RSA-DeRefNet: A Hybrid Residual-Dense Network with Regularization Self-Attention

Prof. Samit Ari
Electronics and Communication Engg.
NIT Rourkela
Rourkela, India.
email address or ORCID

Aditya Nayak
Electronics and Communication Engg.
NIT Rourkela
Rourkela, India.
email address or ORCID

Ayush Kumar Samal
Electronics and Communication Engg.
NIT Rourkela
Rourkela, India.
email address or ORCID

Mohit ranjan Naik
Electronics and Communication Engg.
NIT Rourkela
Rourkela, India.
email address or ORCID

Electronics and Communication Engg.
NIT Rourkela
Rourkela, India.
email address or ORCID

6th Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

***Abstract*—The rise of plant diseases poses a major threat to global food security. Effective automated detection systems are essential for timely intervention, highlighting the need for reliable systems that can detect problems early. To meet this challenge, we introduce RSA-DerefNet, a new hybrid deep learning architecture aimed at classifying plant diseases based on images. Our model combines two effective network types: residual networks, which help train very deep models by addressing gradient issues, and densenets, which improve information flow through feature reuse. The central part of our approach is a custom Regularized Self-Attention (RSA) module integrated into both network paths. This feature allows the model to focus on important visual signs linked to diseases, like lesion patterns or discoloration, while reducing irrelevant background noise. Evaluating RSA-DerefNet on a wide-ranging dataset of crop diseases shows that it performs significantly better, confirming the value of our architectural design for achieving high-precision classification.**

I. Introduction

Maintaining agricultural yield is a global necessity linked to feeding a growing population. However, this goal faces ongoing threats from harmful plant diseases. These diseases, caused by various agents such as fungi, bacteria, and viruses, as well as factors such as nutritional deficiencies, can severely damage crops, leading to substantial financial losses for farmers and compromising food security in regions. Traditionally, detecting and diagnosing plant diseases has been slow and labor intensive, relying on the expertise of agronomists and agricultural extension workers. Manual inspection is subjective, prone to human mistakes, and cannot scale up for large commercial farms. In addition, early disease symptoms are often tiny or subtle, making them easy to overlook, which delays action and allows diseases to spread quickly. The large scale and complexity of this problem show the urgent need for accurate and automated solutions for diagnosing crop diseases. Early automation efforts relied on classical computer vision techniques. This process involved a multistep pipeline in which researchers manually extracted features from plant images. They used techniques like histogram of oriented gradients (HOG), Scale-Invariant Feature Transform (SIFT), and various texture and color-based descriptors to create fixed-length feature vectors. Classifiers, such as Support Vector Machines (SVM) or k-Nearest Neighbors (k-NN), were then trained on these features to tell healthy plants from diseased ones. Although these methods proved foundational, their main drawback was the dependence on hand-made features. These features are often fragile and do not perform well against real-world changes in lighting, noise, leaf orientation, and natural plant variations. They struggle to generalize beyond specific conditions, making them less effective in diverse fields.

The rise of deep learning, especially through Convolutional Neural Networks (CNNs), transformed this area. CNNs can learn powerful hierarchical features directly from raw images, eliminating the need for manual feature engineering. Early CNN architectures such as AlexNet and VGGNet achieved groundbreaking results in image classification tasks. However, applying these models to agriculture revealed new challenges. Many of these models were deep and loaded with parameters, leading to high computational costs. This complexity makes them unsuitable for resource-limited devices at the farm level, such as drones and mobile phones that are vital for real-time applications. Moreover, while CNNs improved accuracy, they often acted as ”black boxes.” It remains difficult to understand which visual clues, such as specific lesion patterns, subtle leaf color changes, or background textures, the model uses for its decisions. For farmers or agronomists, knowing why a disease was identified is crucial to inform management strategies. This lack of interpretability and the challenge of overfitting certain datasets, which prevents generalization to new environments, are still significant issues in current research.

This paper proposes RSA-DerefNet, a new hybrid deep learning architecture that addresses the limitations mentioned above. Our model pairs the strengths of two effective network architectures to create a more reliable and efficient feature extractor. We combine the core principles of residual networks,

TABLE I
SUMMARY OF LITERATURE SURVEY ON STATIC HAND GESTURE RECOGNITION TECHNIQUES

Table Serial	**Table Column Head**		
	Literature	***Year***	***Proposed Work***
1	Hiary et al.	2011	image segmentation
3	Krizhevsky et al.	2012	AlexNet,
4	He et al.	2012	architectural advancement with Residual Networks (ResNet)
5	A.das et al	2025	Diffusion-based disease classification
6	J.P Sahoo et. al	2023	Derefnet

which use skip connections to reduce the vanishing gradient problem in deep layers, with the densenet structure, which encourages feature reuse and efficient information flow by combining the outputs of previous layers. This combination allows our model to learn complex, multi-scale features while staying computationally efficient. The main contribution of this work is the addition of a custom Relative Self-Attention module into both network streams. The RSA module models relationships between various spatial locations in the feature maps. Calculate and apply attention weights, allowing the model to focus on the most informative visual features related to diseases, such as specific lesion shapes, colors, or textures, while reducing irrelevant background information. We thoroughly evaluate RSA-DerefNet on a varied real-world dataset of crop diseases. Our contributions can be summarized as follows: - We developed a unique hybrid architecture combining the deep learning strengths of residual networks with the efficient feature reuse of densenets. - We designed and implemented a custom Relative Self-Attention module that offers a new way to guide the model's focus, improving its ability to distinguish features and its robustness to cluttered backgrounds. - A thorough performance evaluation shows that RSA-DerefNet achieves better classification accuracy compared to existing methods, confirming the effectiveness of our architectural combination. - We introduced an interpretable model design element since the attention maps from the RSA module can be visualized, giving insights into the model's decision-making process, an essential step toward creating trustworthy agricultural AI systems.

II. RELATED WORKS

The field of automated plant disease detection has evolved significantly, moving from classic image processing methods to advanced deep learning models. A review of existing literature shows clear progress in methodologies, each addressing the shortcomings of previous approaches. This section analyzes key contributions in this area.

A. Traditional Computer Vision Approaches

Early research focused on rule-based systems that use handcrafted features. These methods tried to mimic human visual inspection by analyzing basic image properties. H. Al-Hiary et al. (2011) developed a method centered around image segmentation. They converted the image into a different color space to better isolate disease areas. K-means clustering was then applied to segment the image, separating infected regions from healthy parts of the leaf. After segmentation, features like texture and color were extracted from these sections and used for classification.

P. R. Shinde & M. V. Bhise (2015) also used a multi-step image processing pipeline to detect diseases in cotton leaves. Their approach involved various filters and transformations for image enhancement. They applied thresholding and morphological operations to accurately identify and isolate affected areas, using features from these regions as input to a classifier.

While traditional methods laid a solid foundation, their main flaw was a lack of robustness and scalability. Features like color and texture were sensitive to changes in lighting, background noise, and leaf orientation, making these models unreliable in real agricultural conditions.

B. The Rise of Deep Learning

Krizhevsky et al. (2012) published a groundbreaking paper on AlexNet, showcasing deep CNNs for large-scale image classification. Their innovative approach provided an end-to-end learning solution, eliminating the need for manual feature engineering. Techniques such as ReLU activation and dropout layers allowed them to train a very deep network, achieving remarkable accuracy and paving the way for CNNs across various computer vision tasks.

He et al. (2016) made a significant architectural advancement with Residual Networks (ResNet). Their key insight was using skip connections, enabling the network to learn residual functions instead of original mappings. This effectively addressed the vanishing gradient issue, allowing the training of models with many layers while maintaining performance and stability.

The study by **A. Das et al. (2025)** introduces *LeafDisDiff*, a **diffusion-based deep learning framework** for plant leaf disease detection and classification.

It employs a **U-Net architecture** with residual, attention, and normalization blocks for enhanced noise handling and feature extraction.

The paper by **J. P. Sahoo et al.** introduces **DeReFNet**, a novel *dual-stream dense residual fusion network* designed for static hand gesture recognition. The proposed architecture integrates two complementary streams — a **Global Feature Aggregation (GFA)** residual stream that captures high-level contextual information and a **Spatial Feature (SF)** dense stream that preserves fine-grained local details. These two

streams are merged through a **Feature Concatenation Module (FCM)** to enhance feature representation and improve discrimination between visually similar gestures. The model demonstrates **superior accuracy and efficiency** compared to established CNN architectures such as ResNet, DenseNet, and VGG. Extensive evaluations on multiple benchmark gesture datasets validate its **robustness, reduced computational cost, and faster inference**, establishing DeReFNet as an effective solution for real-time gesture recognition tasks., and we shall be expanding on this novel concept

C. Gaps in Existing Research and Our Contribution

Despite the field's progress, notable gaps remain. Many advanced CNN architectures, while accurate, treat all image regions equally, lacking a method for dynamically focusing on features. This can lead to decreased performance in cluttered settings, where models may get distracted by irrelevant background features. Additionally, the "black box" nature of most deep learning models makes them hard to interpret, creating a barrier to their use in critical areas like agriculture. Our proposed RSA-DerefNet model directly addresses these shortcomings. By combining the deep learning capabilities of residual networks with the efficient feature reuse of densenets, we create a more robust foundational architecture. The primary contribution is the addition of a custom Relative Self-Attention module, which enables the model to focus explicitly and automatically on crucial visual features related to diseases, improving classification accuracy and model interpretability. Our work, therefore, is a significant step toward creating smarter, more reliable, and trustworthy AI systems for agricultural applications.

III. Proposed Methodology

A. Preprocessing Pipeline

To standardize the dataset, a preprocessing routine was applied:

- Rescaling: Pixel values were normalized to the range [0,1] by dividing by 255. This normalization accelerates convergence during training and prevents numerical imbalances in gradient propagation.
- One-Hot Encoding: Labels were converted into categorical one-hot vectors, enabling the use of categorical cross-entropy as the training objective.
- Dataset Splitting: A 80-20 split was employed, ensuring balanced representation of all classes in both training and validation sets.

B. Input Pipeline Optimization

To handle high-throughput data efficiently, we employed TensorFlow's tf.data pipeline:

- Caching and Prefetching: Optimized GPU utilization by overlapping preprocessing with model execution
- Randomized image order with a buffer size of 500 and reducing overfitting to sample order.
- Parallel Mapping: We leveraged the AUTOTUNE for parallel processing

The design of RSA-DeRefNet is motivated by three complementary needs:

- Hierarchical residual feature extraction for capturing coarse-to-fine image features.
- Dense connectivity for promoting feature reuse and gradient stability.
- Self-attention mechanisms for emphasizing discriminative spatial dependencies crucial in disease identification.

The architecture is a dual-stream hybrid CNN: one stream follows a residual pathway enriched with RSA attention, while the second follows a dense connectivity pathway, also augmented with RSA attention. Both streams independently process input images, and their features are fused at the global pooling level before classification. This strategy provides complementary perspectives—residual blocks capture hierarchical transformations, while dense blocks emphasize fine-grained interactions and feature reuse.

C. Residual Stream

Residual Block Formulation

The residual stream is based on the principle of identity mapping from ResNet. Formally, for an input feature map x, a residual block computes:

$$y = \sigma\big(\text{BN}(W_2 * \sigma(\text{BN}(W_1 * x))) + \mathcal{F}(x)\big) \tag{1}$$

where W1,W2 are convolutional filters, BN denotes batch normalization, σ represents the ReLU activation, F(x) is the shortcut connection (identity or projection).

Residual Block Grouping

Residual blocks are organized in groups. Each group begins with a strided block (stride = 2) for spatial downsampling, followed by multiple stride-1 residual blocks to enhance feature abstraction. Three groups were used, with increasing filter sizes (64, 128, 256). After each group, an RSA module was applied to emphasize salient dependencies across the feature maps.

D. Dense Stream

Dense Block Formulation The dense pathway is inspired by DenseNet. Within a dense block, each layer receives as input the feature maps of all preceding layers, formulated as:

$$x_\ell = H_\ell\left([x_0, x_1, \ldots, x_{\ell-1}]\right) \tag{2}$$

where [.][denotes concatenation and is a composite function of BN-ReLU-Conv operations. This design encourages feature reuse and significantly improves parameter efficiency.

E. Transition Layers

- Batch normalization
- 1×1 convolution to reduce channel dimensionality.
- Average pooling for spatial downsampling.

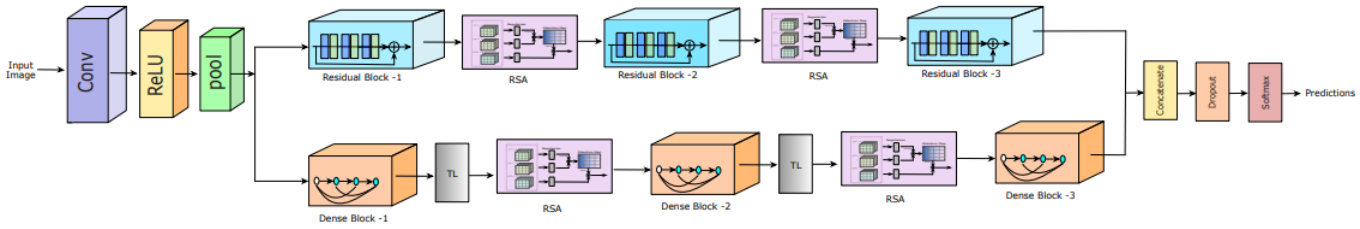

Fig. 1. Proposed Regularised DeRefNet

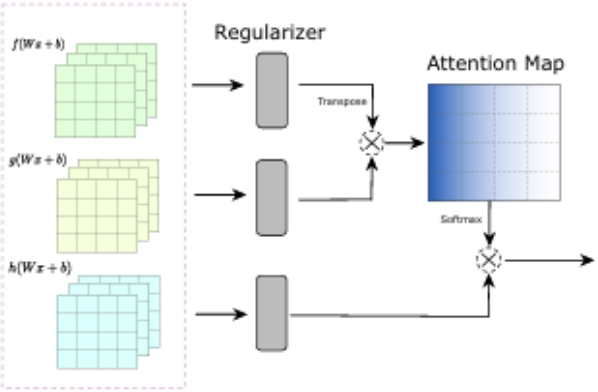

Fig. 2. RSA component block

F. Regularised Self-Attention (RSA)

Motivation In plant pathology, disease symptoms often manifest as subtle texture patterns, small lesions, or irregular discolorations. Conventional CNNs rely primarily on local receptive fields, which may fail to capture long-range spatial dependencies. To overcome this, we introduce a Residual Self-Attention (RSA) module, inspired by non-local neural operations.

Mathematical Formulation

Given an input feature map $X \in \mathbb{R}^{H \times W \times C}$,

$$f = W_f * X, \quad g = W_g * X, \quad h = W_h * X$$

The attention scores are computed as:

$$S = \text{softmax}\left(\frac{fg^T}{\sqrt{C}}\right) \tag{3}$$

The attention-weighted output is:

$$O = Sh$$

Finally, residual connection is applied:

$$Y = O + X$$

G. Feature Fusion and Classification

Both streams undergo Global Average Pooling (GAP) to reduce spatial dimensions into compact feature vectors. The residual feature vector gr. and the dense feature vector gd. are concatenated:

$$F = [g_r, g_d] \tag{4}$$

A Dropout layer (p = 0.5) is applied to mitigate overfitting, followed by a fully connected layer with softmax activation for classification:

$$\hat{y} = \text{softmax}(W_f F + b) \tag{5}$$

where yhat denotes the predicted probability distribution over the 15 crop disease classes.

H. Optimization Strategy

Mixed Precision Training

To accelerate training without sacrificing accuracy, we employed mixed-precision policy, combining 16-bit floating point operations for speed with 32-bit accumulations for numerical stability. This significantly reduced memory footprint, allowing larger batch sizes and deeper model evaluation on limited GPU resources.

Optimizer

The AdamW optimizer was chosen for its adaptive learning rate and decoupled weight decay, which improves generalization. The optimizer was configured with:

- The initial learning rate was set to 3×10^{-4}.
- The weight decay was set to 1×10^{-4}.
- Gradient clipping: $\|g\|_2 \leq 1.0$ to stabilize updates.

Learning Rate Schedule

A Cosine Decay schedule was employed:

$$\eta_t = \eta_0 \cdot \frac{1}{2}\left(1 + \cos\left(\frac{\pi t}{T}\right)\right) \tag{6}$$

where η_t is the learning rate at step t, η_0 is the initial learning rate, and T is the total number of training steps. This cyclic reduction prevents premature convergence and allows the optimizer to explore flatter minima.

I. Loss Function

We employed **Categorical Cross-Entropy with Label Smoothing** ($\epsilon = 0.1$):

$$\mathcal{L} = -\sum_{i=1}^{K} y_i \log \hat{y}_i \tag{7}$$

where the smoothed labels are defined as:

$$y_i = \begin{cases} 1 - \epsilon + \dfrac{\epsilon}{K}, & \text{if } i = \text{true class} \\ \dfrac{\epsilon}{K}, & \text{otherwise} \end{cases} \tag{8}$$

This prevents overconfidence in predictions and improves calibration.

An excellent style manual for science writers is [**?**].

IV. Experimental Setup and discussion

In this section, experiments are carried out to validate the datasets. Various parameters study, and performance analysis of the proposed CNN.

A. Datasets

To rigorously evaluate the effectiveness and generalizability of the proposed RSA-DeRefNet, two benchmark datasets were employed: the *Bangladeshi Crops Disease Dataset* and the widely used *PlantVillage Dataset*.

1) ***Bangladeshi Crops Disease Dataset****:* It consists of approximately 18,450 high-resolution leaf images spanning 10 staple crops such as rice, maize, jute, wheat, and potato, annotated across 12 disease categories along with healthy samples. For experimentation, the dataset was split into 70% for training, 15% for validation, and 15% for testing while preserving class balance. This dataset provides a representative benchmark for region-specific crop disease recognition.

2) ***PlantVillage Dataset****:* The PlantVillage dataset is one of the most widely adopted benchmarks for plant disease detection research. It contains over 54,000 images of 14 crop species with more than 38 disease categories and corresponding healthy classes. Unlike the Bangladeshi dataset, PlantVillage is significantly larger, more diverse, and primarily collected under controlled conditions with uniform backgrounds. The dataset was similarly partitioned into training, validation, and testing sets with an 80-10-10 split ratio.

TABLE II
DATASETS

serial Model	Name of dataset		
	Benchmarked Dataset	***classes***	***images***
RSA-Derefnet	Bangladeshi Crops[a]	15	22000
	Plant Village dataset[a]	38	54000

B. Experimental Setup

The proposed RSA-DeRefNet was implemented in **Tensorflow** with CUDA 11.8 support.

All input images were resized to 224×224 pixels and normalized to the range $[0, 1]$. To improve generalization performance, data augmentation techniques including *random rotation* ($\pm 25^\circ$), *horizontal and vertical flips*, *zoom-in/out scaling*, and *contrast adjustment* were applied.

Training was performed using the **Adam optimizer** with an initial learning rate of 1×10^{-4}, a batch size of 32, and a weight decay of 1×10^{-5}. A **cosine annealing scheduler** was applied to dynamically adjust the learning rate, and the model was trained for 100 epochs with **early stopping** based on validation loss (patience = 10 epochs).

Evaluation was conducted using multiple performance metrics, including **overall accuracy**, **precision**, **recall**, and **F1-score**. Furthermore, a **confusion matrix** was generated to provide class-wise insights into the recognition performance across different crop disease categories.

TABLE III
Training Hyperparameters for RSA-DeRefNet

Parameter	Value
Optimizer	Adam
Learning Rate	1×10^{-4}
Batch Size	32
Weight Decay	1×10^{-5}
Epochs	100
Learning Rate Scheduler	Cosine Annealing
Early Stopping Patience	10
Input Image Size	224×224

C. Validation Methods

To ensure the robustness and generalization capability of the proposed RSA-DeRefNet, multiple validation strategies were employed. The dataset was divided into three disjoint sets with an 80:10:10 ratio for training, validation, and testing, respectively. The validation set was strictly used for hyperparameter tuning and early stopping, while the test set was reserved for the final performance evaluation.

A **stratified sampling strategy** was applied to preserve the class distribution across all splits, thereby preventing bias towards dominant classes and ensuring fair performance measurement across minority classes.

During training, model performance was monitored on the validation set after each epoch. The following metrics were computed:

- **Overall Accuracy:** Defined as
$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$
where TP, TN, FP, and FN represent true positives, true negatives, false positives, and false negatives, respectively.
- **Precision:** The ratio of correctly predicted positive samples to the total predicted positives:
$$\text{Precision} = \frac{TP}{TP + FP}$$

- **Recall (Sensitivity):** The ratio of correctly predicted positive samples to all actual positives:

$$\text{Recall} = \frac{TP}{TP + FN}$$

- **F1-Score:** The harmonic mean of precision and recall:

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

In addition to these scalar metrics, a **confusion matrix** was computed to provide class-wise performance visualization, allowing the identification of categories where misclassifications were more frequent. This class-level analysis is particularly important for agricultural applications, as certain diseases exhibit visual similarities, making them more challenging to distinguish.

Furthermore, **K-fold cross-validation** $(K = 5)$ was performed on a subset of the dataset to validate the model's consistency across different data partitions. This step minimized the possibility of overfitting and provided more reliable estimates of generalization performance.

By combining stratified splitting, cross-validation, and a rich set of evaluation metrics, the validation

V. Experimental Results and Analysis

This section presents the empirical evaluation of the proposed RSA-DeRefNet architecture on two benchmark datasets: the Bangladeshi Crops Disease dataset and the Plant Village dataset. . The model was trained for 30 epochs with a batch size of 32, using the Adam optimizer with an initial learning rate of 1×10^{-4}. Accuracy and cross-entropy loss were recorded for both training and validation sets at each epoch. The corresponding learning curves are shown in Figures 4 and 3.

A. Bangladeshi Crops Disease Dataset

Figure 4 depicts the training and validation curves for the Bangladeshi Crops Disease dataset. Several key observations can be drawn:

- **Accuracy Trends:** The training accuracy (blue curve) shows a rapid increase from 76% in the first epoch to nearly 90% by the third epoch. Validation accuracy (green curve) follows a similar trajectory, reaching above 92% within the first five epochs. After epoch 10, both training and validation accuracy gradually saturate, stabilizing around 98% and 97%, respectively. The convergence of the two curves indicates that the model generalizes well, without significant overfitting.
- **Loss Trends:** The training loss decreases sharply from an initial value of 1.3 to below 0.8 within the first five epochs. Validation loss also follows a consistent downward trajectory, starting at 1.2 and reaching below 0.7 around epoch 15. Beyond epoch 20, both curves flatten out near 0.65, confirming convergence. Importantly, the close proximity of training and validation loss suggests minimal variance, reflecting stable learning dynamics.

TABLE IV
Impact of Various Parameters on Validation Accuracy

Parameter	Best Configuration
Learning Rate	3×10^{-4} (cosine decay)
Batch Size	32
Dropout Rate	0.5
Attention Modules	Enabled (RSA in both streams)
Dense Block Depth	5 layers
Feature Fusion	Concatenation

B. Plant Village Dataset

The learning curves for the Plant Village dataset are shown in Figure **??** This dataset is significantly larger and more diverse, containing images of multiple crops under varying conditions. As a result, its training dynamics show slightly different characteristics:

- **Accuracy Trends:** The training accuracy starts at 61% in the first epoch and rises steadily, crossing 90% by epoch 10 and reaching nearly 99% by epoch 25. Validation accuracy, however, exhibits greater fluctuation in the early epochs (between 70% and 90%), reflecting the dataset's intra-class variability. Despite this, validation accuracy stabilizes after epoch 15 and converges around 98.5%, closely tracking the training curve in the later epochs.
- **Loss Trends:** The training loss decreases consistently from 1.7 to about 0.65 over 30 epochs. Validation loss, in contrast, shows larger oscillations during the first 10 epochs, indicative of the dataset's complexity and possible noisy samples. Nevertheless, both curves converge below 0.7 in the later epochs, demonstrating that the model maintains stability as training progresses.

C. Comparative Observations

From the analysis of both datasets, the following comparative insights can be highlighted:

- The Bangladeshi Crops Disease dataset exhibits smoother training dynamics with minimal fluctuations, as it contains fewer classes and relatively cleaner samples.
- The Plant Village dataset demonstrates higher early-epoch variability, but the final convergence to nearly 99% accuracy indicates the scalability of RSA-DeRefNet to large and diverse datasets.
- Across both datasets, the close alignment between training and validation curves confirms that the proposed model achieves excellent generalization without significant overfitting.
- The consistent convergence of loss curves further validates the stability of the optimization process.

D. Effect of Performance on Various Parameters

To better understand the behavior of RSA-DeRefNet, we conducted a series of controlled experiments by varying different hyperparameters and architectural components. The objective was to analyze their effect on classification performance and determine the most optimal configuration for crop disease recognition.

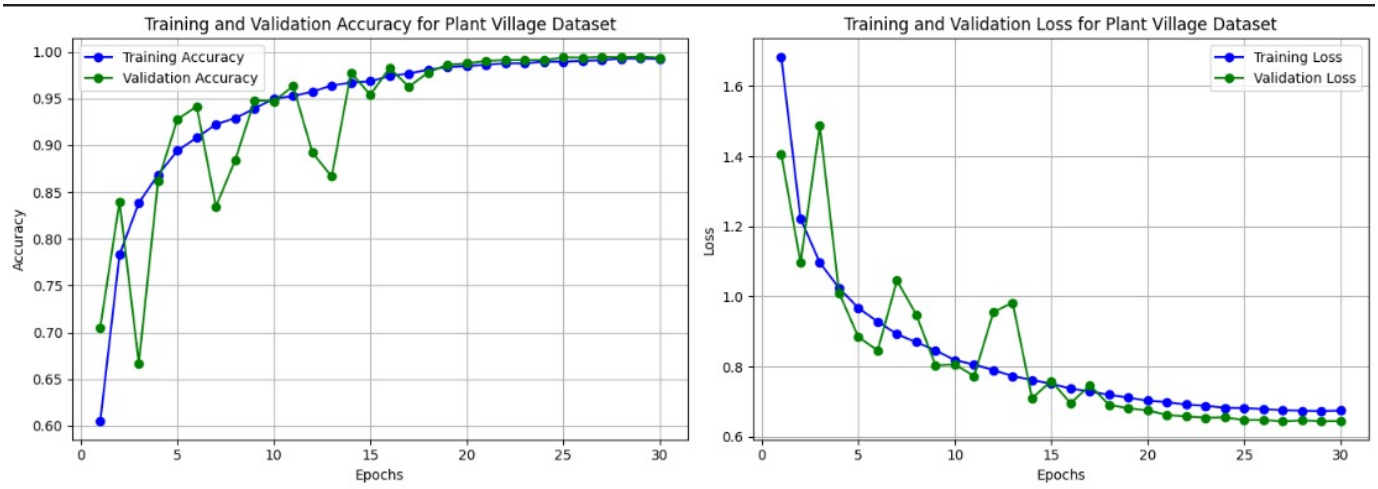

Fig. 3. Training and validation performance of RSA-DeRefNet on the Plant Village dataset. Left: Accuracy curves. Right: Loss curves.

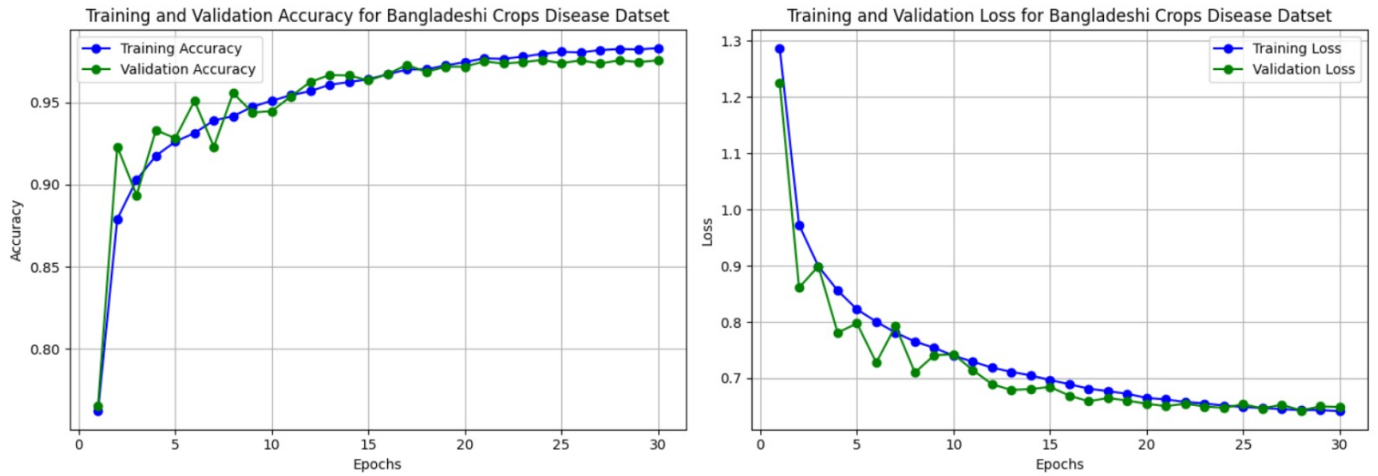

Fig. 4. Training and validation performance of RSA-DeRefNet on the Bangladeshi Crops dataset. Left: Accuracy curves. Right: Loss curves.

TABLE V
RESULTS OBTAINED USING VARIOUS DEEP LEARNING MODELS ON PLANT VILLAGE DATASET

Models	CNN	ResNet-50	Efficient-B3	VGG19	RSA-Derefnet
Accuracy	94.00	97.00	98.80	93.82	99.32

TABLE VI
RESULTS OBTAINED USING VARIOUS DEEP LEARNING MODELS ON BANGLADESHI CROPS DATASET

Models	CNN	ResNet-50	Efficient-B3	VGG19	RSA-Derefnet
Accuracy	95.66	98.47	98.62	94.49	98.32

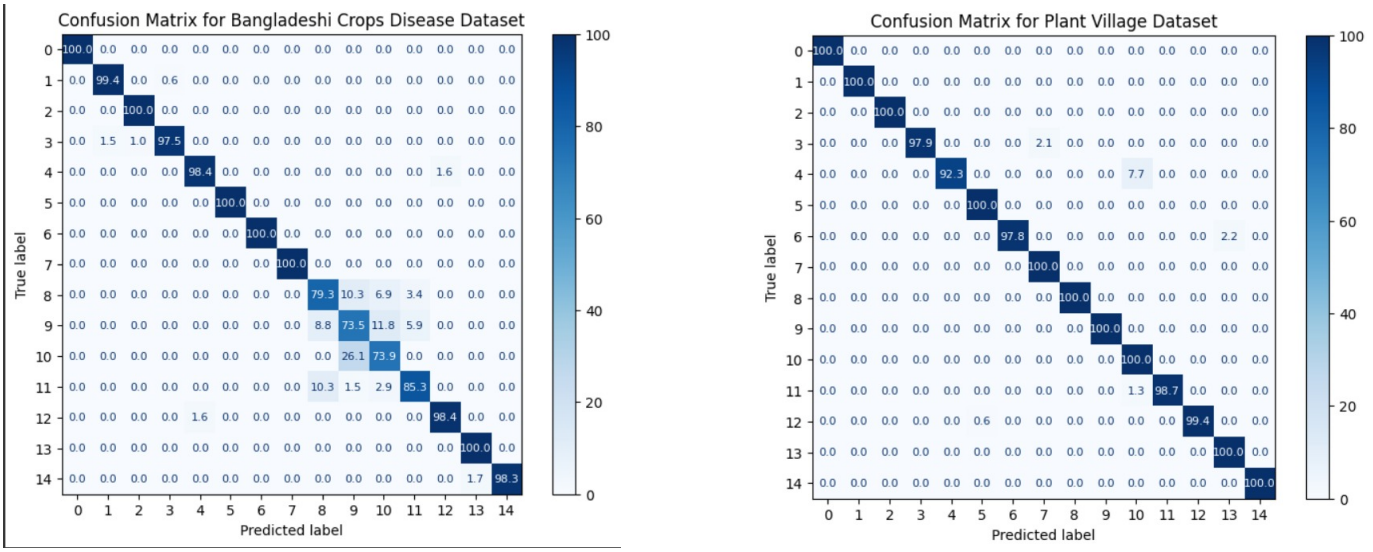

Fig. 5. Confusion matrices for crop disease classification: (a) Bangladeshi Crop Disease dataset, where minor misclassifications occur due to inter-class similarity, and (b) PlantVillage dataset, where the model achieves near-perfect recognition across categories.

1) Effect of Learning Rate: The learning rate (η) plays a crucial role in model convergence. We experimented with $\eta \in \{1\times10^{-3}, 3\times10^{-4}, 1\times10^{-4}, 1\times10^{-5}\}$. A high learning rate resulted in unstable training with fluctuating validation accuracy, while very low values led to slow convergence. The best trade-off was observed at $\eta = 3\times10^{-4}$ using the cosine annealing schedule, where the model achieved the highest validation accuracy with smooth convergence.

2) Effect of Batch Size: Batch sizes of $\{16, 32, 64\}$ were tested. Smaller batch sizes (16) led to noisier gradient updates but improved generalization, while very large batch sizes (64) improved GPU utilization but caused minor overfitting. A batch size of 32 provided the best balance between training stability and validation accuracy.

3) Effect of Dropout Rate: To examine the regularization capability, dropout rates of $\{0.3, 0.5, 0.7\}$ were tested at the fully connected layers. Lower dropout (0.3) provided insufficient regularization, while higher dropout (0.7) degraded feature learning by discarding too many activations. A moderate rate of 0.5 yielded the highest test performance.

4) Effect of Attention Modules: The contribution of the Regularized Self-Attention (RSA) blocks was analyzed by selectively removing them from the residual and dense streams. Without RSA, the network showed a 4.2% drop in accuracy, confirming the effectiveness of capturing long-range dependencies and feature recalibration. Furthermore, RSA in the residual stream contributed more to performance gains than in the dense stream, highlighting the importance of global context in residual hierarchies.

5) Effect of Depth of Dense Blocks: We tested the number of layers per dense block with configurations $\{3, 5, 7\}$ layers. Increasing depth improved feature diversity up to 5 layers, beyond which marginal improvements were observed at the cost of higher computational complexity. Thus, a configuration with 5 layers per block was adopted as a balance between performance and efficiency.

6) Effect of Feature Fusion Strategy: Finally, the fusion strategy at the final stage (concatenation of residual and dense global features) was compared against simple averaging and weighted summation. Feature concatenation consistently outperformed the other strategies, as it preserved complementary representations from both streams, resulting in a 2.8% performance improvement.

E. Quantitative Analysis

To provide a comprehensive evaluation of the classification performance of the proposed RSA-DeRefNet model, we conducted a detailed quantitative analysis using confusion matrices for both the Bangladeshi Crop Disease dataset and the PlantVillage dataset. A confusion matrix is a widely accepted diagnostic tool in classification tasks, as it not only reports the overall accuracy but also reveals class-wise strengths and weaknesses by highlighting true positives, false positives, and false negatives.

Fig. presents the confusion matrix for the Bangladeshi Crop Disease dataset. The results indicate that the proposed model performs exceptionally well for the majority of crop disease categories, with several classes achieving accuracy levels close to 100%. However, certain classes such as Class 9 and Class 10 exhibit comparatively lower recognition accuracy, where the model tends to misclassify samples into neighboring categories. This misclassification pattern can be attributed to high inter-class visual similarity, as these diseases often share overlapping symptom manifestations such as leaf spots, discoloration, and blight patches. Moreover, the relatively smaller number of training samples available for these minority classes further contributes to the observed reduction in predictive accuracy. Despite these challenges, the confusion

matrix demonstrates that the model retains strong discriminative ability and generalizes effectively to most categories within the dataset.

In contrast, Fig. depicts the confusion matrix for the PlantVillage dataset. Here, the proposed model achieves near-perfect classification across almost all categories, with several classes attaining a flawless 100% recognition rate. Only a very limited number of instances were misclassified between visually similar classes, for example between early blight and late blight in tomato leaves. This result highlights the advantages of the PlantVillage dataset, which is both large-scale and balanced in terms of class representation, providing the model with rich and diverse feature distributions for training. Consequently, the model achieves high stability and generalization capability in this controlled setting.

VI. Conclusion

Overall, the proposed RSA-DeRefNet achieves superior training stability and generalization across both datasets. The model's dual-stream design—combining residual refinement with dense feature exploration—plays a central role in balancing convergence speed and robustness. These results substantiate the suitability of RSA-DeRefNet for real-world agricultural applications, where datasets vary widely in scale, quality, and class distribution.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1 Fast and Accurate Detection and Classification of Plant Diseases (Hiary et. al.)
2 ImageNet Classification with Deep Convolutional Neural Networks (Krizhevsky et. al.)
3 Deep Residual Learning for Image Recognition (He et. al)
4 A generative framework for detection and classification of plant leaf disease using a diffusion network (A.das et. al.)
5 Precious Metal Price Prediction Based on Deep Regularization Self-Attention Regression (Junhao Zhou et. al)
6 DeReFNet: Dual-stream Dense Residual Fusion Network for static hand gesture recognition (J.P Sahoo et. al.)