

DSGSU-Net: A U-Net-Based Model for Tomato Leaf Disease Segmentation Using Depthwise Separable Convolutions and Ghost Sampling

jansi e

Vellore Institute of Technology University

kavitha B R

kavitha.br@vit.ac.in

Vellore Institute of Technology University

Article

Keywords: Tomato leaf, Deep learning, enhanced U-net, depthwise separable convolutions, Ghost sampling, dilated convolutions, Segmentation

Posted Date: June 17th, 2025

DOI: <https://doi.org/10.21203/rs.3.rs-6512719/v1>

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

[Read Full License](#)

Additional Declarations: No competing interests reported.

DSGSU-Net: A U-Net-Based Model for Tomato Leaf Disease Segmentation Using Depthwise Separable Convolutions and Ghost Sampling

Jansi E¹ and Kavitha B R^{2*}

^{1,2}School of Computer Science Engineering and Information Systems, Vellore Institute of Technology, Vellore 632014, India

*kavitha.br@vit.ac.in

ABSTRACT

Tomato leaf disease poses a significant threat to global agricultural productivity, underscoring the need for accurate and automated segmentation techniques for early detection and intervention. In this study, we proposed DSGSU-Net, an enhanced U-Net-based architecture explicitly designed for the precise segmentation of tomato leaf diseases. The model incorporates depthwise separable convolutions for efficient feature extraction, Dilated Convolutions in deeper layers for multi-scale context aggregation, and Ghost Sampling in the decoder for improved upsampling. To further enhance segmentation performance, a hybrid loss function combining Dice Loss and Focal Loss is utilized to manage class imbalance and enhance the boundary delineation. Experiments conducted on the PlantVillage dataset (bacterial spot class) demonstrated that DSGSU-Net achieved an accuracy of 0.9572, F1-score of 0.8276, precision of 0.7156, recall of 0.9885, IoU of 0.7102, and a Dice coefficient of 0.9822. The results show that DSGSU-Net outperforms conventional U-Net models in segmentation accuracy and computational efficiency, making it a strong contender for practical use in precision agriculture and disease surveillance.

Keywords: Tomato leaf, Deep learning, enhanced U-net, depthwise separable convolutions, Ghost sampling, dilated convolutions, Segmentation.

Introduction

Agriculture plays a pivotal role in sustaining the global food supply, and tomato cultivation holds substantial economic value across various regions. However, tomato plants are particularly susceptible to a range of diseases that can cause significant yield losses. Among these, leaf diseases—such as bacterial spots—often manifest visibly on the plant’s surface, making them suitable for automated detection using image-based methods. Early and accurate segmentation of disease-affected regions is crucial for targeted intervention and improved crop management.

Deep-learning methods, especially Convolutional Neural Networks (CNNs), have transformed the field of image-based disease detection by providing exceptional feature extraction and classification abilities. In particular, U-Net has become well-known for its capability to achieve pixel-level segmentation with remarkable accuracy. However, conventional U-Net models are often resource-intensive and may find it challenging to identify detailed disease patterns, especially when dealing with small or unbalanced datasets. In response to these challenges, we introduce DSGSU-Net, an improved model based on U-Net that incorporates architectural advancements and a strong loss function to enhance segmentation accuracy and efficiency.

Our contributions can be encapsulated in four primary areas.

1. We incorporate depthwise separable convolutions to minimize computational expense while maintaining feature complexity.
2. Dilated Convolutions were introduced in the deeper layers to enable multi-scale context aggregation.
3. We incorporated a unique spectral sampling module within the decoder to enhance the upscaling resolution and boundary definition.
4. A hybrid loss function was designed by combining Dice and Focal Loss to effectively tackle class imbalance and boundary segmentation problems.

To provide a clear overview of the comprehensive segmentation workflow utilized in our study, Figure 1 illustrates the entire process divided into five essential stages. **Image Collection:** images gathered from the PlantVillage dataset, focusing on the bacterial spot class of tomato leaves. **Mask Creation:** Ground truth masks are semi-automatically generated to annotate disease-affected regions. **Data Preprocessing:** Images and masks undergo resizing, normalization, and augmentation to ensure uniformity and improve model generalization. **Segmentation Process:** The core stage is where our proposed DSGSU-Net



Figure 1. Step-wise pipeline for disease segmentation, comprising image collection, mask creation, preprocessing, segmentation using DSGSU-Net, and final mask prediction

model is trained on the preprocessed dataset to learn spatial features and predict pixel-wise labels. **Mask Prediction:** The trained model is used to infer segmentation masks on images, identifying diseased regions with high precision.

The rest of this paper is organized as follows: Section 2 examines previous research, Section 3 outlines the proposed methodology, Section 4 showcases the experimental results along with a discussion, and Section 5 wraps up with concluding thoughts and suggestions for future research avenues.

Related Work

Traditional Segmentation Methods for Plant Diseases

Traditional segmentation techniques have played a crucial role in the early detection of plant diseases. These methods generally depend on the manually designed features and standard image processing strategies, including color transformation, thresholding, and texture analysis. In the initial research, RGB images of infected tomato leaves were transformed into different color spaces, such as HSI or HSV, to improve the contrast between healthy and diseased areas. Texture-based features, especially those obtained from the Gray Level Co-occurrence Matrix (GLCM), were frequently utilized to describe disease symptoms. For example, in¹, researchers extracted GLCM features from images of tomato leaves and applied a Support Vector Machine (SVM) for classifying fungal infections, achieving an accuracy of up to 86%. In a similar vein,² integrated GLCM features with color histograms and employed an SVM classifier, resulting in a remarkable classification accuracy of 97.29%. Furthermore,³ introduced the Histogram of Oriented Gradients (HOG) along with an SVM for detecting late blight disease in tomato leaves. Despite the strong performance of these techniques, they often faced challenges in generalizing across different lighting conditions and visually similar disease patterns due to their dependence on a restricted set of features.

To improve segmentation precision,^{4,5} implemented a K-means clustering algorithm to identify diseased regions, followed by morphological processing to minimize noise. The features of the extracted lesions were analyzed using the Local Binary Pattern (LBP) operator and subsequently input into a dual-channel network for classification. This approach achieved an accuracy exceeding 96% for conditions like early blight and powdery mildew. However, the high computational complexity presented a notable drawback.

In a separate investigation,⁶ applied a combination of integrated thresholding, edge detection, and K-means clustering to enhance the segmentation of the tomato brown spot disease. They used an advanced binary-tree SVM classifier, which resulted in a roughly 6% increase in accuracy compared to the traditional one-vs-all SVM models. Despite these findings, the study's results were constrained by the relatively small dataset, thus limiting the model's practical application.

A new approach that relies on saliency was proposed in⁷, which entailed choosing seed points for segmenting frog-eye and bacterial spots on tomato leaves. This method was refined using a particle swarm optimization (PSO) algorithm and paired with an SVM classifier, achieving an early disease detection rate of 92%. However, the performance declined when the contrast

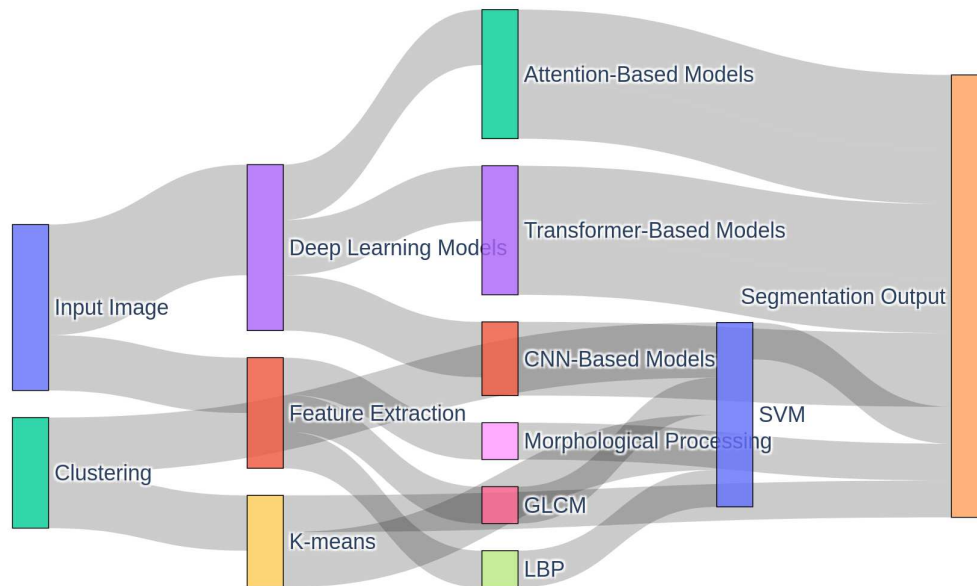


Figure 2. Comparative Progression of Traditional and Deep Learning-Based Segmentation Methods in Plant Disease Detection

between the lesions and the background was minimal, rendering the segmentation unreliable for subtle infections.

^{8,9} used morphological operations in conjunction with wavelet transformations and the Otsu thresholding algorithm to segment the tomato leaf lesions. This technique attained an accuracy of over 90%, but its dependence on grayscale features limited its effectiveness when the intensity of the lesions closely matched that of the background.

Deep Learning-Based Approaches for Tomato Leaf Disease Segmentation

Deep learning segmentation techniques have demonstrated exceptional effectiveness in detecting plant diseases, especially in intricate environments. To address the issues of class imbalance and challenging imaging conditions,¹⁰ introduced an enhanced scSE attention mechanism paired with an asymmetric shuffle convolution module. Their Asymmetric Shuffle Convolutional Neural Network (ASNet) achieved an impressive average segmentation accuracy of 94.7% across various tomato leaf diseases, although its performance for rust disease was hindered by a scarcity of training samples.

In their research,^{11,12} advanced the DeepLabV3+ network by incorporating an Efficient Channel Attention (ECA) module and a Feature Pyramid Network (FPN) to boost the segmentation accuracy for black rot lesions. Their model recorded Dice scores of 91.8% and 86.1% on different test datasets, but difficulties persisted in segmenting overlapping leaves and adapting to fluctuating lighting conditions.

An enhanced version of U-Net, known as RS-UNet, was proposed by¹³, utilizing ResNet50 as its backbone and integrating a Squeeze-and-Excitation (SE) attention mechanism to enhance the segmentation accuracy for early blight and anthracnose in potato leaves. This model achieved a Dice score of 88.86%, yet it faced challenges in generalizing to complex backgrounds.

Additionally,¹⁴ proposed an improved U-Net architecture for tomato leaf disease segmentation, which employed inception modules for multi-scale feature extraction and channel attention during the decoder phase. Their model reached an accuracy of 92.9%, but its high computational demands limited its applicability in real-time scenarios.

Lastly,¹⁵ introduced a two-stage segmentation framework that combined Transformer networks (TRNet) with U-Net to assess the severity of cucumber leaf diseases. This model achieved accuracies of 94.49% and 94.43% for downy mildew and

anthracnose, respectively. However, the substantial size of the model led to increased inference times.

A multi-feature fusion bottleneck pyramid U-Net (MFBP-UNet), which integrates a stable diffusion model for data augmentation, was proposed by¹⁶. This model effectively tackled segmentation issues related to pear leaf lesions with unclear boundaries, although it demanded significant computational power. Additionally,^{17,18} introduced TDSSNet, a model that employs attention modules for the segmentation of tobacco leaves, but it achieved a mere IoU of 64.99%.

Figure 2. The Sankey diagram depicts the progression segmentation methods used in detecting plant leaf diseases. Traditional techniques depend on manually crafted feature extraction methods such as GLCM, LBP, and morphological processing, along with classifiers like SVM, frequently employing clustering techniques like K-means. In contrast, contemporary methods leverage deep learning models, including CNNs, attention-based frameworks, and transformers, to produce direct segmentation results.

Our Proposed Approach for Tomato Leaf Disease Segmentation

Existing deep learning segmentation methods for detecting diseases in tomato leaves face obstacles, including a lack of adaptability to varying environmental conditions and high computational requirements. To address these challenges, we present an innovative U-Net-based segmentation model that incorporates Depthwise Separable Convolution Blocks^{19–22} Ghost Modules, Ghost Sampling^{23,24} during the upsampling process, as well as Dilated Convolutions^{25–27} in the deeper layers to improve feature extraction. In addition, we propose a Hybrid Loss Function²⁸ that combines Dice and Focal Loss²⁹ to enhance the segmentation performance.

Our model is designed to:

- Increase segmentation accuracy in intricate backgrounds.
- Boost computational efficiency through streamlined convolutional operations.
- Mitigate the effects of lighting changes and overlapping leaves.

This strategy seeks to enhance deep learning techniques for the segmentation of tomato leaf diseases, while also ensuring their effective application in real-world agricultural settings.

Materials and Methods

Dataset Acquisition and Annotation

In deep learning, particularly for segmentation models like U-Net, obtaining high-quality image-mask pairs is essential. However, creating accurate segmentation masks is challenging due to the scarcity of annotated datasets and the labor-intensive nature of manually generating masks. This research used tomato leaf images exhibiting bacterial spots from the PlantVillage dataset³⁰. A selection of sample images is presented in Figure 3.

ImageJ software, a well-regarded tool for image processing and analysis³¹, was used to generate accurate segmentation masks. This software provides various features ideal for annotating biological images, including capabilities for mask creation and editing. A related collection of the segmentation masks that have been manually annotated is presented in Figure 4.

Data Augmentation

The dataset was enhanced using multiple data augmentation techniques to boost the performance of the model. A mix of positional and color-based augmentations was used to artificially enlarge the dataset and enhance its diversity. Positional augmentations, which included flipping, rotation, and distortion, allowed the model to learn from different orientations and views of the tomato leaves. Simultaneously, color augmentations such as modifications to brightness, contrast, and channel shuffling boosted the resilience of the model to different lighting scenarios. The augmentation process was carried out using the Albumentations library, which is known for its versatility and effectiveness in image preprocessing. An illustration of the applied augmentation techniques is presented in Figure 5.

Recent developments in data augmentation methods³² have significantly enhanced model generalization. For example, employing diffusion models for data augmentation has demonstrated effective results in increasing the diversity of training datasets. Furthermore, automated augmentation techniques such as RandAugment have optimized the augmentation process by minimizing the search space, resulting in better performance across a range of tasks.

Preprocessing

To ensure efficient preprocessing, all images were loaded in the RGB format, whereas masks were imported in grayscale to enable segmentation as single-channel images. To achieve consistency, both the images and masks were normalized to a range of 0–1 and resized to dimensions of 256×256 pixels. This process resulted in a dataset containing 550 images paired with their respective 550 masks, providing a robust foundation for training the segmentation model.

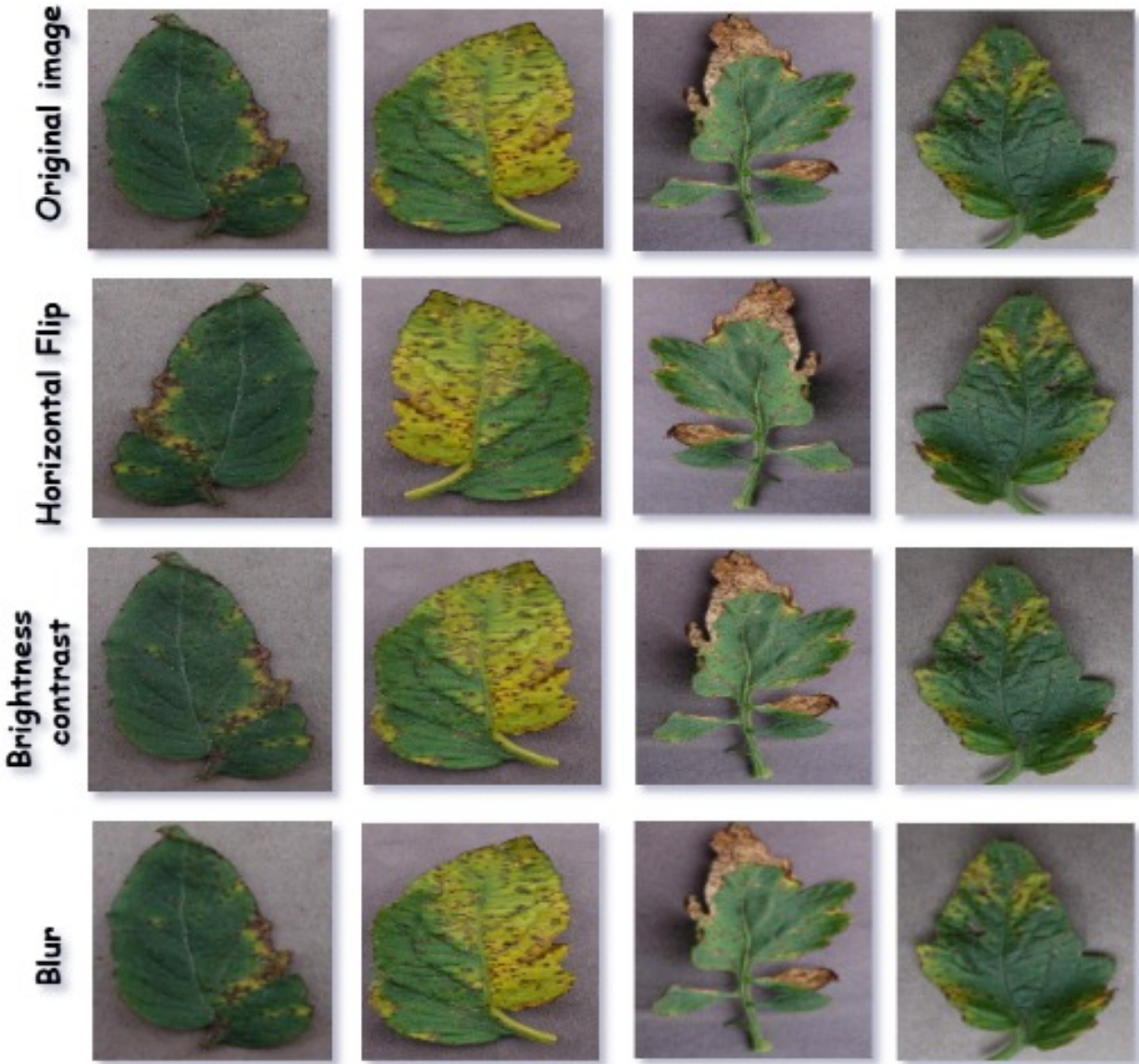


Figure 3. images after applying Augmentation techniques

Overview of the proposed DSGSU-net model

In this research, we propose DSGS-Unet, a deep learning-based architecture for the precise segmentation of tomato leaf diseases. The model is built upon the foundational U-Net³³ architecture and is enhanced with three key components: Depthwise Separable Convolution (DSConv) to improve computational efficiency, Dilated Convolutions for detecting lesions at multiple scales, and Ghost Upsampling (GSU) to facilitate lightweight reconstruction of feature maps. The architecture is organized into four distinct stages: Encoder, Bottleneck, Decoder, and Output, each of which plays a vital role in achieving precise and efficient segmentation. The detailed step-by-step methodology is presented in Algorithm 1 ⁶.

The encoder starts by analyzing the input through a sequence of initial layers, each specifically crafted to capture spatial characteristics with minimal computational demands. Depthwise Separable Convolutions (DSConv) are utilized instead of conventional convolutions, leading to a significant decrease in both the number of parameters and floating-point operations while maintaining strong spatial representation capabilities (Algorithm 1, Steps 2–3). Following each DSConv layer, Batch Normalization and ReLU activation are applied to enhance the training stability and introduce non-linearity (Steps 4–5). Max Pooling is then implemented to downsample the feature maps, which effectively broadens the receptive field (Step 6). In the lower layers of the encoder, Dilated Convolutions are integrated (Step 9). These layers use progressively larger dilation rates to

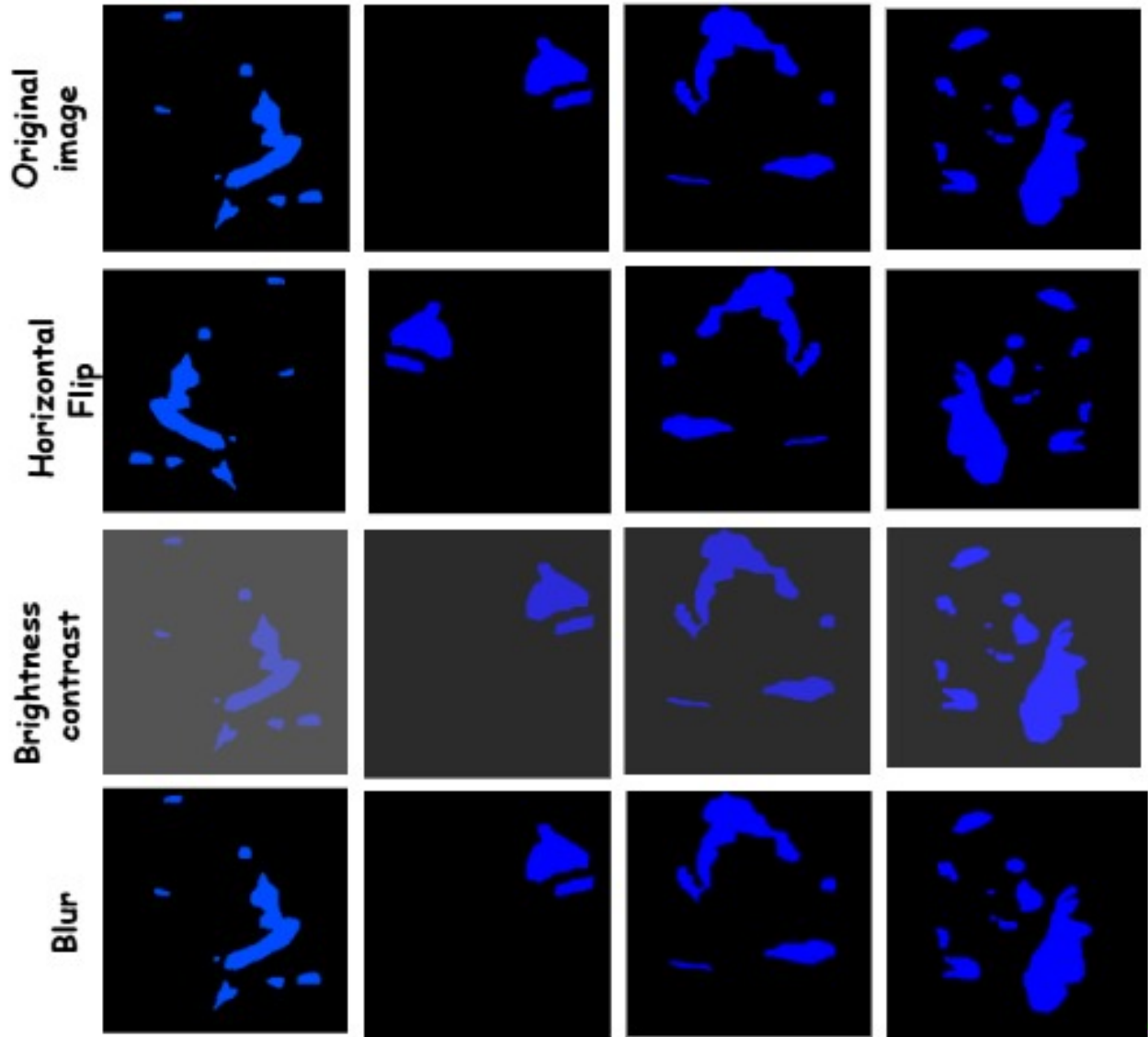


Figure 4. masks after applying Augmentation techniques

broaden the receptive field while preserving the resolution of the feature maps. Each dilated convolution is succeeded by Batch Normalization, ReLU activation, and Max Pooling to ensure normalization and enhance the learning of abstract features (Steps 10–12).

This two-tier encoder structure—employing DSConv in the upper layers and Dilated Convolutions in the lower layers—supports effective and comprehensive multiscale feature extraction (Steps 2–13). At the network’s deepest level, the bottleneck layer employs a conventional convolutional block (Step 15), followed by Batch Normalization (Step 16) and ReLU activation (Step 17). This layer captures highly abstract and compressed representations of the lesion characteristics for the subsequent decoding process (Algorithm 1, Steps 14–17). The decoder phase reconstructs the segmentation mask from the compressed feature representations through a reverse sequence of upsampling and refinement steps, which is divided into two parts:

Deep Decoder Layers (Steps 19–24): Ghost Upsampling (GSU) is used to effectively enhance the spatial resolution (Step 20). Skip connections are implemented to merge the corresponding feature maps from the encoder, thereby retaining the detailed spatial information (Step 21). The merged features are then processed through standard convolution, batch normalization, and

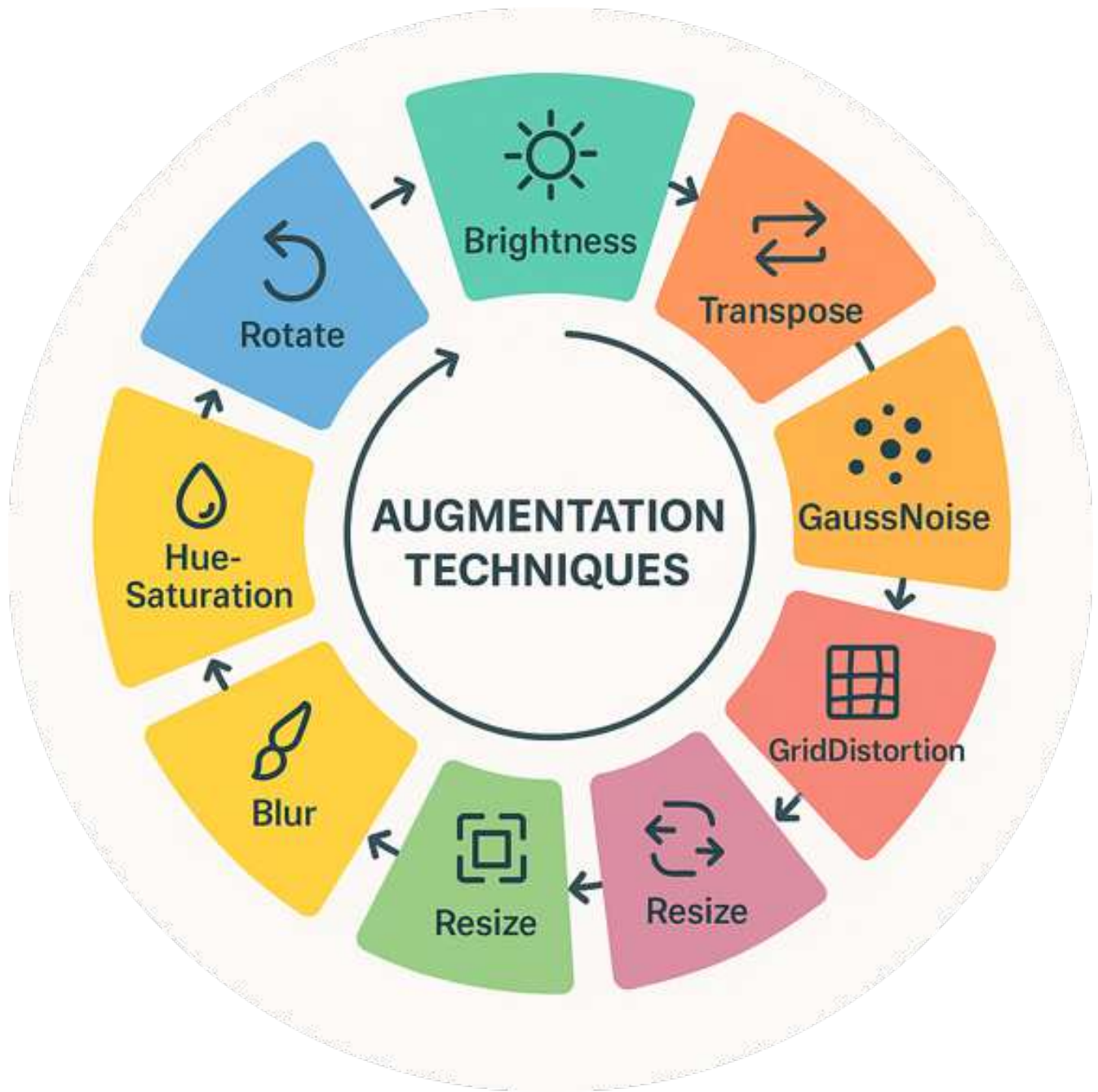


Figure 5. Augmentation techniques used on the images and masks

ReLU activation (Steps 22–24). Initial Decoder Layers (Steps 26–31): The upsampling process continues with the application of the GSU (Step 26). Feature maps from the encoder are concatenated once more (Step 27), and the resulting features undergo refinement through convolution, batch normalization, and ReLU activation (Steps 28–31). This architecture allows for gradual feature refinement while preserving critical spatial information, which is vital for accurately segmenting fine lesion boundaries (Algorithm 1, Steps 18–32). The concluding output layer (Steps 34–35) utilizes a 1×1 convolution to condense the decoded features into a single-channel output. A Sigmoid activation function is then employed to produce a binary segmentation map that emphasizes areas impacted by the disease (Algorithm 1, Steps 34–35).

Require: Preprocessed tomato leaf image
Ensure: Segmentation map highlighting disease-affected regions

- 1: **Encoder Phase:**
- 2: **for** each initial layer **do**
- 3: Perform Depthwise Separable Convolution (DSConv)
- 4: Execute Batch Normalization
- 5: Utilize ReLU activation
- 6: Implement Max Pooling
- 7: **end for**
- 8: **for** each deeper layer **do**
- 9: Conduct Dilated Convolution with increasing dilation rate
- 10: Carry out Batch Normalization
- 11: Employ ReLU activation
- 12: Apply Max Pooling
- 13: **end for**
- 14: **Bottleneck Layer:**
- 15: Execute standard Convolution
- 16: Perform Batch Normalization
- 17: Utilize ReLU activation
- 18: **Decoder Phase:**
- 19: **for** each deeper layer in reverse order **do**
- 20: Implement Ghost Upsampling (GSU)
- 21: Concatenate with corresponding encoder output (skip connection)
- 22: Conduct standard Convolution
- 23: Carry out Batch Normalization
- 24: Employ ReLU activation
- 25: **end for**
- 26: **for** each initial layer in reverse order **do**
- 27: Execute Ghost Upsampling (GSU)
- 28: Merge with corresponding encoder output (skip connection)
- 29: Perform standard Convolution
- 30: Implement Batch Normalization
- 31: Utilize ReLU activation
- 32: **end for**
- 33: **Output Layer:**
- 34: Apply 1×1 Convolution
- 35: Use Sigmoid activation to obtain the final segmentation map

Algorithm.1 Process of the proposed DSGSU-net model.

Figure 6. Proposed segmentation algorithm

Improvement of the DSGSU-net model

DSGS-Unet is an advanced deep learning framework specifically developed for the segmentation of plant diseases. As illustrated in the Figure 7, the model utilizes Depthwise Separable Convolutions (DSConv), Dilated Convolutions (DConv), and Ghost Upsampling (GSUp) as integral parts of the U-Net architecture. The following mathematical representation highlights the key elements of the model.

Depthwise Separable Convolution Block

Depthwise Separable Convolution (DSConv) offers a more efficient approach compared to traditional convolution, focusing on reducing computational costs while maintaining effective feature extraction. The DSConv structure consists of a depthwise convolution that is succeeded by a pointwise convolution. The mathematical formulation and corresponding definitions for each stage in the DSConv block, including batch normalization and ReLU activation, are presented in Table 1.

$$Y_d = X * K_d \tag{1}$$

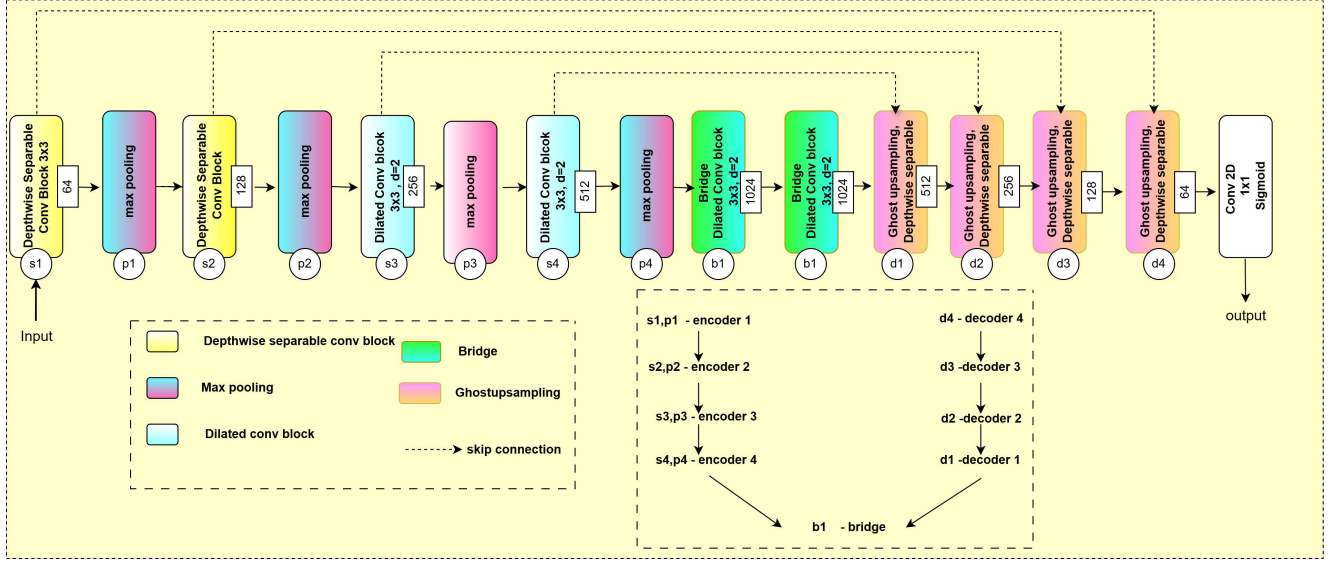


Figure 7. DSGSU-net model architecture

$$Y_p = Y_d * K_p \quad (2)$$

$$Y = \text{ReLU}(\text{BN}(Y_p)) \quad (3)$$

Table 1. Mathematical operations and symbol definitions for depthwise separable convolution with normalization and activation

Symbol / Operation	Description
X	Input feature map
K_d	Depthwise convolution kernel
$Y_d = X * K_d$	Output from depthwise convolution
K_p	Pointwise convolution kernel
$Y_p = Y_d * K_p$	Output from pointwise convolution
$Y = \text{ReLU}(\text{BN}(Y_p))$	Final output after batch normalization and ReLU activation

Dilated Convolution Block

Dilated convolutions enhance the receptive field while keeping the parameter count constant, which makes them efficient for capturing multi-scale context in segmentation tasks. A dilated convolution with a dilation factor of d alters the standard convolution process, as illustrated in Equation (4). The symbols and definitions used in the formulation are outlined in Table 2.

$$Y = \sum_{i,j} X(i + d \cdot i', j + d \cdot j') K(i', j') \quad (4)$$

Ghost Upsampling Block

Ghost Upsampling (GSUp) is used in the decoder to accurately reconstruct high-resolution segmentation maps. It produces additional feature maps by performing a depthwise convolution on the upsampled feature map. The operations involved in GSUp are defined mathematically in Equations (5)–(7), and the symbol definitions are summarized in Table 3.

Table 2. Symbol definitions for the dilated convolution operation

Symbol	Description
Y	Output feature map after dilated convolution
X	Input feature map
$K(i', j')$	Convolution kernel indexed by (i', j')
d	Dilation rate controlling the spacing between kernel elements
(i, j)	Spatial indices over the input feature map
(i', j')	Spatial indices over the convolution kernel

$$Y_u = \text{Upsample}(X) \quad (5)$$

$$Y_g = \text{ReLU}(\text{BN}(\text{DepthwiseConv}(Y_u))) \quad (6)$$

$$Y = \text{Concat}(Y_u, Y_g) \quad (7)$$

Table 3. Symbol definitions for the Ghost Upsampling (GSUp) process

Symbol	Description
X	Input feature map to the upsampling block
$Y_u = \text{Upsample}(X)$	Feature map after transposed or nearest-neighbor upsampling
$Y_g = \text{ReLU}(\text{BN}(\text{DepthwiseConv}(Y_u)))$	Ghost feature map generated via depthwise convolution, BN, and ReLU
$Y = \text{Concat}(Y_u, Y_g)$	Final upsampled feature map formed by concatenating original and ghost features

Encoder-Decoder Architecture

The encoder comprises a sequence of meticulously optimized convolutional blocks. This design integrates Depthwise Separable Convolutions (DSConv) in the initial phases and employs Dilated Convolutions in the subsequent layers to improve efficiency and expand feature representation.

$$S_l, P_l = f_{\text{Enc}}(X_l) \quad (8)$$

In this context, S_l represents the output from the l -th encoder block, while P_l refers to the downsampled feature map that is carried forward to the subsequent layer.

Encoder Structure and Design Rationale

As depicted in Figure 8(a), Encoder Blocks 1 and 2 use Depthwise Separable Convolutions instead of traditional convolutions. Every block is organized with a depthwise convolution followed by batch normalization and ReLU activation, a pointwise convolution (1×1), additional batch normalization, and a max-pooling layer. This design greatly reduces the number of parameters and computational requirements while efficiently capturing the spatial features. The spatial resolution is reduced by half after each encoder block due to maximum pooling (i.e., spatial size $\rightarrow$ spatial size / 2), facilitating hierarchical feature learning.

As we delve further into the network, we encounter Encoder Block 3, as depicted in Figure 8(b), and Encoder Block 4, shown in Figure 8(c), which employ dilated convolutions with dilation rates of 2 and 4, respectively. These blocks improve the receptive field while maintaining the spatial resolution during the initial encoding phases. The enlarged receptive field enables the model to grasp a wider array of global and contextual features, which are crucial for detecting subtle disease patterns. This hierarchical framework employs Depthwise Separable Convolutions in the early layers and Dilated Convolutions in the later layers, effectively balancing the retention of local details with the improvement of global feature learning. This methodology is especially beneficial for agricultural segmentation tasks, where lesions may appear at different scales and intensities.

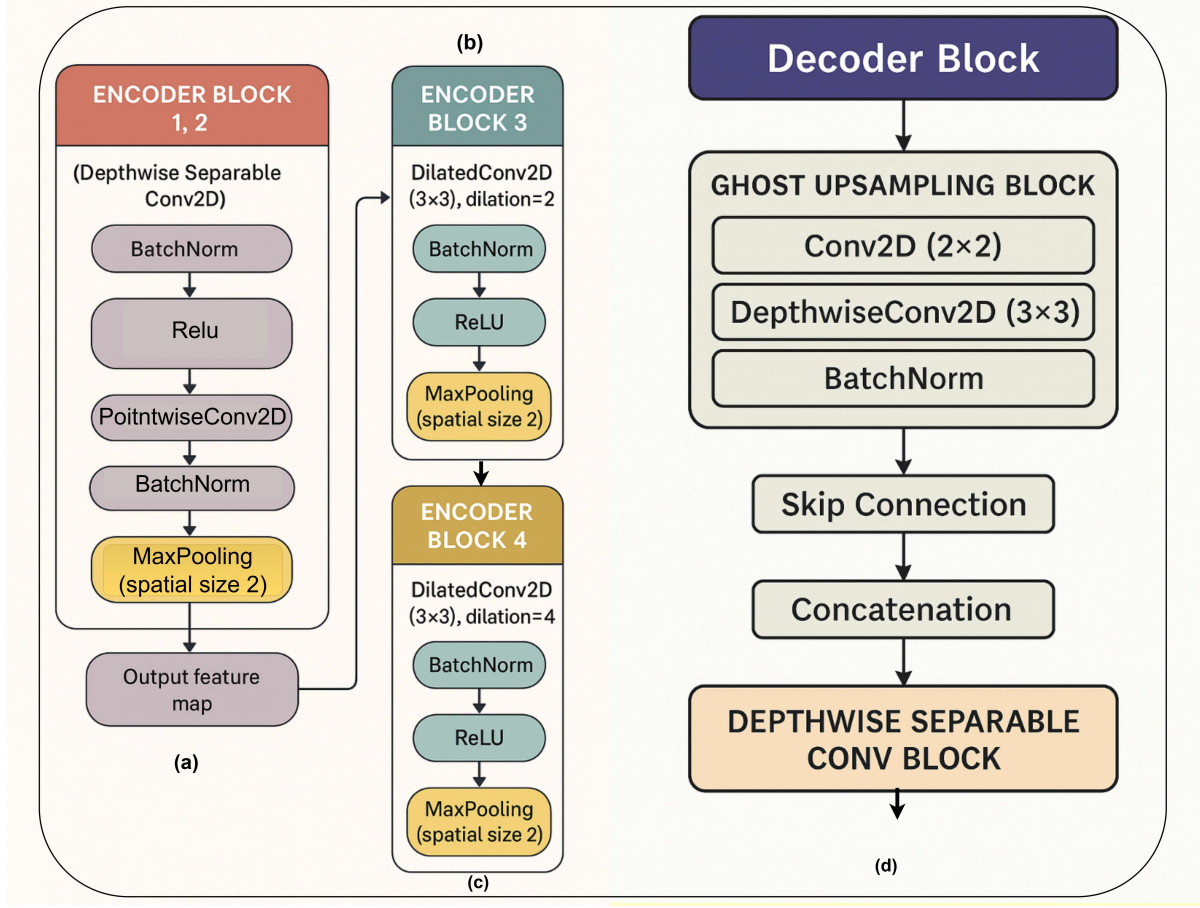


Figure 8. DSGSU-net model encoder decoder blocks

Decoder structure and design

As depicted in Figure 8(d), the Decoder Block is structured to incrementally restore the spatial resolution while maintaining the semantic information acquired during the encoding phase. Each decoder block initiates with a Ghost Upsampling Block, which comprises a 2×2 transposed convolution, followed by a 3×3 depthwise convolution, batch normalization, and a ReLU activation function. This arrangement facilitates effective upsampling with reduced computational demands.

After the upsampling process, a skip connection from the corresponding encoder block is integrated to strengthen the fine-grained spatial details that may have been diminished during downsampling. The combined features are subsequently processed through a Depthwise Separable Convolution Block, which incorporates both depthwise and pointwise convolutions to enhance the refined features. This integration improves the spatial localization while ensuring that the model remains efficient.

The decoder employs Ghost Upsampling blocks to reconstruct the segmentation map from the encoded features.

$$D_l = f_{\text{Dec}}(D_{l+1}, S_l) \quad (9)$$

In this context, D_l denotes the output from the l -th decoder block, while S_l signifies the skip connection from the associated encoder layer. Ghost Upsampling enables the efficient reconstruction of high-resolution segmentation maps by leveraging the inherent redundancy of features.

Output Layer

The final segmentation map is produced using a 1×1 convolution with a sigmoid activation function:

$$Y_{\text{out}} = \sigma(\text{Conv}_{1 \times 1}(D_1)) \quad (10)$$

In this formulation, $\sigma(x) = \frac{1}{1+e^{-x}}$ ensures pixel-wise classification.

Loss Function

In the field of plant disease segmentation, accurate pixel-level classification is crucial, especially for small and irregularly shaped disease-affected areas. Traditional loss functions such as Dice Loss and Focal Loss each present limitations when used independently. Dice Loss assesses the similarity between the predicted and actual masks but may struggle with small lesion targets, leading to training instability. On the other hand, Focal Loss effectively addresses class imbalance by focusing on difficult-to-classify pixels. To leverage the strengths of both methods, we implement a hybrid loss function that combines Dice Loss and Focal Loss, defined as follows:

$$L_{\text{total}} = \alpha L_D + \beta L_F \quad (11)$$

where L_D represents the Dice Loss, L_F denotes the Focal Loss, and α and β are weight coefficients, both set to 1 in our experiments.

Dice Loss

Dice Loss, which originates from the Dice coefficient, is frequently utilized in image segmentation tasks to address class imbalance. Its definition is as follows:

$$L_D = 1 - \frac{2 \sum_{i=1}^N p_i g_i}{\sum_{i=1}^N p_i + \sum_{i=1}^N g_i} \quad (12)$$

In this context, p_i represents the predicted probability, while g_i indicates the actual label for pixel i . Here, N refers to the total number of pixels within the image. Dice Loss improves segmentation precision by increasing the overlap between predicted and actual areas; however, it does not effectively emphasize smaller or more difficult regions to classify.

Focal Loss

Focal Loss is an enhancement of the cross-entropy loss designed to tackle class imbalance by giving more importance to challenging pixels that are difficult to classify. It is defined as

$$L_F = -\lambda_t (1 - q_t)^\gamma \log(q_t) \quad (13)$$

In this context, q_t represents the predicted probability of the correct class, while λ_t serves as a balancing factor that modifies the weight assigned to the foreground and background classes. Additionally, γ acts as a focusing parameter that regulates the intensity of the weighting effect. An increased value of γ places greater emphasis on misclassified pixels, thereby improving the segmentation performance for regions affected by disease that are either small or exhibit low contrast.

Evaluation Metrics for Plant Disease Segmentation

Evaluation Metrics

To thoroughly evaluate the segmentation performance, we used various metrics such as the Dice Coefficient, Intersection over Union (IoU), F1 Score, Precision, Recall, and Accuracy. These metrics evaluate how well the predicted segmentation masks align with the ground truth, offering a reliable measure of the model accuracy in identifying diseased regions on plant leaves.

Dice Coefficient(DICE)

The Dice Coefficient assesses the resemblance between the predicted and actual segmentation masks by quantifying the overlap in the following manner:

$$\text{DICE} = \frac{2|P \cap G|}{|P| + |G|} \quad (14)$$

where P and G represent the predicted and ground truth segmentation masks, respectively. To ensure numerical stability, a small smoothing factor ϵ is added:

$$\text{DICE} = \frac{2 \sum (Y_{\text{true}} \cdot Y_{\text{pred}}) + \epsilon}{\sum Y_{\text{true}} + \sum Y_{\text{pred}} + \epsilon} \quad (15)$$

where Y_{true} and Y_{pred} denote the actual and predicted binary values. A threshold of **0.5** is applied to convert predicted probabilities into binary values before computation.

Intersection over Union (IoU)

IoU measures the degree of overlap between the predicted and ground truth masks:

$$\text{IoU} = \frac{|P \cap G|}{|P \cup G|} \quad (16)$$

A higher IoU indicates enhanced segmentation accuracy, signifying a greater overlap between the predicted and actual regions impacted by the disease.

F1 Score

F1 Score represents the harmonic mean of Precision and Recall, effectively balancing the trade-off between false positives and false negatives.

$$\text{F1 Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (17)$$

Precision Precision refers to the ratio of accurately identified diseased pixels to the total number of pixels predicted as diseased.

$$\text{Precision} = \frac{TP}{TP + FP} \quad (18)$$

Recall

Recall represents the proportion of accurately identified diseased pixels out of the total number of actual diseased pixels.

$$\text{Recall} = \frac{TP}{TP + FN} \quad (19)$$

Accuracy

Accuracy quantifies the ratio of accurately identified pixels within the segmentation map.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (20)$$

TP (True Positives) refers to the pixels that are accurately recognized as diseased, TN (True Negatives) denotes the pixels that are correctly identified as non-diseased, FP (False Positives) indicates healthy pixels that have been mistakenly classified, FN (False Negatives) represents diseased areas that have not been detected.

Experimental environment and parameter settings

The DBDPU-Net architecture was used for segmenting plant leaf diseases. This implementation was executed on a system with 32 GB of RAM and an NVIDIA RTX A4000 GPU, using Python version 3.4 in the Jupyter Notebook environment. The operating system was Windows 11 64-bit. Python was the programming language employed, and the Adam optimizer was implemented with a starting learning rate of 0.001, completing 50 training epochs.

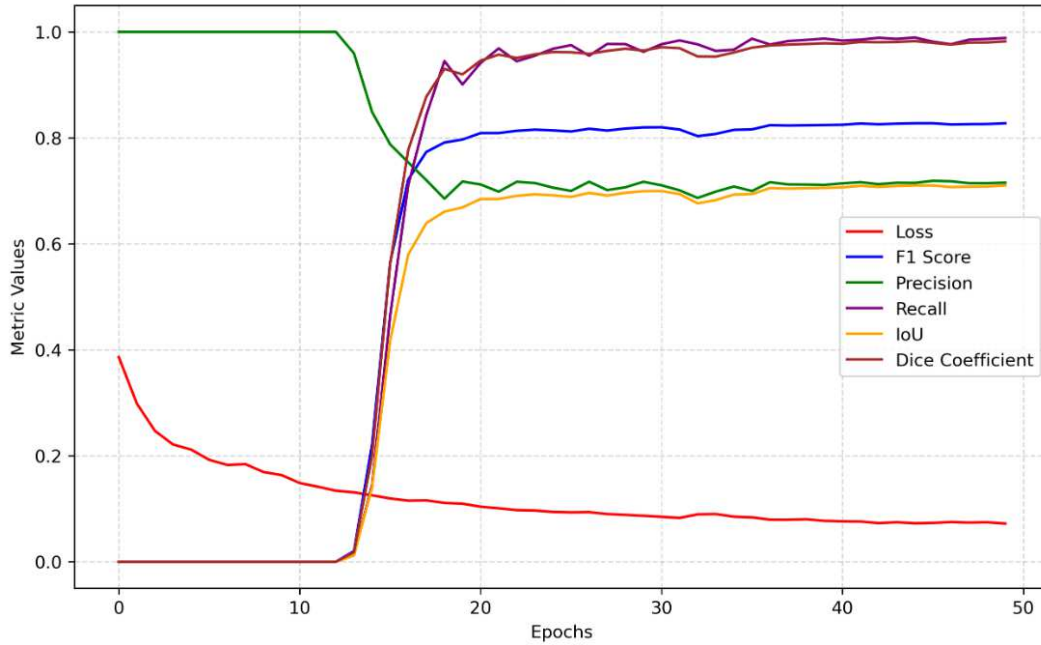
Results and Discussions

Quantitative Results

We assessed the effectiveness of DSGSU-Net by employing various standard segmentation metrics: Accuracy, Precision, Recall, F1-Score, Intersection over Union (IoU), and Dice Coefficient. Figure 9 and Table 4 provide a summary of the outcomes achieved by the DSGSU-net model. The model attains a high recall (0.9885), suggesting that it effectively identifies nearly all diseased regions. Although the precision was marginally reduced (0.7156), the balance of precision and recall resulted in a robust F1-score and an excellent Dice Coefficient, confirming accurate pixel-wise segmentation and effective boundary definition.

Table 4. Results of DSGSU-Net Model

Metrics	DSGSU-Net
Accuracy	0.9572
Precision	0.7156
Recall	0.9885
F1-Score	0.8276
IoU	0.7102
Dice Coefficient	0.9822

**Figure 9.** DSGSU-net model various metrics values

Comparison with Baseline Models

We conducted a comparative analysis of DSGSU-Net's performance against baseline architectures such as U-Net, U-Net++, SegNet, and PSPNet. The results, presented in Figure 10, demonstrate that DSGSU-Net outperforms these models across diverse segmentation metrics.

Visual Analysis

Figure 11 illustrates visual comparisons between the ground truth and the predictions made by U-Net and DSGSU-Net. The predictions from DSGSU-Net exhibit more defined boundaries, reduced over-segmentation, and a lower incidence of false positives when compared with the baseline. The incorporation of Ghost Upsampling has effectively preserved intricate details in leaf boundaries, while Dilated Convolutions have enhanced the contextual understanding of dispersed disease patterns. In addition, depthwise separable convolutions have facilitated these enhancements without imposing a significant computational load.

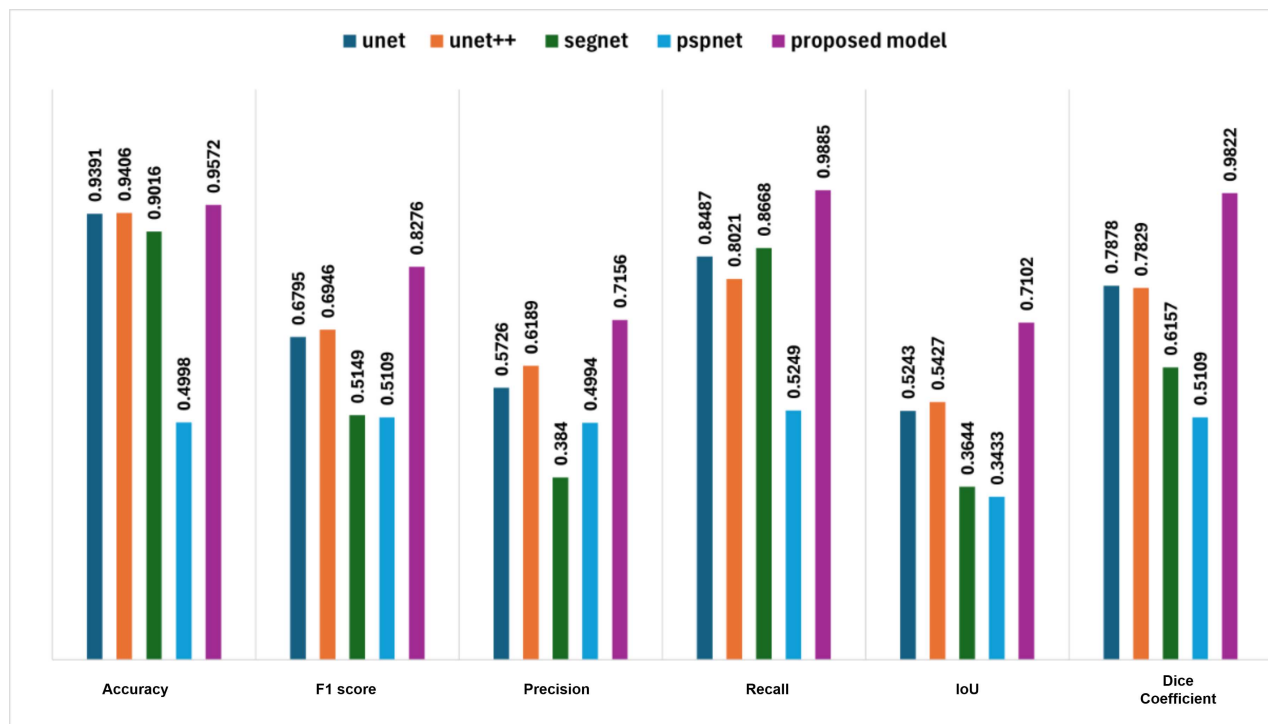


Figure 10. comparative analysis with the baseline models

Visual Analysis

Computational Performance Analysis

Figure 12. An evaluation of the computational efficiency among different U-Net variants was performed by analyzing the total, trainable, and non-trainable parameters, as outlined in Table X. The use of Depthwise Convolutions, Ghost Modules, Dilated Convolutions, and Ghost Upsampling significantly affects the model complexity, leading to a reduction in the parameter count while maintaining or improving the segmentation accuracy. The baseline U-Net variant has a total of 310.55M parameters (310.43M of which are trainable), making it computationally demanding. In contrast, incorporating Depthwise Convolutions reduced the total parameter count to 178.55M, greatly enhancing the efficiency. Furthermore, adding Ghost Modules increases the total to 208.70M parameters, achieving a favorable trade-off between complexity and representational power. The U-Net+Depthwise+Dilated+GhostUpsampling model emerges as the most efficient setup, offering an excellent compromise between computational efficiency and segmentation accuracy. With 195.44 million parameters (74.55MB) and 19.52 million trainable parameters (74.49MB), it significantly reduces computational costs while outperforming other models in key segmentation metrics such as IoU, Dice Coefficient, and F1 Score. Consequently, the U-Net model, augmented with Depthwise Separable Convolutions, Dilated Convolutions, and Ghost Upsampling, is identified as the optimal architecture. Its balanced parameter count, coupled with exceptional segmentation performance, makes it an ideal choice for applications that demand both efficiency and accuracy.

Ablation study

To evaluate the impact of different architectural modifications on the segmentation performance, we conducted a series of ablation studies. We started with the baseline U-Net model and progressively integrated Depthwise Separable Convolutions, Ghost Modules, Ghost Upsampling, and Dilated Convolutions, while also experimenting with various loss functions such as Binary Cross Entropy (BCE), Focal Loss, Dice Loss, and a Hybrid BCE + Dice Loss. The replacement of standard convolutions with Depthwise Separable Convolutions resulted in a notable decrease in the total parameters, reducing from 31M in the U-Net to 17.8M, thus compressing the model while maintaining comparable accuracy (train accuracy 0.7943 after 100 epochs). However, after just 50 epochs, the accuracy dropped slightly to 0.7681, indicating that depthwise convolutions may necessitate additional training iterations for optimal performance. The incorporation of Ghost Modules further enhanced the efficiency of feature extraction while keeping the parameter count low. The U-Net + Depthwise + Ghost model achieved a superior accuracy of 0.7963 compared to the U-Net baseline. Nonetheless, substituting BCE with Focal Loss resulted in a slight performance decline (0.787), suggesting that Ghost Modules are more effective with Binary Cross Entropy in this segmentation scenario.

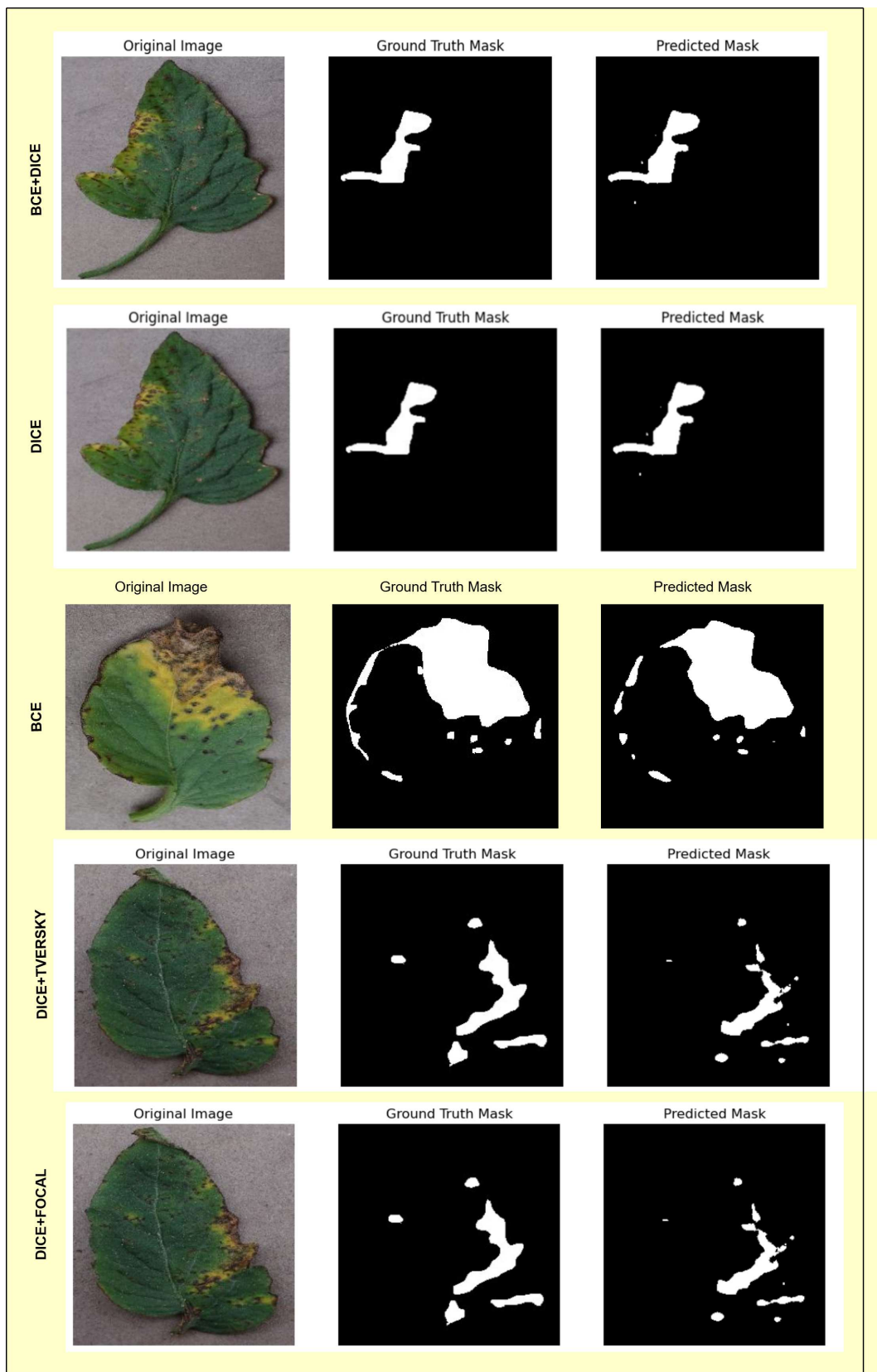


Figure 11. segmentation results of different loss functions

Table 5. Quantitative comparison of segmentation performance metrics for U-Net model variants using various loss functions

Model Variants	Epochs	Loss Function	Accuracy	Final Loss	F1 Score	Precision	Recall	IoU	Dice coefficient
unet + depthwise	50	BCE	0.9457	0.0547	0.7395	0.7134	0.7859	0.5998	0.8304
unet + depthwise	50	FOCAL	0.9264	0.0084	0.4641	0.7482	0.4374	0.3647	0.4971
unet + depthwise + ghost	50	BCE	0.9488	0.0528	0.7677	0.7164	0.8360	0.6313	0.8681
unet + depthwise + ghost	50	FOCAL	0.9380	0.0084	0.6627	0.7278	0.6291	0.5110	0.7037
unet + depthwise + ghost upsampling	50	BCE	0.9483	0.0551	0.7594	0.7035	0.8396	0.6223	0.8612
unet + depthwise + ghost upsampling	50	FOCAL	0.9331	0.0054	0.5798	0.7074	0.5779	0.4543	0.6441
unet + depth + dilated + ghost upsampling	50	BCE	0.9523	0.0399	0.7967	0.6881	0.9542	0.6688	0.9410
unet + depth + dilated + ghost upsampling	50	DICE	0.9563	0.0244	0.8181	0.7084	0.9751	0.6973	0.9684
unet + depth + dilated + ghost upsampling	50	BCE + DICE	0.9550	0.0431	0.8017	0.6820	0.9837	0.6749	0.9543
unet + depth + dilated + ghost upsampling	50	DICE + FOCAL	0.9572	0.0150	0.8276	0.7156	0.9885	0.7102	0.9822

By introducing Ghost Upsampling, we further decreased the total parameters to 15.2M while attaining competitive accuracy (0.7898 with BCE and 0.7884 with Focal Loss). This demonstrates that Ghost Upsampling effectively streamlines the model without significantly sacrificing accuracy, making it a valuable enhancement for efficient segmentation models. The addition of Dilated Convolutions in the deeper layers improved the spatial feature extraction, albeit with a minor increase in the parameters to 19.5 M. Among the various models evaluated, the highest-performing variant of this architecture employed Dice + Focal

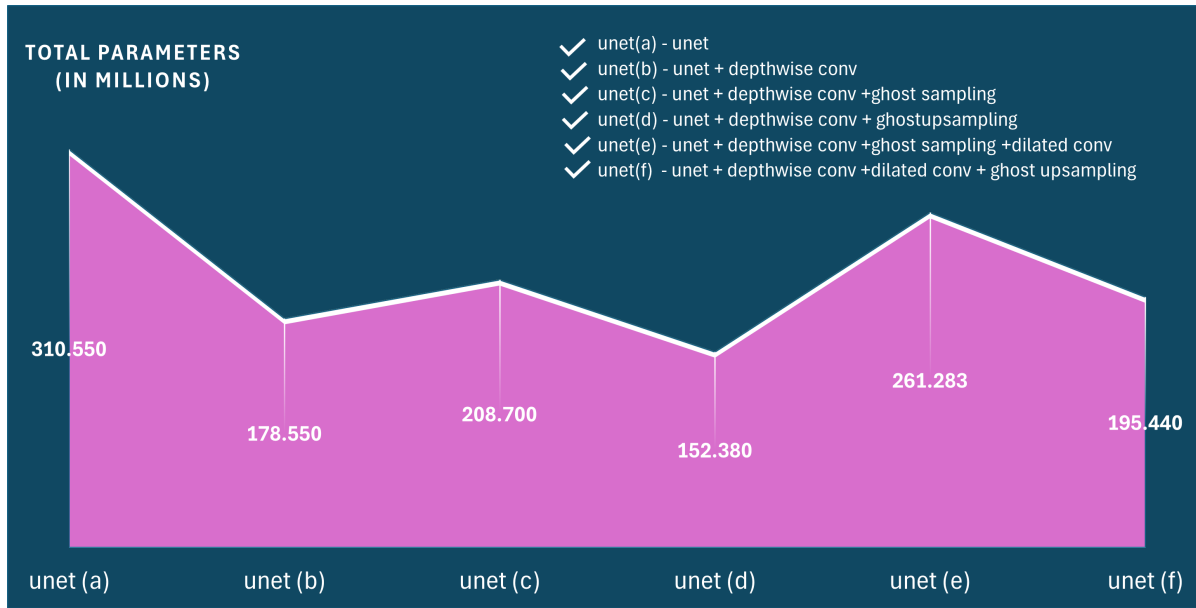


Figure 12. computational graph

Loss, achieving a train accuracy of 0.7937.

This suggests that the use of Dilated Convolutions in conjunction with Hybrid Loss leads to better boundary segmentation and increased robustness. The gradual enhancements indicate that the depthwise separable convolutions and ghost modules improve the efficiency without compromising the segmentation quality. Furthermore, Ghost Upsampling greatly reduces computational costs, while Dilated Convolutions help in capturing contextual information. The integration of Dice and Focal Loss has proven beneficial for performance, achieving the highest accuracy of 0.7937 in the final configuration. Table 5 shows the results of the segmentation performance.

Limitations and Future Work

While DSGSU-Net demonstrates robust performance, it has certain drawbacks that need to be addressed. The model sometimes produces false positives in areas with intricate textures, such as distinct leaf veins or shadows. Although it maintains a high recall rate, its precision falls short compared with that of U-Net++, suggesting a propensity to over-segment healthy regions. Furthermore, DSGSU-Net is limited to binary segmentation and does not differentiate between various disease types. Its assessment is based on the PlantVillage dataset, which features pristine, laboratory-captured images, thereby restricting its applicability to real-world scenarios with complex backgrounds and diverse lighting conditions. Future developments will aim to incorporate attention mechanisms to reduce false activations, utilize post-processing methods like Conditional Random Fields (CRF) or edge refinement to enhance boundary precision, and extend the model for multi-class and multi-label segmentation. Additionally, adapting the model for real-world field images and optimizing it for lightweight deployment through techniques such as quantization and pruning will significantly improve its usability in mobile and edge computing settings.

Conclusion

In this research, we introduced DSGSU-Net, an advanced architecture based on U-Net, specifically tailored for the accurate segmentation of tomato leaf diseases. By incorporating Depthwise Separable Convolutions, Dilated Convolutions, and ghost-sampling techniques, the model achieves a commendable balance between high segmentation accuracy and computational efficiency. A specially designed hybrid loss function that merges Dice and Focal Loss further enhanced the ability of the model to address class imbalance and accurately capture fine lesion boundaries. Comprehensive assessments on the PlantVillage dataset revealed that DSGSU-Net surpasses traditional models such as U-Net and U-Net++ across various metrics, attaining a Dice coefficient of 0.9822, an IoU of 0.7102, and a recall of 0.9885. The ablation study validated the significance of each architectural improvement, while the visual results showcased the proficiency of the model in precisely delineating intricate and subtle lesion patterns. However, the model occasionally misclassifies textured healthy regions and is currently restricted to binary segmentation. Future research will aim to enhance lesion classification through the implementation of attention mechanisms, post-processing methods, real-world adaptation, and strategies for more lightweight deployment. Overall, DSGSU-

Net represents a promising and effective approach for automated disease segmentation in precision agriculture, facilitating early disease detection and informed crop management.

Data Availability

The datasets generated and analyzed during the current study are available in the Kaggle repository: <https://www.kaggle.com/datasets/charuchaudhry/plantvillage-tomato-leaf-dataset>

References

1. Nyasulu, L., Munthali, C. & Banda, M. Classification of tomato fungal leaf diseases using texture-based features and machine learning algorithms. *Bioinformation* **19**, 896–902, DOI: [10.6026/97320630019896](https://doi.org/10.6026/97320630019896) (2023).
2. Prem Kumar, K. & Soori, R. Real-time tomato leaf disease detection system using glcm and color features with svm. *J. Surv. Fish. Sci.* **10**, 2138–2152 (2023).
3. Alhwaiti, M., Alnuaimi, N., Saba, T. & Rehman, A. Early detection of tomato late blight using hog feature extraction and svm classifier. *arXiv preprint arXiv:2306.08326* (2023).
4. Jamjoom, M. A., Khan, R. A., Khan, S., Alaskar, H. & Shakya, D. Tomato leaf disease detection and classification using hybrid feature extraction techniques and machine learning algorithms. *J. King Saud Univ. - Comput. Inf. Sci.* **35**, 7661–7672, DOI: [10.1016/j.jksuci.2023.01.031](https://doi.org/10.1016/j.jksuci.2023.01.031) (2023).
5. Jamjoom, M., Elhadad, A., Abulkasim, H. & Abbas, S. Plant leaf diseases classification using improved k-means clustering and svm algorithm for segmentation. *Comput. Mater. & Continua* **76** (2023).
6. Thimmegowda, T. G. M. & Jayaramaiah, C. Cluster-based segmentation for tobacco plant detection and classification. *Bull. Electr. Eng. Informatics* **12**, 75–85 (2023).
7. James, S. & Punitha, S. Detection and classification of plant leaf diseases using pso optimized svm classifier. *Int. J. Comput. Appl.* **144**, 25–29, DOI: [10.5120/ijca2016910270](https://doi.org/10.5120/ijca2016910270) (2016).
8. Saraf, R., Raut, S. & More, A. Wavelet based image segmentation for leaf disease detection using otsu thresholding. *Eudoxus Press.* **5**, 101–109 (2024).
9. Shao, Y. *et al.* Research on automatic identification system of tobacco diseases. *The Imaging Sci. J.* **65**, 252–259 (2017).
10. Zulkifley, M. A., Moubark, A. M., Saputro, A. H. & Abdani, S. R. Automated apple recognition system using semantic segmentation networks with group and shuffle operators. *Agriculture* **12**, 756 (2022).
11. Yuan, H., Zhu, J., Wang, Q., Cheng, M. & Cai, Z. An improved deeplab v3+ deep learning network applied to the segmentation of grape leaf black rot spots. *Front. plant science* **13**, 795410 (2022).
12. Zhang, W., Li, M. & Zhao, H. Adaptive multilevel attention deeplabv3+ with heuristic-based frame selection for semantic segmentation. *Inf. Process. Agric.* **12**, 45–56, DOI: [10.1016/j.inpa.2024.01.019](https://doi.org/10.1016/j.inpa.2024.01.019) (2025).
13. Chen, Z. *et al.* Md-unet for tobacco leaf disease spot segmentation based on multi-scale residual dilated convolutions. *Sci. Reports* **15**, 2759 (2025).
14. Xiaohu, Z., Xiao, L., Sheng, Y., Wei, F. & Xingyi, Y. Multi-scale tomato disease segmentation algorithm based on improved u-net network. *J. Comput. Eng. & Appl.* **58** (2022).
15. Yao, H. *et al.* A cucumber leaf disease severity grading method in natural environment based on the fusion of trnet and u-net. *Agronomy* **14**, 72 (2023).
16. Fu, J., Li, X.-x., Chen, F.-h. & Wu, G. Pear leaf disease segmentation method based on improved deeplabv3+. *Cogent Food & Agric.* **10**, 2310805 (2024).
17. Ou, J. *et al.* Tobacco leaf disease segmentation based on tdssnet. In *2023 Twelfth International Conference on Image Processing Theory, Tools and Applications (IPTA)*, 1–5 (IEEE, 2023).
18. Chen, Z. *et al.* Md-unet for tobacco leaf disease spot segmentation based on multi-scale residual dilated convolutions. *Sci. Reports* **15**, 2759 (2025).
19. Yu, Y., Zhang, Y., Yu, J. & Yue, J. Lightweight decoder u-net crack segmentation network based on depthwise separable convolution. *Multimed. Syst.* **30**, 295, DOI: [10.1007/s00530-024-01509-3](https://doi.org/10.1007/s00530-024-01509-3) (2024).
20. Hasan, M. A. & Dey, K. Depthwise separable convolutions with deep residual convolutions. *arXiv preprint arXiv:2411.07544* (2024).

21. Wang, B. *et al.* Dsml-unet: Depthwise separable convolution network with multiscale large kernel for medical image segmentation. *Biomed. Signal Process. Control.* **97**, 106731 (2024).
22. Wang, X., Liu, Y., Chen, L. & Zhao, P. Dsml-unet: Depthwise separable convolution network with multiscale learning for medical image segmentation. *Biomed. Signal Process. Control.* **87**, 104789, DOI: [10.1016/j.bspc.2024.104789](https://doi.org/10.1016/j.bspc.2024.104789) (2024).
23. Han, K. *et al.* Ghostnet: More features from cheap operations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2020).
24. Zhang, W., Li, M. & Zhao, H. Msgu-net: A lightweight multi-scale ghost u-net for image segmentation. *Front. Neurobotics* **18**, 1480055, DOI: [10.3389/fnbot.2024.1480055](https://doi.org/10.3389/fnbot.2024.1480055) (2024).
25. Ma, W., Lorenzana, M. B., Dai, W., Sun, H. & Chandra, S. S. Multi-scale mri reconstruction via dilated ensemble networks. *arXiv preprint arXiv:2310.04705* (2023).
26. Li, X., Wang, Y., Zhang, H. & Liu, Q. Ground-based cloud image segmentation method based on improved u-net. *Appl. Sci.* **14**, 11280, DOI: [10.3390/app142311280](https://doi.org/10.3390/app142311280) (2024).
27. Xu, C., Yu, C. & Zhang, S. Lightweight multi-scale dilated u-net for crop disease leaf image segmentation. *Electronics* **11**, 3947 (2022).
28. Chen, Y., Zhao, L. & Wang, Q. Hybrid loss functions for addressing class imbalance in medical image segmentation. *Med. Image Analysis* **85**, 102345, DOI: [10.1016/j.media.2023.102345](https://doi.org/10.1016/j.media.2023.102345) (2023).
29. Li, X. *et al.* Generalized focal loss for dense object detection. *IEEE Transactions on Pattern Analysis Mach. Intell.* **45**, 3673–3688, DOI: [10.1109/TPAMI.2022.3183135](https://doi.org/10.1109/TPAMI.2022.3183135) (2023).
30. Chaudhry, C. Plantvillage tomato leaf dataset. <https://www.kaggle.com/datasets/charuchaudhry/plantvillage-tomato-leaf-dataset> (2025). Accessed: 2025-04-16.
31. ImageJ Development Team. ImageJ 1.54h Release Notes. <https://imagej.net/ij/notes.html> (2023). Accessed: 2025-04-16.
32. Alimisis, P., Mademlis, I., Radoglou-Grammatikis, P., Sarigiannidis, P. & Papadopoulos, G. T. Advances in diffusion models for image data augmentation: a review of methods, models, evaluation metrics and future research directions. *Artif. Intell. Rev.* **58**, DOI: [10.1007/s10462-025-11116-x](https://doi.org/10.1007/s10462-025-11116-x) (2025).
33. Ronneberger, O., Fischer, P. & Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)* (2015).

Acknowledgements

This work was supported by the Vellore Institute of Technology, Vellore

Author contributions statement

As this is part of a Ph.D. research project, the primary work was conducted by Miss Jansi E under the supervision of Prof. Kavitha B R. All authors contributed to the review of the manuscript.

Funding

Open access funding provided by Vellore Institute of Technology.

Declarations

Competing interests

The authors declare no competing interests.