where the original image and segmented ROI were processed in parallel without attention-based fusion, further improved accuracy to 96.6%, highlighting the benefit of combining global and localized features. Finally, the complete model integrating segmentation, dual-branch learning, and attention-based fusion achieved the highest accuracy of 97.8%, confirming the effectiveness of comprehensive feature integration for improved classification performance.

4.8 Cross-validation and statistical reliability analysis

To further assess the robustness and statistical reliability of the developed framework, a five-fold cross-validation experiment was conducted using the PlantVillage grape leaf dataset. The dataset was randomly partitioned into five mutually exclusive folds of approximately equal size. During each iteration, four folds were used for training and one fold was reserved for testing. This process was repeated five times, ensuring that each sample was utilized exactly once for testing. The same preprocessing operations, augmentation strategy, network architecture, hyperparameters, and evaluation protocol were maintained across all folds to ensure consistency and fairness.

The developed framework was trained using the Adam optimizer with a learning rate of 0.0001, batch size of 16, and 100 training epochs. Classification performance was evaluated using Accuracy, Precision, Recall, and F1-score. The fold-wise results obtained from the cross-validation experiment are summarized in Table 10.

Table 9. Ablation study results

Configuration	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
RegNet (Preprocessed Image)	93.8	93.3	92.9	93.1
RegNet (Segmented ROI)	95.1	94.6	94.2	94.4
Dual-Branch RegNet (Without Attention Fusion)	96.6	96.2	96.0	96.1
Proposed Model	97.8	97.7	97.9	97.8

Table 10. Five-fold cross-validation performance of the developed framework

Fold	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
Fold 1	97.4	97.2	97.5	97.3
Fold 2	98.0	97.8	98.1	97.9
Fold 3	97.7	97.5	97.8	97.6
Fold 4	98.2	98.0	98.3	98.1
Fold 5	97.8	97.7	97.9	97.8
Mean $\pm$ SD	97.82 ± 0.30	97.64 ± 0.30	97.92 ± 0.29	97.74 ± 0.29

The results demonstrate consistent performance across all folds, with classification accuracy ranging from 97.4% to 98.2%. The mean accuracy of 97.82% closely aligns with the accuracy obtained using the fixed train-test evaluation protocol. Furthermore, the low standard deviations observed for all metrics indicate limited performance variability across different dataset partitions. These findings confirm the stability, reproducibility, and generalization capability of the developed framework, suggesting that the reported performance is not dependent on a specific train-test split but is consistently achieved across multiple validation folds.

To further assess the statistical reliability of the obtained results, a 95% confidence interval (CI) was estimated using the observed classification accuracy and the size of the independent test set. The confidence interval was calculated as:

$$CI = p \pm 1.96 \sqrt{\frac{p(1-p)}{n}} \quad (20)$$

where p denotes the observed classification accuracy and n represents the total number of test samples.

For the developed framework, an accuracy of 97.8% was obtained on 813 test samples. The corresponding 95% confidence interval was calculated as 96.8%–98.8%, as summarized in Table 11.

The relatively narrow confidence interval indicates limited uncertainty in the estimated classification accuracy and suggests stable predictive performance on unseen samples drawn from the same data distribution. The confidence interval analysis further supports the reliability of the developed framework and complements the cross-validation results by providing an estimate of the expected variation in classification performance.

4.9 Performance comparison

Recent grape disease analysis studies can generally be categorized into classification, segmentation-assisted classification, and object detection frameworks. Since the primary objective of the developed framework is disease classification, only studies reporting disease classification performance were included in the quantitative comparison. Object detection approaches were excluded because they employ fundamentally different evaluation criteria, such as localization accuracy and mean Average Precision (mAP), which are not directly comparable with classification accuracy metrics.

Table 12 presents a comparison of recent grape leaf disease classification approaches reported between 2024 and 2025 that are methodologically comparable to the developed framework in terms of disease recognition objectives. To improve transparency, the table summarizes the datasets, disease categories, model architectures, evaluation metrics, and available evaluation protocol information reported in the original studies. Because the reviewed methods employ different dataset compositions, imaging conditions, disease categories, and validation strategies, direct numerical comparison should be interpreted cautiously. The comparison is intended to position the developed framework relative to recent grape disease recognition approaches rather than establish a strictly standardized benchmark under identical experimental conditions.

It should also be noted that the PlantVillage dataset contains images from multiple crop species and disease categories. The present study utilizes only the grape leaf subset comprising Black Rot, Esca, Leaf Blight, and Healthy classes. To improve task relevance and comparison fairness, Table 12 includes only studies specifically addressing grape disease recognition. Studies focusing on other crop species or multi-crop disease classification tasks were intentionally excluded from the comparative analysis.

Studies reported in [21-26] demonstrate the continued evolution of grape leaf disease recognition through hybrid feature engineering, convolutional neural networks, ensemble learning, attention-enhanced architectures, and Transformer-based frameworks. The reported classification accuracies range from 93.00% to 96.36% across different datasets and experimental conditions. Attention-enhanced architectures employing CBAM and SENet modules achieved accuracies up to 96.36% [26], whereas Transformer-based multimodal fusion strategies reported an accuracy of 94.00% [25]. Ensemble learning approaches and hybrid feature extraction frameworks reported accuracies between 95.34% and 96.10% [21-23].

The developed segmentation-guided dual-branch attention framework achieved an accuracy of 97.8% on the PlantVillage grape leaf subset and demonstrated performance comparable to recent

Table 11. Confidence interval analysis of the developed framework

Metric	Value
Accuracy (%)	97.8
95% CI Lower Bound (%)	96.8
95% CI Upper Bound (%)	98.8

Table 12. Performance comparison of recent grape leaf disease classification approaches

Author(s)	Year	Dataset	Disease Classes	Model / Approach	Accuracy (%)
Kaur & Devendran [21]	2024	PlantVillage (Grape subset)	Black Rot, Black Measles, Leaf Blight, Healthy	K-means + GWO + Hybrid Features + Ensemble	95.69
Khan et al. [22]	2024	Farm-based Images	Powdery Mildew, Blotches, Healthy	CNN + ML + Stacking	96.10
Ishengoma & Lyimo [23]	2024	PlantVillage Grape (4040 images)	Black Rot, Esca, Leaf Blight, Healthy	VGG16-Xception + Random Forest	95.34
Yücalar & Yıldırım [24]	2025	Hybrid Dataset (Public + Field Images)	Powdery Mildew, Downy Mildew, Rust Mite	Inception-ResNet-V2	93.00
Li et al. [25]	2025	Vineyard Dataset (RGB + Hyperspectral Data)	Downy Mildew, Powdery Mildew, Gray Mold, Anthracnose, Spot Anthracnose	Transformer-Based Multimodal Fusion Network	94.00
Unal [26]	2025	Vineyard Dataset	Downy Mildew, Grape Leaf Blister Mite, Healthy	DenseNet121 + CBAM + SENet	96.36
Proposed Model		PlantVillage Grape	Black Rot, Esca, Leaf Blight, Healthy	Segmentation + Dual-Branch Learning + Attention Fusion	97.8

Note: Evaluation protocols are reported according to the information available in the corresponding publications. Direct protocol-level equivalence is not always possible because of differences in dataset composition, disease categories, image acquisition conditions, and validation strategies.

grape disease recognition approaches while additionally incorporating segmentation-guided disease localization, dual-branch feature learning, and adaptive attention-based feature fusion. The ablation analysis presented in Table 9 further demonstrates the contribution of each architectural component, showing progressive performance improvements as segmentation, dual-branch learning, and attention fusion mechanisms are incorporated.

The observed performance can be attributed to the complementary utilization of global contextual information from original leaf images and localized disease-specific characteristics extracted from segmented regions. However, because the compared studies utilize different datasets, disease categories, image acquisition conditions, and validation protocols, the reported results should be interpreted as a representative overview of recent developments rather than a direct ranking of state-of-the-art performance. A fully standardized comparison would require implementation and evaluation of all competing methods under identical datasets, disease categories, preprocessing procedures, and validation protocols, which is beyond the scope of the present study.

5. Conclusion

This study presents a DL-based framework for plant phenotyping aimed at accurate classification of leaf disease conditions. The method addresses key limitations of existing approaches, particularly the

inability to simultaneously capture global structural features and localized disease-specific characteristics. By integrating segmentation-based region extraction, dual-branch feature learning, and attention-based fusion, the framework enables effective extraction of discriminative phenotypic features. The segmentation module demonstrates strong performance, achieving an IoU of 94.1% and a Dice score of 96.8%, indicating accurate isolation of disease-relevant leaf regions. The dual-branch learning strategy captures complementary information from both the original image and the segmented region, while the attention-based fusion mechanism enhances feature representation through adaptive integration of global and local features. As a result, the model achieves an overall classification accuracy of 97.8%, with precision, recall, and F1-score values in the range of 97.7%–97.9%, demonstrating consistent and balanced performance. Although the proposed framework demonstrates strong performance on the PlantVillage grape leaf dataset, it is important to acknowledge that the dataset consists of images acquired under controlled conditions with limited background variability. Consequently, the reported results should be interpreted within the context of controlled image-based disease classification. Therefore, the current study primarily establishes the effectiveness of the proposed framework under standardized imaging conditions rather than demonstrating full real-world deployment capability. This framework contributes to image-based plant phenotyping by providing an

effective solution for automated disease analysis under controlled imaging conditions. Future work includes validation on field-acquired datasets containing complex backgrounds, illumination variations, occlusions, scale changes, and other environmental challenges, as well as the integration of multi-modal imaging data and optimization for deployment on edge devices to enable real-time monitoring in precision agriculture.

Conflicts of Interest

The authors declare that there are no conflicts of interest related to this manuscript.

Author Contributions

The paper conceptualization, methodology, software, validation, formal analysis, investigation, resources, data curation, writing—original draft preparation, writing—review and editing, visualization, have been done by 1st author. The supervision and project administration have been done by 2nd author.

Acknowledgments

I would like to express my sincere gratitude to all those who contributed to the completion of this research paper. I extend my heartfelt thanks to my supervisor, my family, my colleagues and fellow researchers for their encouragement and understanding during the demanding phases of this work.

References

- [1] K. Pawlak and M. Kołodziejczak, “The role of agriculture in ensuring food security in developing countries: Considerations in the context of the problem of sustainable food production”, *Sustainability*, Vol. 12, No. 13, p. 5488, 2020.
- [2] S. S. Acharya, “Sustainable agriculture and rural livelihoods”, *Agricultural Economics Research Review*, Vol. 19, No. 2, pp. 205–217, 2006.
- [3] S. C. Gosa, Y. Lupo, and M. Moshelion, “Quantitative and comparative analysis of whole-plant performance for functional physiological traits phenotyping: New tools to support pre-breeding and plant stress physiology studies”, *Plant Science*, Vol. 282, pp. 49–59, 2019.
- [4] M. Haworth, G. Marino, G. Atzori, A. Fabbri, A. Daccache, D. Killi, *et al.*, “Plant physiological analysis to overcome limitations to plant phenotyping”, *Plants*, Vol. 12, No. 23, p. 4015, 2023.
- [5] A. M. Mutka and R. S. Bart, “Image-based phenotyping of plant disease symptoms”, *Frontiers in Plant Science*, Vol. 5, p. 734, 2015.
- [6] C. Kaya, “Optimizing crop production with plant phenomics through high-throughput phenotyping and AI in controlled environments”, *Food and Energy Security*, Vol. 14, No. 1, p. e70050, 2025.
- [7] J. W. Clohessy, S. Sanjel, G. K. O’Brien, R. Barocco, S. Kumar, S. Adkins, *et al.*, “Development of a high-throughput plant disease symptom severity assessment tool using machine learning image analysis and integrated geolocation”, *Computers and Electronics in Agriculture*, Vol. 184, p. 106089, 2021.
- [8] R. Zenkl, B. A. McDonald, A. Walter, M. Unvericht, C. Saintenac, and J. Anderegg, “From canopy images to organ-level disease assessments: A scalable approach to measure quantitative resistance in the field”, *bioRxiv*, 2025, doi: 10.1101/2025.04.30.651476.
- [9] A. Upadhyay, N. S. Chandel, K. P. Singh, S. K. Chakraborty, B. M. Nandede, M. Kumar, *et al.*, “Deep learning and computer vision in plant disease detection: A comprehensive review of techniques, models and trends in precision agriculture”, *Artificial Intelligence Review*, Vol. 58, No. 3, p. 92, 2025.
- [10] L. Zhu, L. Tang, and S. Ren, “Deep learning-based text generation for plant phenotyping and precision agriculture”, *Frontiers in Plant Science*, Vol. 16, p. 1564394, 2025.
- [11] M. Shafay, T. Hassan, M. Owais, I. Hussain, S. G. Khawaja, L. Seneviratne, and N. Werghi, “Recent advances in plant disease detection: Challenges and opportunities”, *Plant Methods*, Vol. 21, No. 1, p. 140, 2025.
- [12] T. Zhang, S. Huang, G. Kezierbieke, Y. Halimu, and H. Li, “FewMedical-XJAU: A challenging benchmark for fine-grained medicinal plant classification”, *Sensors*, Vol. 25, No. 17, p. 5499, 2025.
- [13] T. Wang, R. Tong, T. Xu, Y. Li, and Y. Chen, “Artificial intelligence in plant science: From image-based phenotyping to yield and trait prediction”, *Frontiers in Plant Science*, Vol. 16, p. 1732979, 2026.
- [14] E. Omia, E. Park, D. Semyalo, R. Joshi, and B. K. Cho, “Advancements in 3D field-crop phenotyping using point clouds: A comparative review of sensor technology, target traits and challenges under controlled and field

- conditions”, *Frontiers in Plant Science*, Vol. 17, p. 1731852, 2026.
- [15] W. Shafik, A. Tufail, L. C. De Silva, R. A. A. H. M. Apong, and K. H. Kim, “Deep learning technique for plant disease classification and pest detection and model explainability elevating agricultural sustainability”, *BMC Plant Biology*, Vol. 25, No. 1, p. 1491, 2025.
- [16] A. E. Ewiekpaefe, D. T. Chinyio, and N. P. Nwanna, “Comparative analysis of binary and multiclass potato leaf disease classification using VGG19 model”, *FUDMA Journal of Sciences*, Vol. 9, No. 9, pp. 207–217, 2025.
- [17] W. Shafik, A. Tufail, C. De Silva Liyanage, and R. A. A. H. M. Apong, “Using transfer learning-based plant disease classification and detection for sustainable agriculture”, *BMC Plant Biology*, Vol. 24, No. 1, p. 136, 2024.
- [18] M. H. Najim, S. K. Abdulateef, and A. H. Alasadi, “Early detection of tomato leaf diseases based on deep learning techniques”, *International Journal of Artificial Intelligence*, Vol. 13, No. 1, pp. 509–515, 2024.
- [19] I. Bouacida, B. Farou, L. Djakhadjakha, H. Seridi, and M. Kurulay, “Innovative deep learning approach for cross-crop plant disease detection: A generalized method for identifying unhealthy leaves”, *Information Processing in Agriculture*, Vol. 12, No. 1, pp. 54–67, 2025.
- [20] M. Naveenkumar and S. Nandagopal, “Adaptive hybrid segmentation combined with meta heuristic optimization in transfer learning for plant leaf disease classification”, *Scientific Reports*, Vol. 15, No. 1, p. 9838, 2025.
- [21] N. Kaur and V. Devendran, “A novel framework for semi-automated system for grape leaf disease detection”, *Multimedia Tools and Applications*, Vol. 83, No. 17, pp. 50733–50755, 2024.
- [22] K. H. Khan, A. Aljaedi, M. S. Ishtiaq, H. Imam, Z. Bassfar, and S. S. Jamal, “Disease detection in grape cultivation using strategically placed cameras and machine learning algorithms with a focus on powdery mildew and blotches”, *IEEE Access*, Vol. 12, pp. 139505–139523, 2024.
- [23] F. S. Ishengoma and N. N. Lyimo, “Ensemble model for grape leaf disease detection using CNN feature extractors and random forest classifier”, *Heliyon*, Vol. 10, No. 12, p. e33377, 2024, doi: 10.1016/j.heliyon.2024.e33377.
- [24] F. Yücalar and R. Yildirim, “Disease detection from grape plant leaves using transfer learning methods”, *Sinop Üniversitesi Fen Bilimleri Dergisi*, Vol. 10, No. 2, pp. 497–512, 2025.
- [25] W. Li, B. Zhou, Y. Zhou, C. Jiang, M. Ruan, T. Ke, H. Wang, and C. Lv, “Grape disease detection using transformer-based integration of vision and environmental sensing”, *Agronomy*, Vol. 15, No. 4, p. 831, 2025.
- [26] Y. Unal, “Integrating CBAM and squeeze-and-excitation networks for accurate grapevine leaf disease diagnosis”, *Food Science & Nutrition*, Vol. 13, No. 6, p. e70377, 2025, doi: 10.1002/fsn3.70377.
- [27] TensorFlow, “PlantVillage Dataset”, [Online]. Available: https://www.tensorflow.org/datasets/catalog/plant_village
- [28] R. Song, Y. Wu, L. Wan, S. Shao, and H. Wu, “SE-TransUNet-based semantic segmentation for water leakage detection in tunnel secondary linings amid complex visual backgrounds”, *Applied Sciences*, Vol. 15, No. 14, p. 7872, 2025.
- [29] S. Acosta-Jiménez, V. Maeda-Gutiérrez, C. E. Galván-Tejada, M. M. Mendoza-Mendoza, L. C. Reveles-Gómez, J. M. Celaya-Padilla, et al., “Assessing ResNeXt and RegNet models for diabetic retinopathy classification: A comprehensive comparative study”, *Diagnostics*, Vol. 15, No. 15, p. 1966, 2025.