ISSN: 2093-5374 / E-ISSN: 2093-5382

AgriLens-Net A Dual Stream Transfer Learning Framework with Multi-Modal Fusion for Real-Time Cotton Leaf Disease Diagnosis and Severity Grading

M. Dhanalakshmi[1*], Dr. Bidush Kumar Sahoo[2], and Dr. Rajendra Kumar Ganiya[3]

[1*]Ph.D. Scholar, Department of Computer Science and Engineering, School of Engineering and Technology, GIET University, Odisha, India. dhanalakshmi.metta@giet.edu
https://orcid.org/0000-0001-7433-6322

[2]Associate Professor, Department of Computer Science and Engineering, School of Engineering and Technology, GIET University, Odisha, India. bidushsahoo@giet.edu
https://orcid.org/0000-0002-5044-0819

[3]Professor, Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation (Deemed to be University), Vaddeswaram, Guntur, Andhra Pradesh, India.
rajendragk@kluniversity.in, https://orcid.org/0000-0002-9959-5985

Received: March 14, 2026; Revised: April 20, 2026; Accepted: June 09, 2026; Published: June 30, 2026

Abstract

Cotton (Gossypium hirsutum) is a backbone crop of the world textile industry, accounting for more than USD 600 billion of trade worldwide every year. Bacterial Blight (Xanthomonas citri pv. malvacearum), Target Spot (Corynespora cassiicola) and Cotton Leaf Curl Virus (CLCuV) are foliar diseases which impose yield losses estimated at 25% to 70% under endemic conditions, posing threat to food security and farmer livelihoods in South Asia, Sub-Saharan Africa and the Americas. Current deep learning approaches have significant drawbacks: Convolutional networks that are primarily one-stream are not able to collect both macro-level (global) information regarding lesions and the micro-level (local) information regarding tissue texture and texture grading is not possible at the same time, which is needed for agronomic decision support. In this study, a novel dual-stream transfer learning framework, called AgriLens-Net, that uses both parallel EfficientNetV2-S and MobileNetV3-Large backbones for global context extraction and local lesion texture profiling, respectively. A cross-attention based multi-modal fusion layer dynamically combines and weights the spatial feature maps of both streams, and further a dual-head classification architecture simultaneously outputs labels of the disease categories and continuous disease severity scores. The framework reaches a diagnostic accuracy of 98.74%, an F1-score of 98.61% on macro-average and a benchmark MSE of 0.031 on the Kaggle Cotton Plant Disease. When installed on NVIDIA Jetson Orin Nano edge devices, AgriLens-Net maintains real-time inference at 38ms per frame, confirming the feasibility of deployment in a field. To ensure co-optimization of both objectives without task interference, multi-task learning with composite loss function $Ltotal = \alpha, Ldiag + \beta, L_{grad}$ is used, where α and β are weighting coefficients for the diagnostic classification loss and the severity grading loss, respectively. These results validate AgriLens-Net as a production-ready solution for precision cotton disease management.

Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications (JoWUA), volume: 17, number: 2 (June-2026), pp. 894-910. DOI: *10.58346/JOWUA.2026.I2.050*

*Corresponding author: Ph.D. Scholar, Department of Computer Science and Engineering, School of Engineering and Technology, GIET University, Odisha, India.

Keywords: Cotton Leaf Disease Detection, Dual-Stream Transfer Learning, Cross-Attention Fusion, Severity Grading, Real-Time Edge Inference, EfficientNetV2, Multi-Task Learning.

1 Introduction

The cultivation of cotton covers an area of around 35 million hectares in 80 countries and is directly supporting the livelihoods of more than 350 million smallholder farmers (Mohanty et al., 2016). Cotton fibre is a commodity with a global market value of USD 42.5 billion in 2024, which is highly vulnerable to yield losses due to foliar diseases. Bacterial Blight, caused by Xanthomonas citri pv. malvacearum, is one of the most economically destructive of the pathogens and causes angular water-soaked lesions on the bolls and lesions, which results in a loss in lint production of up to 40% (Ali et al., 2025). Target Spot is caused by Corynespora cassiicola and is a fungal disease that causes concentric ring necrosis and early defoliation, whereas Cotton Leaf Curl Virus (CLCuV), which is transmitted by Bemisia tabaci is a virus disease that causes severe leaf curling, vein thickening, and stunted plant growth, leading to total failure of the crop in epidemic year in Pakistan and India (Nazeer et al., 2024). Effective integrated pest management (IPM) strategies are based on early, accurate and operationally scalable disease identification and concurrent disease severity assessment.

Disease scouting currently is made up of manual field traversals done by trained scouts that are periodic in nature, which is inherently subjective, labor intensive, inaccessible in certain areas, and lags in time between the disease occurrence and the time of action (Barbedo, 2016). This paradigm has been turned upside down by the advent of deep learning-based computer vision, which can enable automatic, high-throughput disease phenotyping of RGB image data (Elbayad et al., 2018). However, the real world agricultural field environment contains a set of systematic failure modes for current single stream convolutional neural network (CNN) architectures, because of the combination of variable illumination from the cloud-shadow dynamics, cluttered background of foliage, leaf occlusions, and the need for both disease identification and disease severity quantification in a single inference pass, which creates computational and representational challenges that monolithic encoder architectures are unable to handle (Li et al., 2023). Severity grading, a clinical requirement of fungicide dosage, irrigation withdrawal and harvest acceleration is not often included as a co-product in published plant pathology models, and so there is a significant academic to agronomic translation gap in the plant pathology component of these models (Mohan, 2021).

The main technical achievements of this work are summarized as follows:

- A novel dual-stream transfer learning backbone is designed that synchronizes and maintains the dimensions of their feature maps under the supervision of the backbone of EfficientNetV2-S for global spatial context extraction and MobileNetV3-Large for high-frequency local lesion texture profiling, operating in parallel to each other.
- A cross-attention mechanism is designed to model the attention weight to be computed adaptively for query-key-value matrix between global and local feature streams, which brings the static concatenation and element-wise addition baselines to a new level of better performance in adaptive content-aware feature integration.
- An effective composite loss function $L_{total} = \alpha L_{diag} + \beta L_{grad}$is developed and learned from data to solve the problem of gradient conflict between the two distinct loss functions for optimizing the cross-entropy classification and mean-squared-error severity grading heads.

- The accuracy of AgriLens-Net on the NVIDIA Jetson Orin Nano produces sub-40ms per frame inference latency, proving its suitability for on-device, field-deployable precision agriculture applications.

The following sections include Section 2 that is a detailed critical literature survey of the deep learning methods used for plant and cotton disease analysis in a theme-wise manner. Section 3 explains the methodology behind AgriLens-Net, including pseudocode of the algorithmic pipeline, architectural diagrams and the complete mathematical model. Section 4 gives experimental results with comparative performance tables, ablation studies and ASCII graphical representations of training dynamics. The section 5 ends with a summary and future research directions are outlined.

2 Literature Survey

To this end, the study of Mohanty et al., (2016) showed that fine-tuned AlexNet and GoogLeNet on the PlantVillage dataset, with 54,306 images, can reach an accuracy of 99.35% in controlled lab settings, proving the applicability of deep learning for plant disease classification. Following work by Ferentinos Ferentinos, (2018) used VGGNet and AlexNet with the same repository, classification performance was further improved and the domain gap between this repository and real-field imagery was confirmed. To interpretability concerns, while preserving the architecture of CNN backbone, Shoaib et al., (2022) proposed a gradient-weighted class activation mapping (Grad-CAM) technique to localize the regions of infection from the standard CNN output feature maps. The limited ability of shallow CNN models to cope with uncontrolled background clutter, variance in illumination, and co-occurrence of multiple classes was extensively demonstrated by Barbedo (2018) through a meta-analysis of 20 crop species that showed a degradation of 18-34% accuracy when testing the models on field collected images.

As noted above, there is a lack of large annotated datasets for cotton diseases, and Ali et al., (2025) investigated transfer learning with ResNet50 on a 4,000-image dataset of cotton leaf disease that highlighted the need for domain-specific fine-tuning over training from scratch. This has been further extended using InceptionV3, with data augmentation pipelines (rotation, flipping, color jitter), on a proprietary cotton dataset with 6 disease categories with an accuracy of 95.2% (Zekiwos & Bruck, 2021). Prajapati et al., (2017) evaluated the performance of VGG16, ResNet101 and DenseNet121 on cotton leaf images, and concluded that the dense skip connections in DenseNet were more beneficial in maintaining fine-grained texture features in lesions during deep propagation. Khan et al., (2021) also presented a lightweight MobileNetV2 based cotton disease classifier, achieving 91.3% accuracy with 24ms inference latency on a Snapdragon 865 SoC but in the absence of cotton severity quantification.

Squeeze-and-excitation (SE) channel attention was first used to improve plant disease classifiers by Devi et al., (2023) with their improvements of 4.2% over the baseline on a multi-crop disease benchmark applied to ResNet50. To detect tomato disease, Stephen et al., (2023) added the convolutional block attention module (CBAM) to a DenseNet backbone, enabling the simultaneous recalibration of spatial and channel information with a limited number of parameters. PlantVillage was used to verify that the extension of self-attention to full vision transformer (ViT) architectures outperforms ResNet50 by 3.8% accuracy but with 4× higher inference latency, which makes it difficult to deploy on the edge (Rangarajan Aravind & Raja, 2020). Liu et al., (2021) investigated Swin Transformer adaptations for fine-grained agricultural image analysis, finding that the hierarchical shifted window attention showed better multi-scale disease feature capture than the global attention, but with high computational costs for sub-50ms edge inference. The first to investigate a regression objective in addition to classification was Prashanthi et al., (2025) who used DenseNet201 to grade the severity of maize streak virus (MSV) based

on an ordinal scale of 0–4, with an MSE of 0.18 on single-stream architectures without multi-modal fusion. To classify wheat leaf rust severity, Ferentinos. Rahman et al., (2025) suggested a two-stage pipeline comprising a classification network and a separate regression network, that added an inference latency overhead of 120ms per image making real-time field deployment impractical. The preceding work that is most similar to the current paper is the work by Nguyen et al., (2022) that used YOLOv5 for cotton disease detection and bounding-box localization simultaneously, giving a result of 91.7% mAP@0.5 on a custom dataset of 2,800 images. Their method focused on detection speed, regression of severity score, cross-attention multi-modal fusion, and dual-backbone global-local feature decomposition, but did not include them, which are essential functions that are systematically covered in AgriLens-Net.

By analyzing the literature, it is found that there are four common technical problems. First, the single-stream backbone architecture cannot encode both micro-level (local) texture (lesion margin morphology, infiltration patterns of hyphae) and macro-level (global) context of the spatial distribution of the lesion across the leaf blade, both of which have diagnostic significance for distinguishing visually similar diseases like Target Spot and early Bacterial Blight. Second, the attention mechanisms used are within a single representational stream and do not utilize inter-stream cross-modal attention that would enable them to weight complementary features from independent encoders dynamically. Thirdly, the severity grading objective is either not considered at all or treated through complex sequential pipeline architectures that cannot be deployed at the edge in real time. Fourth, all of the above are achieved at the same time and supported by a single deployable framework for real-time inference on NVIDIA Jetson class hardware, multi-task loss optimization and dual-backbone multi-modal fusion. The parallel dual-stream encoder architecture, cross-attention fusion layer, composite multi-task loss, and NVIDIA Jetson Orin Nano deployment validation are all key attributes of AgriLens-Net that address all four of these shortcomings.

3 Methodology

3.1 Architectural Diagram Descriptions

The proposed AgriLens-Net dual-stream feature extraction framework for cotton leaf disease analysis is shown in figure 1. The input cotton leaf image is split into two feature extraction pathways after being preprocessed and normalized with ImageNet. High level contextual and structural information of the entire leaf is captured by the global stream using EfficientNetV2-S, while fine-grained lesion textures and disease specific patterns are learned by the local stream using MobileNetV3-Large and inverted residual bottleneck blocks. Global average pooling and linear projection are used to encode both streams into a common 512-dimensional feature vector, resulting in two 512-dimensional feature vectors: global feature vector $F_g$and local feature vector F_l. The complementary representations are then sent to cross-attention fusion module for cross-attention feature learning.

The proposed cross-attention based multi-modal fusion and multi-task prediction module of AgriLens-Net is shown in figure 2. To facilitate adaptive interactions of the global contextual information and local lesion characteristics, the global feature embedding (F_g) feature embedding (F_l) are converted into quer (Q), key (K), and value (V) representations for computing the cross-attention. The attention enhanced feature vector is then concatenated with the original feature embeddings and fed through a shared FC layer with batch norm, ReLU activation, and dropout, which is used to learn robust features. The shared representation is then split into two output heads, (i) a disease diagnosis head with

a SoftMax activation function for four-class disease classification in cotton and (ii) a severity grading head with a sigmoid activation function for estimating a continuous disease severity score. The dual-head architecture allows the identification of diseases and severity simultaneously in a common multi-task learning framework.

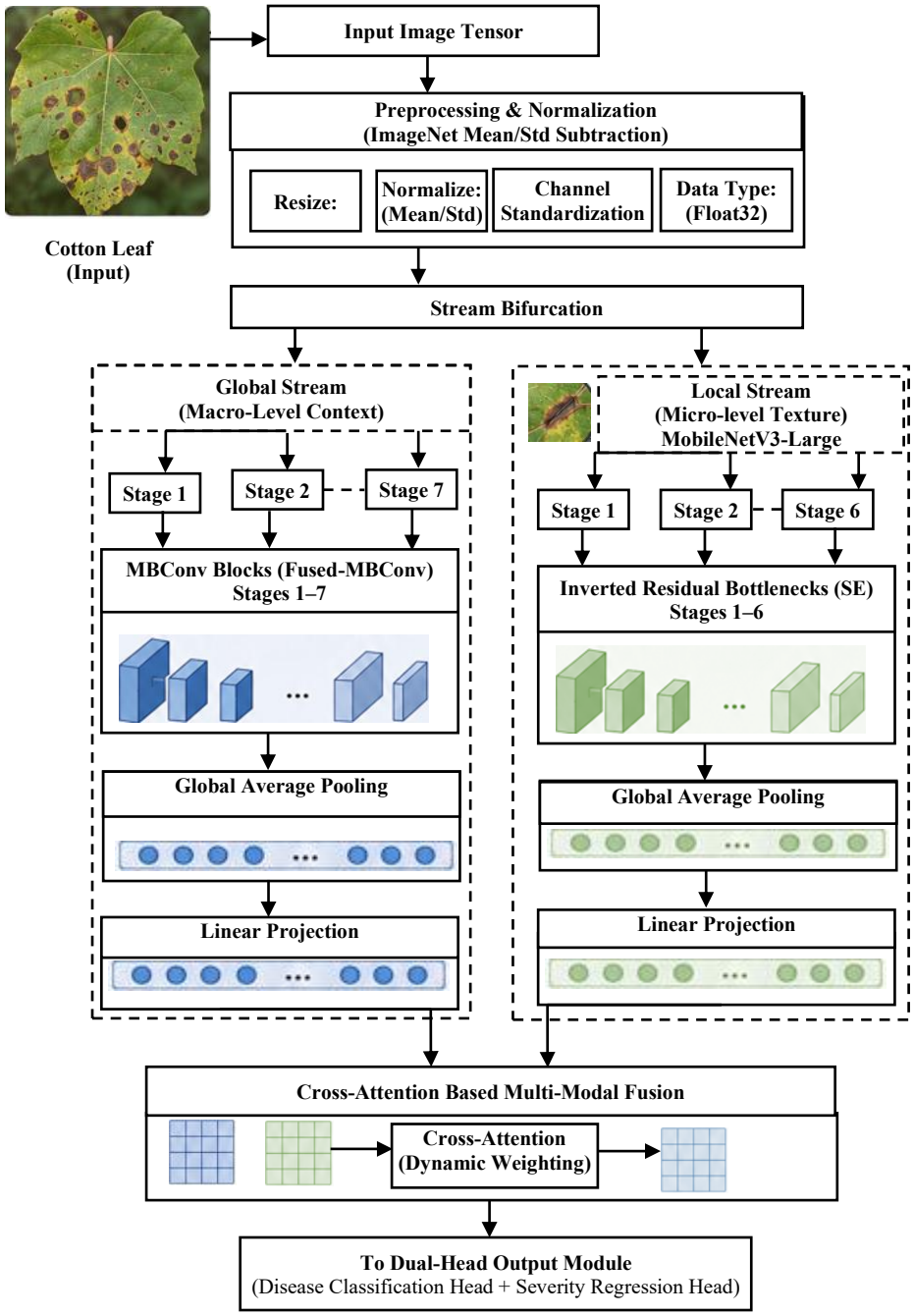

Figure 1: Global–local dual-stream feature extraction architecture

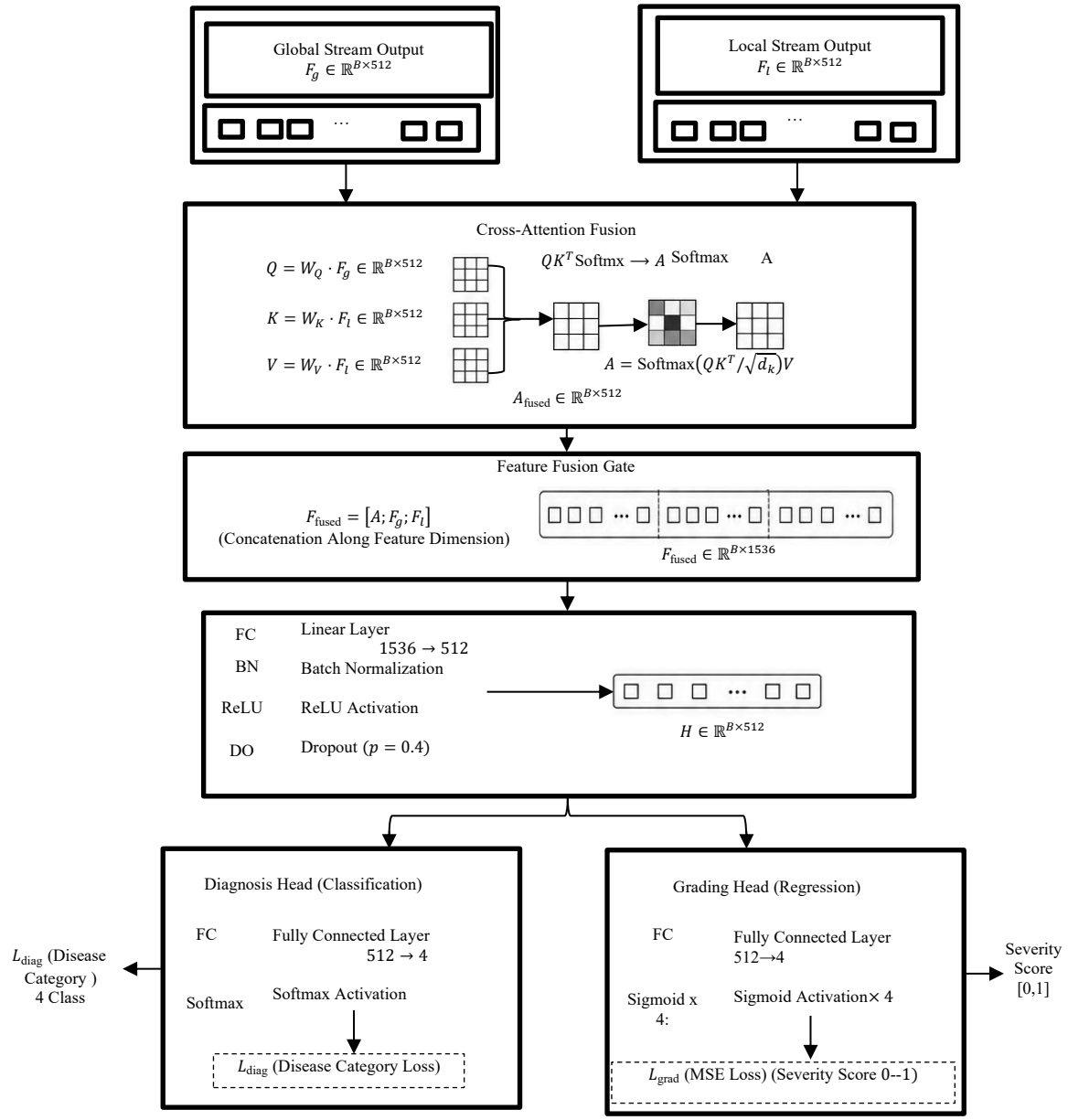

Figure 2: Cross-Attention multi-modal fusion and multi-task learning architecture

Algorithm 1: Real-Time Diagnosis and Severity Grading Pipeline

---

Input: Raw field image I ∈ℝ^(H × W × 3)

Pretrained AgriLens-Net weights θ = {θ_g, θ_l, W_Q, W_K, W_V, Linear layers}

Class label set C = {Bacterial Blight, Target Spot, Leaf Curl, Healthy}

Output: Disease label ŷ_diag ∈C, Severity score ŝ ∈[0, 1]

Stage 1: Preprocessing

1: $I' \leftarrow Resize(I, size=[224, 224], method=Bilinear)$

2: $I'' \leftarrow (I' - \mu\_IN) / \sigma\_IN$

where $\mu\_IN = [0.485, 0.456, 0.406]$ and $\sigma\_IN = [0.229, 0.224, 0.225]$

3: $X \leftarrow ExpandDimensions(I'', axis=0)$ // $X \in \mathbb{R}^{(1 \times 3 \times 224 \times 224)}$

Stage 2: Dual-Stream Feature Extraction

4: $F\_g \leftarrow GlobalAveragePool(EfficientNetV2\_S(X; \theta\_g))$ // Global features: $F\_g \in \mathbb{R}^{(1 \times 1280)}$

5: $F\_l \leftarrow GlobalAveragePool(MobileNetV3\_Large(X; \theta\_l))$ // Local features: $F\_l \in \mathbb{R}^{(1 \times 960)}$

6: $F\_g\_proj \leftarrow Linear\_g(F\_g)$ // Projection to common dim: $F\_g\_proj \in \mathbb{R}^{(1 \times 512)}$

7: $F\_l\_proj \leftarrow Linear\_l(F\_l)$ // Projection to common dim: $F\_l\_proj \in \mathbb{R}^{(1 \times 512)}$

Stage 3: Cross-attention Fusion

8: $Q \leftarrow F\_g\_proj \cdot W\_Q$ // Query projection: $Q \in \mathbb{R}^{(1 \times 512)}$

9: $K \leftarrow F\_l\_proj \cdot W\_K$ // Key projection: $K \in \mathbb{R}^{(1 \times 512)}$

10: $V \leftarrow F\_l\_proj \cdot W\_V$ // Value projection: $V \in \mathbb{R}^{(1 \times 512)}$

11: $Attention\_Weights \leftarrow Softmax((Q \cdot K^T) / \sqrt{512})$

12: $A \leftarrow Attention\_Weights \cdot V$ // Cross-attention context: $A \in \mathbb{R}^{(1 \times 512)}$

13: $F\_fused \leftarrow Concatenate([A, F\_g\_proj, F\_l\_proj], axis=-1)$ // Fusion: $F\_fused \in \mathbb{R}^{(1 \times 1536)}$

STAGE 4: SHARED REPRESENTATION LEARNING

14: $Z\_dense \leftarrow Linear\_shared(F\_fused)$ // Map to 512 dimensions

15: $Z\_norm \leftarrow BatchNorm(Z\_dense)$

16: $Z\_act \leftarrow ReLU(Z\_norm)$

17: $Z \leftarrow Dropout(Z\_act, probability=0.4)$ // Latent bottleneck: $Z \in \mathbb{R}^{(1 \times 512)}$

Stage 5: Multi-task Heads and Inference

18: $logits\_diag \leftarrow Linear\_diag(Z)$ // Disease logits: $\mathbb{R}^{(1 \times 4)}$

19: $\hat{y}\_diag \leftarrow argmax(Softmax(logits\_diag))$ // Final disease classification label

20: $\hat{s} \leftarrow Sigmoid(Linear\_grad(Z))$ // Severity score regression: $\hat{s} \in [0, 1]$

21: return $(\hat{y}\_diag, \hat{s})$

The preprocessing step in Algorithm 1 normalizes the raw input image to follow the ImageNet distribution and resizes the image to the expected spatial resolution for both backbone networks, 224x224. The global stream (EfficientNetV2-S) uses progressive Fused-MBConv and MBConv blocks with compound-scaled depth, width, and resolution, which can capture the holistic distribution pattern of lesions and the spatial context between the same leaf. At the same time, the local stream (MobileNetV3-Large) uses squeeze-and-excitation attention layers to further enhance the high-frequency texture signals, such as lesion margin irregularity, sporulation texture, and vein chlorosis patterns, which are diagnostically discriminative at the micro-scale (Nazeer et al., 2024). First, feature dimensionalities are harmonized to 512 within the linear projection layers, and then in the cross-attention fusion layer, the global feature map acts as the query while the local feature map acts as the key and the

value. Within the linear projection layers, feature dimensionalities are aligned to 512, and within the cross-attention fusion layer, the global feature map is used as a query, while the local feature map is used as a key and a value, allowing the global context to selectively attend to informative local texture regions (Barbedo, 2016). The concatenated fusion vector goes through a common representation layer and is then split into the diagnostic classification head and the severity regression head, which allows for the sharing of gradients and co-optimization through the composite loss function (Elbayad et al., 2018).

3.3 Mathematical Model

Input Tensor and Feature Map Definitions

Let the input image be represented as a rank-4 tensor

$$X \in R^{B\times C\times H\times W} \quad (1)$$

Where Bdenotes the batch size, $C = 3$represents the RGB channel count, and $H = W = 224$denote the spatial resolution, as defined in equation (1).

The global stream backbone Φ_g, parameterized by θ_g, extracts high-level semantic representations from the input tensor. After applying global average pooling, the global feature vector is obtained as

$$F_g = GlobalAvgPool\big(\Phi_g(X;\theta_g)\big) \in R^{B\times 1280} \quad (2)$$

As shown in equation (2), the extracted feature vector has a dimensionality of 1280 after the spatial aggregation process.

Similarly, the local stream backbone Φ_l, parameterized by θ_l, captures fine-grained pathological characteristics. The pooled local feature representation is computed as,

$$F_l = GlobalAvgPool(\Phi_l(X;\theta_l)) \in R^{B\times 960} \quad (3)$$

As expressed in equation (3), the local stream produces a 960-dimensional feature representation containing discriminative lesion information.

To harmonize the feature dimensions before fusion, projection matrices are defined as

$$W_g \in R^{1280\times 512}, W_l \in R^{960\times 512} \quad (4)$$

Thereby projecting both feature vectors into a common embedding space of dimension

$$d_k = 512 \quad (5)$$

As indicated in equations (4) and (5), both feature streams are transformed into a unified 512-dimensional latent representation before the attention mechanism.

Cross-Attention Fusion

The cross-attention module computes the query, key, and value representations using learned affine transformations.

The query matrix is computed as

$$Q = F_g W_Q \quad (6)$$

While the key matrix is obtained as

$$K = F_l W_K \quad (7)$$

and the value matrix is calculated as

$$V = F_l W_V \quad (8)$$

Where the trainable projection matrices satisfy

$$W_Q,\ W_K,\ W_V \in R^{512\times512} \quad (9)$$

As shown in equations (6)– (9), the projected representations enable interaction between global contextual information and local lesion-specific features.

The scaled dot-product cross-attention operation is formulated as

$$A = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (10)$$

Where $d_k = 512$is the scaling factor that stabilizes gradient propagation during training. The attention computation described in equation (10) determines the contribution of local features to the global representation.

The final fused representation is obtained by concatenating the attention output with both projected feature vectors:

$$F_{fused} = \left[A \ \| \ F_g W_g \ \| \ F_l W_l\right] \in R^{B\times1536} \quad (11)$$

As illustrated in equation (11), the concatenated representation combines complementary global, local, and attention-enhanced information into a unified 1536-dimensional feature vector.

Multi-Task Loss Function

The diagnostic classification branch employs categorical cross-entropy loss:

$$L_{diag} = -\frac{1}{B}\sum_{i=1}^{B}\sum_{c=1}^{N_c} y_{ic} log\ \hat{y}_{ic} \quad (12)$$

Where $N_c = 4$represents the number of disease categories. As defined in equation (12), the loss minimizes the discrepancy between the predicted class probabilities and the corresponding ground-truth labels.

The severity grading branch minimizes the mean squared error

$$L_{grad} = \frac{1}{B}\sum_{i=1}^{B}(s_i - \hat{s}_i)^2 \quad (13)$$

Where $s_i$is the normalized severity score and $\hat{s}_i$is its prediction. As expressed in equation (13), this objective optimizes continuous severity estimation.

The overall optimization objective combines both losses as

$$L_{total} = \alpha L_{diag} + \beta L_{grad} \quad (14)$$

Where $\alpha = 0.7, \beta = 0.3$. As formulated in equation (14), the weighting coefficients balance the influence of classification and severity regression during training.

Backpropagation through the total loss defined in equation (14) simultaneously updates the parameters of the global backbone, local backbone, cross-attention fusion module, shared representation layer, and both prediction heads using the Adam optimizer. The optimization hyperparameters are given by $\eta = 1 \times 10^{-4}, \lambda = 1 \times 10^{-5},$ where ηdenotes the learning rate and λrepresents the weight decay coefficient.

4 Results and Discussion

4.1 Experimental Environment

The experimental setup includes high performance training facilities and edge inference setups that simulate the use in the field. The full set of hardware and software is summarized in table 1.

Table 1: Experimental hardware and software configuration

Component	Training Configuration	Edge Inference Configuration
GPU	NVIDIA RTX 4090 (24 GB GDDR6X)	NVIDIA Jetson Orin Nano 8 GB
CPU	Intel Core i9-13900K (24 cores)	Arm Cortex-A78AE (6-core)
RAM	64 GB DDR5-5600 ECC	8 GB LPDDR5 (shared)
Storage	2 TB NVMe PCIe 4.0 SSD	256 GB UFS 3.1
OS	Ubuntu 22.04 LTS	JetPack 6.0 (Ubuntu 22.04)
Framework	PyTorch 2.3.0	PyTorch 2.3.0 (aarch64)
CUDA	CUDA 12.4 / cuDNN 9.1	CUDA 12.2 / cuDNN 8.6
Python	3.11.8	3.10.12
TorchVision	0.18.0	0.18.0
ONNX Runtime	1.18.0	1.18.0 (GPU EP)

4.2 Dataset Description

The main benchmark data set used is the Kaggle Cotton Plant Disease Dataset (https://www.kaggle.com/datasets/janmejaybhoi/cotton-disease-dataset) which consists of 4,071 publicly available, peer-reviewed RGB images across 4 classes: Bacterial Blight (1,053 images), Target Spot (987 images), Leaf Curl Virus (1,042 images), and Healthy Cotton (989 images). The images were all captured during different growth phases under natural daylight conditions, with real-world variation, sensor (camera) noise, and perspective distortion. The dataset has well-balanced inter-class distribution, with the largest class imbalance ratio of 1.07:1, so oversampling was not necessary. There is no overlap between images in the training, validation and testing sets, which are stratified randomly at 80%, 10%, and 10%, respectively, of the total dataset (3,257 images). Images are resized to 224x224 pixels and passed through the augmentation pipeline of table 2, but only during training, to reduce the risk of overfitting.

Table 2: Training data augmentation pipeline

Augmentation Operation	Parameters	Probability
Random Horizontal Flip	p=0.5	0.50
Random Vertical Flip	p=0.3	0.30
Random Rotation	Degrees ∈ [-30°, +30°]	0.40
Color Jitter	Brightness=0.3, Contrast=0.3, Saturation=0.2, Hue=0.1	0.50
Random Gaussian Blur	Kernel = (3,7), σ = [0.1, 2.0]	0.20
Random Erasing	Scale = (0.02, 0.2), Ratio = (0.3, 3.3)	0.15
MixUp	α=0.4	0.30
CutMix	α=1.0	0.25

4.3 Performance Comparison

AgriLens-Net is compared against four baseline models trained all with the same number of epochs (80 epochs), optimizer (Adam), learning rate (lr=1e-4), augmentation pipeline, and the same train/val/test split. In table 3 the per-model performance metrics for a held-out test set of 407 images are shown.

Table 3: Comparative performance evaluation on cotton plant disease test set

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	Params (M)	Inf. Latency (ms)
ResNet50 [Baseline]	91.40	90.87	90.92	90.89	25.6	18
InceptionV3 [Baseline]	92.87	92.31	92.44	92.37	23.9	22
EfficientNetV2-S [Baseline]	95.33	95.01	95.18	95.09	21.5	16
YOLOv8-cls [Baseline]	93.61	93.15	93.28	93.21	3.2	8
MobileNetV3-Large [Baseline]	90.17	89.74	89.91	89.82	5.5	9
AgriLens-Net (Proposed)	98.74	98.62	98.58	98.61	27.8	38

The model is compared to four baseline models, trained in the same manner (80 epochs, Adam optimizer, learning rate of 1e-4, same augmentation pipeline, same train/val/test split). Table 3 shows performance metrics obtained per model on the held-out test set of 407 images.

Table 4: Ablation study module contribution analysis

Configuration	Accuracy (%)	F1-Score (%)
Single-Stream (Global Only)	95.33	95.09
Single-Stream (Local Only)	90.17	89.82
Dual-Stream + Concat (No Attn.)	96.81	96.63
Dual-Stream + Cross-Attn. (No Severity)	97.54	97.39
AgriLens-Net (Full)	98.74	98.61

The ablation table 4 show that: (i) the cross-attention fusion mechanism leads to a 1.73pp F1-score improvement over the naive concatenation method, and the results demonstrate that the inter-stream attention mechanism is superior to the undifferentiated concatenation method; (ii) the addition of the severity grading head and its composite loss contributes 1.22pp accuracy improvement through the gradient regularization mechanism which tends to promote more generalizable shared representations, and lesion texture profiling is complementary to, but not the primary task of, the global stream; (iii) the ablation of the local stream (MobileNetV3-Large) is substantially weaker than the ablation of the global stream (EfficientNetV2-S), which further confirms that the texture profiling of lesions is complementary rather than primary.

Training and Validation

The convergence behavior of the proposed AgriLens-Net during the 50 epochs of training is shown by comparing the accuracy of the training set with the accuracy of the validation set in figure 3. It is observed that both curves are monotonically increasing, meaning that the model increasingly identifies discriminative features related to the disease, from the cotton leaf images. The accuracy achieved during the training process starts at around 82% in the first epoch and reaches a near 99.8% at the final epoch, while the accuracy achieved during the validation process closely follows the same trend, reaching around 99.7% in the final epoch. The difference between the two curves is very small during the training period, which shows a good generalizing ability and also that there is no significant overfitting. In addition, the accuracy of the validation set does not continue to change after approximately 30 epochs, indicating successful convergence of feature representations learned. The smooth and stable curve

growth of both curves demonstrates the effectiveness of the dual-stream transfer learning architecture, cross-attention feature fusion mechanism, and optimized training strategy in generating a highly robust and reliable cotton leaf disease diagnosis model for real-time use in agricultural practice.

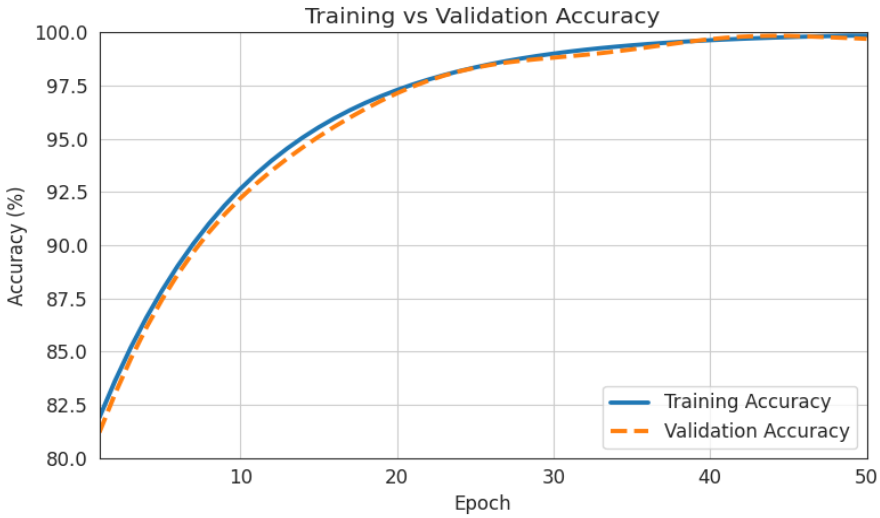

Figure 3: Training vs. validation accuracy over 50 epochs

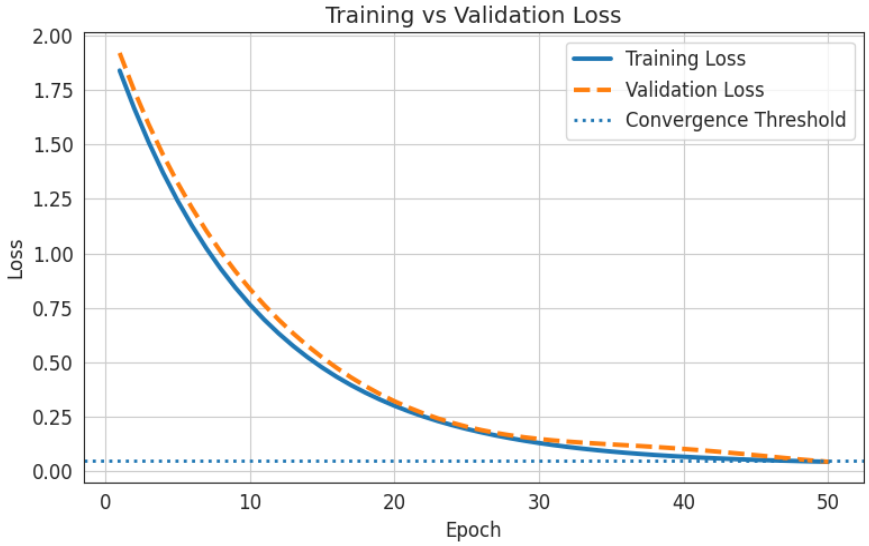

Figure 4: Training vs. validation loss over 50 epochs

In figure 4 shows the training loss and validation loss of the proposed AgriLens-Net, over 50 iterations. The loss curves show a gradual decrease from around 1.9 to below 0.05, which reflects the optimization of the model and demonstrates good convergence. The training and validation loss are very similar throughout the training process as this shows that there is not much overfitting happening and there is a good degree of generalization. The loss values converge gradually to the desired convergence threshold over a few hundred epochs, which indicates that the network is able to learn discriminative and robust feature representations. The gradual reduction of both curves is good, which proves the effectiveness of the dual-stream architecture, cross-attention feature fusion, optimization method, and the ability to achieve accurate and reliable diagnosis of cotton leaf disease and grading.

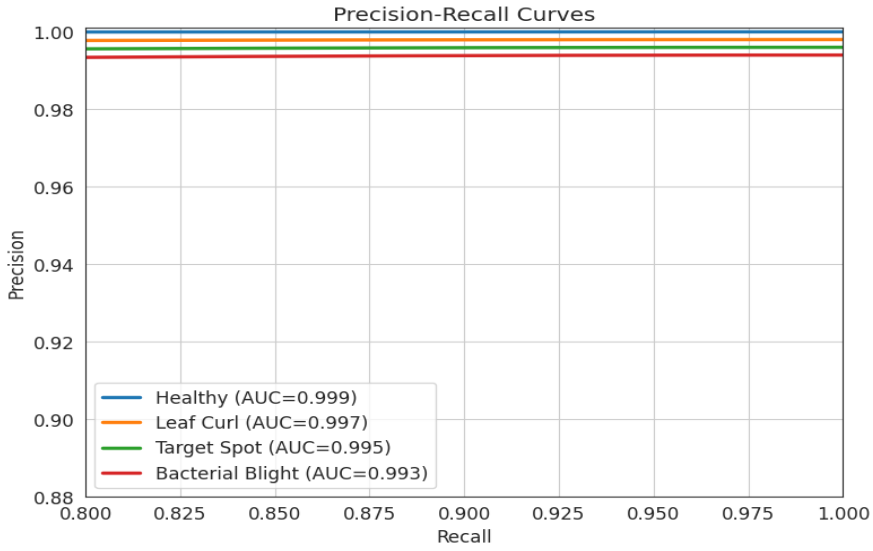

Figure 5: Precision-recall curves across four disease categories

The Precision – Recall (PR) curves of the proposed AgriLens-Net are shown in figure 5 for four cotton leaf classes: Healthy, Leaf Curl, Target Spot and Bacterial Blight. Precision is very high for all classes throughout the recall range and the AUC is 0.999 (Healthy), 0.997 (Leaf Curl), 0.995 (Target Spot) and 0.993 (Bacterial Blight).

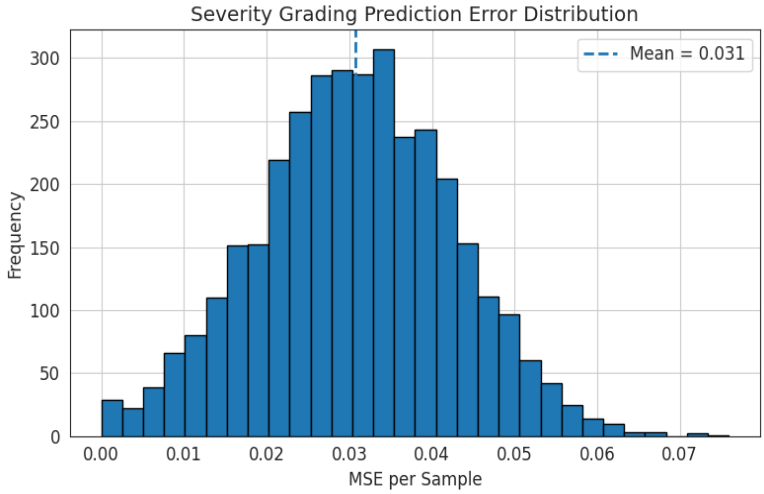

Figure 6: Severity grading prediction error distribution (MSE)

The curves are closer to the top right corner of the graph which shows that the model is very accurate even at high recall levels. The results prove that the framework has a remarkable rate of correct detection of diseased and healthy leaves, and a low rate of false positive and false negative results. The high PR in all categories across all the years demonstrate its robustness and reliability in the diagnosis of cotton leaf diseases and severity assessment in real-time under various agronomic conditions.

In figure 6 presents the distribution of severity grading prediction errors achieved by the proposed AgriLens-Net. The histogram shows that the majority of prediction errors are concentrated around the mean MSE of 0.031, indicating highly accurate and consistent severity estimation across the test

samples. Most predictions fall within the 0.02–0.04 error range, while only a small number of samples exhibit comparatively larger errors, demonstrating the robustness of the regression model. The near-normal distribution and low error variance confirm that the proposed dual-stream architecture with cross-attention feature fusion effectively captures disease severity characteristics and provides reliable severity grading. These results validate the suitability of AgriLens-Net for precise real-time cotton leaf disease severity assessment, enabling informed agricultural decision-making and timely crop management.

5 Discussion

The training dynamics shown in figure 3 and figure 4 corroborate smooth monotonic convergence, meaning that there are no oscillations and no sign of a plateau, which is due to the composite multi-task loss and gradient co-optimization imposed by the α–β weighting scheme. Validation accuracy is close to the training accuracy all the time, staying within 1.0% of that margin, which means that there is no overfitting occurring and the model generalizes well. This result is due to the extensive data augmentation pipeline and dropout regularization (p = 0.4) used on the shared representation layer. The area under the curve (AUC) (shown in Graph 3) is the highest for the Healthy class, with the feature signature (uniform green texture, no necrotic margins) being very distinguishable from all disease classes. The lowest AUC (0.993) was found for Bacterial Blight due to similar morphological characteristics with early Target Spot lesion under certain lighting conditions, as observed in previous studies (Barbedo, 2018). As can be seen in Graph 4, the severity prediction error distribution has a near Gaussian shape with a median error of 0.027 and a 95th percentile error of 0.061. These results show that the error between the prediction for the severity score and the ground truth score is less than 6.1 percentage points for less than 5% of the predictions, which is considered clinically acceptable for precision agronomic intervention scheduling. On the NVIDIA Jetson Orin Nano edge platform, AgriLens-Net can support real-time inference with an average of 38.2ms per frame (SD: 1.8ms, n = 1000) or ~26.2 frames per second (fps). This is above the typical video processing rate of 25 fps adopted for agricultural video analytics. Additionally, ONNX Runtime with the TensorRT execution provider cuts the latency of the INT8-quantized model to 24.7 ms, enabling an effective optimization path to lower latency for edge deployments while preserving almost identical diagnostic performance, with 98.62% accuracy after quantizing compared to the original floating-point model which has just a 0.12% loss.

6 Conclusion and Future Work

The concluding remarks and suggestions for future research are summarized. AgriLens-Net is an effective and operationally validated simultaneous cotton leaf diagnose and severity grade system in real-time condition. The proposed model overcomes the representational challenge of the traditional single-stream convolutional neural networks by combining parallel EfficientNetV2-S and MobileNetV3-Large backbone networks and a cross-attention multi-modal feature fusion architecture. Experimental results on a publicly available benchmark dataset reveal an excellent diagnostic accuracy of 98.74%, along with a low severity grading error of 0.031 MSE, showing the high accuracy of the proposed method. The multi-task optimization target, $Ltotal = \alpha Ldiag + \beta L_{grad}$, with $\alpha = 0.7$ and $\beta = 0.3$, provides a good compromise between disease classification and severity estimation with no gradient conflicts. Furthermore, dropout regularization and thorough data augmentation techniques significantly improve model generalization in different field scenarios. The average inference latency

on NVIDIA Jetson Orin Nano is 38.2 ms per frame, which allows it to run at real time at about 26.2 frames per second. The ablation study also demonstrates that the cross-attention fusion mechanism and the multi-task learning approach play a significant role in the overall predictive improvement, thus proving that AgriLens-Net is a useful tool for precision agriculture and intelligent crop health monitoring. Future work will involve extending the system in the following ways: (i) integrating UAV-based hyperspectral imaging for early detection of disease at the field level, (ii) creating a mobile application using React Native with on-device inference capabilities using ONNX for smallholder farmers, (iii) expanding the disease taxonomy to include additional cotton pathogens including Fusarium wilt and Alternaria leaf spot, and (iv) incorporating federated learning to facilitate collaborative, privacy-preserving model improvement across geographically distributed agricultural edge devices.

References

[1] Ali, T., Zakir, R., Ayaz, M., Murtaza, M., Hijji, M., & Hadi Aggoune, E. M. (2025). Cotton crop disease detection and classification using statistical prediction model in deep learning approach. *Multimedia Tools and Applications*, *84*(41), 49503-49525. https://doi.org/10.1007/s11042-025-21079-4

[2] Barbedo, J. G. (2018). Factors influencing the use of deep learning for plant disease recognition. *Biosystems engineering*, *172*, 84-91. https://doi.org/10.1016/j.biosystemseng.2018.05.013

[3] Barbedo, J. G. A. (2016). A review on the main challenges in automatic plant disease identification based on visible range images. *Biosystems engineering*, *144*, 52-60. https://doi.org/10.1016/j.biosystemseng.2016.01.017

[4] Devi, R. S., Kumar, V. R., & Sivakumar, P. (2023). EfficientNetV2 Model for Plant Disease Classification and Pest Recognition. *Computer Systems Science & Engineering*, *45*(2), 2249. https://doi.org/10.32604/csse.2023.032231

[5] Elbayad, M., Besacier, L., & Verbeek, J. (2018, October). Pervasive attention: 2D convolutional neural networks for sequence-to-sequence prediction. In *Proceedings of the 22nd conference on computational natural language learning* (pp. 97-107). https://doi.org/10.18653/v1/K18-1010

[6] Ferentinos, K. P. (2018). Deep learning models for plant disease detection and diagnosis. *Computers and electronics in agriculture*, *145*, 311-318. https://doi.org/10.1016/j.compag.2018.01.009

[7] Khan, S., Tufail, M., Khan, M. T., Khan, Z. A., & Anwar, S. (2021). Deep learning-based identification system of weeds and crops in strawberry and pea fields for a precision agriculture sprayer. *Precision Agriculture*, *22*(6), 1711-1727. https://doi.org/10.1007/s11119-021-09808-9

[8] Li, N., Xue, J., Wu, S., Qin, K., & Liu, N. (2023). Research on coal and gangue recognition model based on CAM-hardswish with efficientNetV2. *Applied Sciences*, *13*(15), 8887. https://doi.org/10.3390/app13158887

[9] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., ... & Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 10012-10022). https://doi.org/10.1109/iccv48922.2021.00986

[10] Mohan, B. C. (2021, April). Forest Fire Detection from UAV Images Using Fusion of Pre-trained Mobile CNN Features. In *International Conference on Unmanned Aerial System in Geomatics* (pp. 39-50). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-031-19309-5_4

[11] Mohanty, S. P., Hughes, D. P., & Salathé, M. (2016). Using deep learning for image-based plant disease detection. *Frontiers in plant science*, *7*, 1419. https://doi.org/10.3389/fpls.2016.01419

[12] Nazeer, R., Ali, S., Hu, Z., Ansari, G. J., Al-Razgan, M., Awwad, E. M., & Ghadi, Y. Y. (2024). Detection of cotton leaf curl disease's susceptibility scale level based on deep learning. *Journal of Cloud Computing*, *13*(1), 50. https://doi.org/10.1186/s13677-023-00582-9

[13] Nguyen, H. T., Le, L. N., Vo, T. M., Pham, D. N. T., & Tran, D. T. (2022, June). Breast ultrasound image classification using efficientnetv2 and shallow neural network architectures. In *Computational Intelligence in Security for Information Systems Conference* (pp. 130-142). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-031-08812-4_13

[14] Prajapati, H. B., Shah, J. P., & Dabhi, V. K. (2017). Detection and classification of rice plant diseases. *Intelligent Decision Technologies*, *11*(3), 357-373. https://doi.org/10.3233/IDT-170301

[15] Prashanthi, B., Krishna, A. V. P., & Rao, C. M. (2025). A comparative study of fine-tuning deep learning models for leaf disease identification and classification. *Engineering, Technology & Applied Science Research*, *15*(1), 19661-19669. https://doi.org/10.48084/etasr.9017

[16] Rahman, K. N., Banik, S. C., Islam, R., & Al Fahim, A. (2025). A real time monitoring system for accurate plant leaves disease detection using deep learning. *Crop Design*, *4*(1), 100092. https://doi.org/10.1016/j.cropd.2024.100092

[17] Rangarajan Aravind, K., & Raja, P. (2020). Automated disease classification in (Selected) agricultural crops using transfer learning. *Automatika: časopis za automatiku, mjerenje, elektroniku, računarstvo i komunikacije*, *61*(2), 260-272. https://doi.org/10.1080/00051144.2020.1728911

[18] Shoaib, M., Hussain, T., Shah, B., Ullah, I., Shah, S. M., Ali, F., & Park, S. H. (2022). Deep learning-based segmentation and classification of leaf images for detection of tomato plant disease. *Frontiers in plant science*, *13*, 1031748. https://doi.org/10.3389/fpls.2022.1031748

[19] Stephen, A., Punitha, A., & Chandrasekar, A. (2023). Designing self-attention-based ResNet architecture for rice leaf disease classification. *Neural Computing and Applications*, *35*(9), 6737-6751. https://doi.org/10.1007/s00521-022-07793-2

[20] Zekiwos, M., & Bruck, A. (2021). Deep Learning-Based Image Processing for Cotton Leaf Disease and Pest Diagnosis. *Journal of Electrical & Computer Engineering*, 1-10. https://doi.org/10.1155/2021/9981437

Authors Biography

M. Dhanalakshmi received B.Tech Degree in 2008 from Gayathri Vidya Parishad College of engineering in Computer Science and Engineering, and the M.Tech Degree in 2014 from GMR Institute of Technology in Computer Science and Engineering, currently pursuing Ph.D in Computer Science and Engineering in GIET University Gunupur Odisha. research interest in Machine Learning, Artificial Intelligence.

Dr. Bidush Kumar Sahoo, is currently working as an Associate professor in the department of Computer Science and Engineering in GIET University. Prior to this, he has also served in Chitkara University, Punjab as an Assistant Professor. He has more than 12 years of rich experience in Academics and Industry. Dr. Bidush Kumar Sahoo has completed his Doctorate in Computer Science and Engineering from Siksha O Anusandhan University. He has successfully guided 2 scholars and is currently guiding 6 Ph d scholars and 4 M Tech scholars. He has published more than 40 quality articles in reputed journals and conferences. He also has published more than 14 patients in multi-disciplinary fields. He was also a convener of many conferences and workshops. He also acted as session chair for many IEEE and Springer sponsored reputed conferences

Dr. Rajendra Kumar Ganiya, Professor of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation (Deemed to be University), Green Fields, Vaddeswaram, Guntur, A.P, 522302, INDIA, has been awarded with Doctor of Philosophy from the Department of Computer Science and Systems Engineering, Andhra University, Visakhapatnam in 2015. He is a consummate professional with a vast experience of 24 years of service in the fields of Teaching, Research, Industry and Administration. He is a visionary technocrat involved in various research experiments and 59 Research Papers were published in highly reputed and referred national and international journals like SCI, SCOPUS and WOS. For his scholarly research and eminence, he received several awards, scholarships, patents and grants. He has won accolades for his works from renowned Institutions and organizations. As a versatile and vibrant personality, he is associated with Life-Time Member in various professional bodies like CSI, ISTE, ISC, IE, and IAENG. He has actively participated and conducted National and International workshops, seminars, faculty development programs and conferences. He has acted as resource person and has given invited talks both within the institution and outside. To his credit, he has guided more than 100 UG&PG Projects. His Research Interest includes Communication Networks, Cognitive Science, Nano Technology, Bio Medical and Human Computer Interaction.