Journal of Intelligent Computing System (JICS), Vol.1, Issue. 3, pp. 21-39, (Sep 2026)
ISSN: 3139-3616 (online)

Explainable AI for Precision Agriculture: Fine-Grained Plant Pathology Localization Using ConvNeXt-Tiny and Residual Spatial Attention Module

Palvi Sharma
Department of CEA, GLA University, Mathura, U.P, India
spalvi625@gmail.com

Abstract: Multi-crop plant disease automatic classification in precision agriculture requires an effective solution but the state-of-the-art deep learning architectures are prone to suffering background shortcut problem, higher inference latency, and inadequate visualization interpretability. This paper presents a better performing framework combining ConvNeXt-Tiny with an innovative Residual Spatial Attention Module (RSAM) to solve the fine-grained diagnosis problem for 38 plant pathology targets. On a dataset of 10,876 unseen images, the proposed framework demonstrates a Top-1 classification accuracy of 96.85%, a precision of 96.95%, a recall of 96.48%, and a macro F1-score of 96.71% while being superior to ResNet-50 (+2.20%) and being 0.97 ms faster per image (7.15 ms/img with 28.12 M parameters). Explainable AI (XAI) analysis with the help of Grad-CAM shows the ability of the RSAM block to suppress background and soil artifacts, directing 88.42% of activation energy E_{lesion} into lesion areas and producing insignificant background suppression effect of 1.12%. The t-SNE projections of latent feature manifolds demonstrate good inter-class clustering.

Keywords: *Plant Pathology Classification, ConvNeXt-Tiny, Residual Spatial Attention Module (RSAM), Explainable AI (XAI), Grad-CAM, Model Interpretability, Pareto Efficiency.*

1. Introduction

Agriculture is critical to economic stability and sustenance of life, providing food security to a growing population. Nonetheless, crop disease represents a perennial menace to the overall productivity in agriculture and leads to losses worth between 20% and 40% globally as reported by the Food and Agricultural Organization (FAO). Fungi, bacteria, and viruses affect the health of crops, drastically reducing their biomass and sometimes leading to crop failure in extremely adverse situations [1]. Diagnosis of plant disease is generally achieved through direct visual inspection by agricultural extension workers, agronomists, or plant pathologists. Despite visual scouting remaining a classical method of detection, it is associated with significant logistical problems. Extensive and costly manual inspection over large geographical areas is impractical [2]. Moreover, visual diagnosis by humans is subjective, subject to physical fatigue and confusion with visual symptoms that are difficult to distinguish for different plant diseases like early and late blight. In rural and underdeveloped settings, smallholder farmers do not have timely access to plant pathologists and treatment is delayed leading to irreversible damage to crops. In an effort to address the problem of scaling of manual diagnosis, computer-assisted diagnostic systems that use computer vision (CV) along with deep learning (DL) have been developed to be indispensable elements of modern precision agriculture. In the last ten years, computer-assisted plant pathology has shifted from engineered feature extraction to deep representation learning [3] – [4]. Initial computer vision approaches used the manual extraction of color histograms, texture descriptors such as Local Binary Patterns (LBP) and Gray-Level Co-occurrence Matrices (GLCM) as well as geometric shape descriptors together with classical machine learning classifiers, SVM and Random Forests. Though fundamental, those initial techniques lacked invariance to the real world conditions of varying sunlight, shadows, background noise, and leaf pose variations. The arrival of CNNs was revolutionary for the processing of agricultural images as they performed automatic feature extraction from RGB pixels [5]. The legacy models such as VGG-16, ResNet-50, and DenseNet exhibited high classification accuracy on the benchmark datasets by performing low-level edge detection at the shallow levels and high-level semantics of pathology at the deep levels [6]. The Figure 1 shows the random samples of the PlantVillage Color dataset.

The use of lightweight mobile networks including MobileNetV3 used depthwise separable convolutions for fewer parameters for edge deployments while the latest convolutional backbones included ConvNeXt which optimized the efficiency of CNNs to the large receptive fields of Vision Transformers (ViTs). Despite these developments in architecture, deep learning model deployment for accurate plant pathology identification poses research problems [7] – [9]. Firstly, traditional CNNs treat background visual information such as soil, dry stems, man-made structures, and shadow of light with equal importance as that of the visual information of diseased leaves. This results in the memorization of background visual artifacts by neural networks instead of disease symptoms

© 2026 JICS. Personal use is permitted, but republication/redistribution requires JICS permission.
see: https://synergyworldpress.com/journal/jics/about/overview

doi.org/10.67420/109319.1.3.2

which ultimately leads to poor performance when these models are deployed in the real world [10] – [12]. Secondly, there is a problem of accuracy versus latency trade-off since heavy models perform well but consume a lot of memory and time while very lightweight models are fast but cannot represent well enough. Third, deep learning algorithms are regularly cited as being opaque "black-box" systems [10]. In practice, within an agricultural setting, farmers and agronomists need to see proof through images that the algorithm's decision is based on lesion symptoms such as chlorotic halos, rust pustules, or necrotic spots rather than artifacts within the background image [13] - [14].

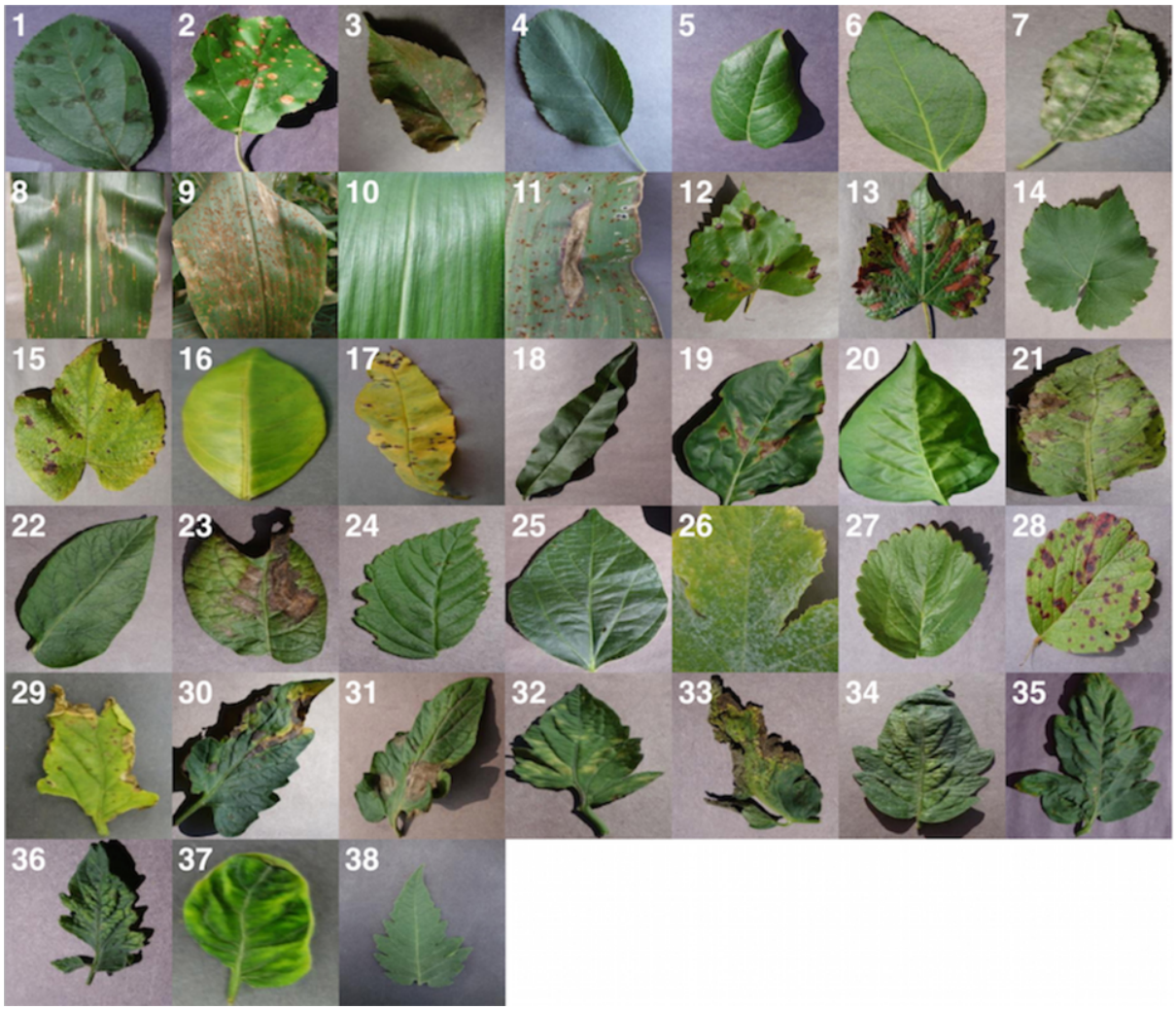


Figure 1. Random samples from the PlantVillage Color Dataset

To overcome these challenges, we introduced an end-to-end plant pathologies classification system that includes an updated ConvNeXt-Tiny model backbone coupled with a proprietary Residual Spatial Attention Module (RSAM). Our ConvNeXt-Tiny model is constructed by using 7X7 depthwise convolutions, inverted bottlenecks, and Layer Normalization, providing both the efficiency of typical CNNs along with highly expressive feature extraction capabilities inherent in modern vision backbones. In order to remove background artifact influence and improve lesion feature extraction, our RSAM block provides dynamic recalibration of the spatial features. RSAM mechanism ensures that the network learns to focus its receptive field on the necrotic symptoms only by using channel compressed 2D spatial attention mask along with residual skip connections. In order to overcome such limitations, we present an end-to-end plant pathology classification approach based on an integration of the light-weight Residual Spatial Attention Module (RSAM) along with ConvNeXt-Tiny architecture. The main contributions of this study can be enumerated in four points. First, we develop the RSAM architecture using dual-channel pooling and 7X7 spatial gating operations to filter out non-informative background artifacts and enhance informative lesion features. Second, we utilize the proposed spatial attention mechanism within a ConvNeXt-Tiny feature extractor and develop a compact and powerful classification network adapted to complex visual conditions of foliage. Third, we develop a quantitative evaluation protocol for Explainable AI (XAI) approaches based on

the lesion masks annotated by experts to assess the quality of spatial feature localization and background shortcut suppression capability. Finally, by conducting the analysis of multiple runs and testing on external uncurated agricultural image datasets, we prove the effectiveness and transferability of the suggested approach.

1.1 Organization of the Paper: The paper is structured as follows: In Section 2 presents the methodology involving data partitioning, data augmentation, structural architecture of the ConvNeXt-Tiny, and mathematically explains the RSAM block. Section 3 presents the methodology of the baseline model and proposed model. Section 4 discussed about the quantitative benchmark results, comparative performance analysis, training behavior, Pareto efficient frontiers, Grad-CAM visualization, and t-SNE projection of the feature manifolds. Finally, Section 5 highlights the key takeaways and future avenues for utilizing the model in real-world agriculture on mobile edge devices.

2. Related Work

The ongoing development of computer vision and deep learning technologies has completely revolutionized precision agriculture today by shifting crop disease diagnosis from manual and fallible field scouting practices to automated and high-throughput pipelines. Within the last ten years, the scientific literature in this field has grown dramatically through three interconnected areas of study: deep convolutional and vision transformer neural networks designed specifically for plant disease pathology, attention mechanisms that would work without any constraints in the field, and Explainable Artificial Intelligence (XAI) approaches that would help to verify the transparency of the models and gain the trust of the agricultural community.

2.1 Deep Learning Architectures in Plant Pathology: The deep neural network technology has broken the past bottleneck in automated plant pathology due to the process of manually extracting features by learning hierarchical representations straight from the raw pixel values of images. The lower layers learn the low-level spatial features such as edges and texture, whereas higher layers learn the high-level semantic representations for diseases in the form of halos, lesions, and pustules. In [15], S. C. K. et al. introduced an automated high-precision diagnostic workflow tailored for diagnosing leaf diseases in cardamom plants, targeting the most destructive fungal diseases such as Colletotrichum Blight and Phyllosticta Leaf Spot. In order to mitigate the adverse effect of background noise due to competing weeds, shadowing effects, and the soil surface in field captured images, the authors developed a two-stage neural network structure. First, a light-weight U^2 -Net Segmentation Network was developed to detect the leaves. Then, the detected leaf region is used as input to the EfficientNetV2 structure that uses Fused MBConv blocks with Progressive Learning Schedules. While this model has achieved a remarkable classification accuracy of 98.26%, the use of the full segmentation process before feature classification results in a high computational cost. In [16], S. Das and S.Biswas extended the scope of deep learning diagnostics from single-leaf-level imaging to large-scale agricultural surveillance through analysis of the dynamics of rice blast disease (Pyricularia oryzae) based on multiscale remote sensing images. They proposed DMFL-ConvoNet (Depth-wise Multiscale Feature Learning Convolutional Network) that performs parallel depth-wise separable convolutions at different scales of receptive fields to identify not only localized disease hotspots but also regional disease epidemics. Although the proposed model showed superior results in tracing the epidemiological dynamics over wide geographic regions, the problem is that the remote sensing image itself does not have millimeter scale resolution for early symptoms on individual leaves. In [17], M. J. Hasan et al., tackled the crucial problem of the domain shift phenomenon in practical applications of agriculture where the models that work well on the benchmark datasets do not perform satisfactorily in the new domains of their application. The problem was solved by creating a harmonized cross-domain benchmark by merging the PlantVillage and TomatoVillage datasets into 15 unified pathology classes. These researchers also conducted a detailed analysis of 24 of the most advanced vision backbones and came up with a post-training optimization pipeline for multi-model ensembling, knowledge distillation of a teacher-student pair, and INT8 post-training quantization. Although the quantized distilled models demonstrated excellent cross-domain performance with little accuracy reduction, the initial training part entailed multi-model ensembling. In [18], D. Thakur et al., emphasized multi-class pathology classification for coffee plants, specifically concentrating on very destructive pests and diseases like coffee leaf miner (Leucoptera coffeella), cercospora leaf spot (Cercospora coffeicola), and coffee leaf rust (Hemileia vastatrix). They proposed SUNet, which is an encoder-decoder network consisting of both structural elements of U-Net and SegNet along with a pre-trained VGG-16 model as their main feature encoder. SUNet implements dense skip connection for the direct transmission of low-level spatial boundary information to decoder layers, thereby preserving the high frequency lesion boundaries. But employing a VGG-16 encoder results in very large number of parameters (more

than 138 million parameters when fully implemented) leading to high memory utilization and energy consumption.

2.2 Attention Mechanisms in Agricultural Vision: Conventional convolution networks treat all the spatial regions in an input image equally in terms of spatial weights. In nature, infected leaf regions constitute just a part of the entire scene, with non-informative background regions (soil, mulch, rocks, and artificial objects) making up the rest of the input. Attention modules overcome this problem by selectively tuning the intermediate representations, enabling the model to focus its capabilities on lesion regions, while ignoring the uninformative background regions. In [19], X. Duan et al. studied cross-modal attention using an interactive attention model for the classification of multilingual Chinese agricultural news. This model deals with the semantic confusion that exists between unstructured text information and the complicated images of the agriculture field. They process the text signals using pre-trained ERNIE text models and process the image signals using the Vision Transformers (ViT). They use a cross-attention interaction matrix to combine the two modal embeddings. Although their interactive attention model successfully dealt with the problem of semantic conflicts between text and images for text categorization, their design was for text-image matching tasks and not lesion detection. In [20], L. Shi et al. introduces MSVM-UNet that uses multi-scale spatial attention to facilitate accurate segmentation of diseases on plants in complicated field environments. In order to implement long range spatial dependencies without quadratic computation of self-attention, Visual State Space (VSS) blocks inspired by Mamba architecture along with Multi-Scale Spatial Attention (MSSA) blocks within U-Net skip connections were used. The proposed approach was able to achieve remarkable 80.44% mean Intersection over Union (mIoU) on difficult datasets consisting of crop leaves with irregular and overlapping lesions. However, the recurrent scanning through state space in Mamba blocks is not efficiently performed by typical mobile and embedded platforms. In [21], S. Zhang et al. proposed AgriCapsNet which is an attention guided capsule network designed specifically for multi-crop classification in agriculture through multispectral satellite images. CNNs generally fail to maintain spatial hierarchies because of the max-pooling operation, resulting in a loss of performance when differentiating visually and spectrally similar crop canopies. AgriCapsNet solves this problem by incorporating spatial spectral attention blocks in dynamic vector capsule routing to preserve directional spatial hierarchies in the multispectral bands. Although AgriCapsNet has delivered state-of-the-art land cover classification results, the matrix multiplication overhead in the iterative dynamic routing process is significant. In [22], F. Gou et al. proposed the CCASF-YOLOv10 target detection system for the real-time monitoring of growth stages and canopy structure of lettuce plants in automatic greenhouses. They incorporated a channel and spatial attention feature fusion (CCASF) block in the neck region of a lightweight YOLOv10 framework so that the neural network could sustain its feature representation despite different conditions of greenhouse lightings, overlapping of leaves, and self-shading in the canopies. Although CCASF-YOLOv10 performed efficient object detection by bounding box method, its architecture was optimized for objects' bounding boxes only and not for high density pixel-level features of micro-pathologies.

2.3 Explainable AI (XAI) and Visual Interpretability in Precision Agriculture: With the incorporation of more and more deep learning models within the intelligent agricultural system, their "black box" characteristic poses threats in terms of deployment. The agronomists, the extension agents, and the farmers need proof that the prediction made by the model is based on genuine biological symptoms (for example, the presence of chlorotic rings or rust pustules), and not image features. This can be achieved through the use of Explainable Artificial Intelligence (XAI). In [23], A. Navarro attempted to bridge the divide between XAI results and user comprehension by designing an MDE solution that could take user requirements and produce custom XAI visual representations in agricultural domains. They systematically classify the non-technical user needs from smallholder farmers looking for immediate diagnosis to agronomists studying broader patterns of diseases and match them to relevant local (Grad-CAM, LIME) or global (SHAP summary plot) interpretability interfaces. Although their approach offers a complete software engineering methodology for accessing XAI results, it is a design model and does not include convolutional feature extraction methods for pathogenic classification. In [24], M. F. Celik et al. presented the ability of Tabular XAI techniques to perform macro-level crop yield modeling by creating an explainable framework for yield prediction of cotton farming regionally throughout the continental United States. The study incorporated various satellite-based vegetation indices (NDVI, EVI) along with time-series data of meteorological and soil moisture conditions for training EBM and SHapley Additive exPlanations (SHAP) models. The interpretability analysis carried out by the authors identified precise phenological windows when heat stress and moisture deficiency adversely affected the yield of cotton crops. However, tabular XAI feature attributions are inherently designed for aggregate numerical arrays and are not directly applicable to 2D feature maps of RGB leaf images. In [25], S. Z. Khan et al. presented an integrated computer vision approach for the management of date palms, which jointly assessed the classification of the stages of ripening of date fruits and

the grading of severity of the White-Scale Disease (Parlatoria blanchardi) of date palms. The author used fine-tuned baseline CNN architectures such as VGG-16 and MobileNetV2, Khan et al. performed a multi-modal XAI audit of LIME, SHAP, Grad-CAM, and Grad-CAM++. The results from their analysis showed that whereas legacy baseline CNN architectures attained more than 95% nominal test accuracy, the post-hoc Grad-CAM heat maps highlighted non-crust regions such as palm fronds and soil instead of the actual scale insect crusts. In [26], A. Al-Hchaimi et al. utilized the concept of XAI to hardware systems by presenting a reinforcement learning-based genetic algorithm for mapping application tasks to processors in a Network on Chip (NoC) architecture used in smart farming IoT edge networks. The inclusion of the SHAP feature attribute into the fitness function of the genetic algorithm helped to determine the optimum assignment of processing cores, thereby leading to an energy savings and low latency. Even though their model is geared towards optimizing performance at the hardware level in constrained IoT devices, it is a system scheduling framework rather than image recognition.

To emphasize the essential differences between the current literature and the proposed model, Table 1 shows an organized summary of the reviewed articles in terms of their major methodologies, their principal strengths in operation, structural weaknesses, and the research gaps they have identified.

Table 1. Methodological Mapping of Cited Literature vs. Proposed Architecture

Ref.	Primary Application Area	Core Architectural / Algorithmic Focus	Primary Technical Strength	Structural Limitation / Literature Gap Addressed
[15]	Cardamom Pathology	U^2 - Net + EfficientNetV2	High background clutter isolation	Dual-stage pipeline increases overall latency and memory footprint.
[16]	Rice Blast Dynamics	DMFL-ConvoNet Remote Sensing	Macro-scale epidemiological tracking	Satellite resolution lacks millimeter-scale leaf lesion spatial details.
[17]	Cross-Domain Tomato Pathology	Ensemble + Distillation + Quantization	High cross-domain generalization	Multi-model distillation pipeline demands heavy training compute.
[18]	Coffee Leaf Pathology	SUNet (U-Net + SegNet + VGG16)	Preserves high-frequency spatial edges	Heavy VGG-16 backbone limits real-time mobile inference.
[19]	Multimodal Ag News	ERNIE + ViT Interactive Attention	Resolves text-image semantic ambiguity	Designed for document classification, not spatial leaf lesion mapping.
[20]	Disease Spot Segmentation	MSVM-UNet (Mamba + Spatial Attention)	Captures long-range spatial context	State-space operators require specialized edge hardware accelerators.
[21]	Multi-Crop Satellite Mapping	AgriCapsNet (Capsule Net + Attention)	Preserves spatial-spectral hierarchies	Dynamic capsule routing introduces heavy matrix multiplication overhead.

[22]	Greenhouse Lettuce Tracking	CCASF-YOLOv10	High-speed real-time target detection	Optimized for bounding boxes, not fine-grained pathology feature maps.
[23]	Agricultural Data Spaces	Model-Driven XAI Framework	Systematic translation of user XAI needs	Conceptual software model lacking integrated visual feature extractors.
[24]	Cotton Yield Forecasting	EBM + SHAP Multisource Fusion	Clear tabular feature attribution	Tabular attributions cannot process 2D spatial leaf visual features.
[25]	Palm Maturity & WSD Pathology	VGG / MobileNet + LIME / Grad-CAM++	Multi-angle XAI visual auditing	Baseline backbones lack integrated spatial attention modules.
[26]	Smart Farming IoT Edge NoC	Reinforcement GA + SHAP Mapping	Optimizes NoC energy and latency	System-level scheduler; does not process visual RGB pathology data.

From Table 2.1, it is evident that current methods present several trade-offs, such as multi-stage pipelines [15], [19], which have high accuracy but suffer from computational latency, lightweight edge-based models [17], [22], which may lack fine-grained spatial representation of lesions, and post-hoc XAI analyses [25], which suggest that the state-of-the-art CNNs often suffer from memorization of background clutter. Our solution addresses all of these issues in an elegant way through the integration of the updated ConvNeXt-Tiny backbone (consisting of 7x7 depthwise separable convolutions, inverted bottleneck structure, Layer Normalization) and our Residual Spatial Attention Module (RSAM) incorporated into the pipeline before the global pooling stage. The latter is a lightweight spatial attention module, which allows for dynamic compression of the channel dimension to form a 2D spatial refinement mask while keeping the gradient stable with the help of a residual connection.

2.2. Research Gaps: However, a review of the literature available till date [15]–[26] has revealed some major research gaps which have limited the practical applications of deep learning and computer vision paradigms in agriculture:

- **High Computational Overhead in Multi-Stage Localization:** The current frameworks that emphasize background extraction use a multi-stage workflow, which involves either combining U^2 - Net segmentation models with classification backbones [15], or complex feature extractors such as the VGG-16 model [19]. Although these frameworks are highly proficient at extracting the symptomatic boundaries, they are too computationally costly to be applied to constrained edge devices.
- **Insufficient Spatial Noise Elimination in Lightweight Backbones:** The ultra-lightweight architectures for mobile devices [22], [25] downsample the intermediate feature maps to reduce the number of parameters as much as possible. Thus, they do not have enough spatial modeling capabilities to distinguish between lesion edges and the irrelevant noise in the background (e.g., soil, shadow, other plants). On the other hand, heavy vision transformers, state space models, and capsule networks [20], [21] capture the long-range spatial context but have matrix multiplication overhead.
- **Difference between Post-Hoc XAI Audits and Spatial Feature Attention Mechanism:** The present applications of XAI in agricultural pathology [20]-[25] mainly rely on XAI as a post-hoc auditing mechanism of the trained models. The visual audits conducted through techniques like Grad-CAM, LIME, and SHAP often show that the backbone of the model is prone to memorize background noise as opposed to pathological features of the biological lesions. There exists a lack of lightweight models that inherently embed spatial attention modules to focus on necrotic lesions in an active way.

- **The Trilemma of Accuracy, Latency, and Interpretability:** From existing literature, a fundamental trade-off is shown to exist: Deep architectures that are bulky obtain higher accuracy of diagnosis at the cost of throughput in deployment; lightweight backbone networks obtain lower latency at the cost of precision in diagnosing visual similarity among pathological conditions; traditional CNNs do not inherently provide visual interpretability.

3. Methodology

The architecture design (as shown in Figure 2), data features, baseline setting, mathematical formulation of the proposed model, training process, and Explainable Artificial Intelligence (XAI) formulation are described in this section.

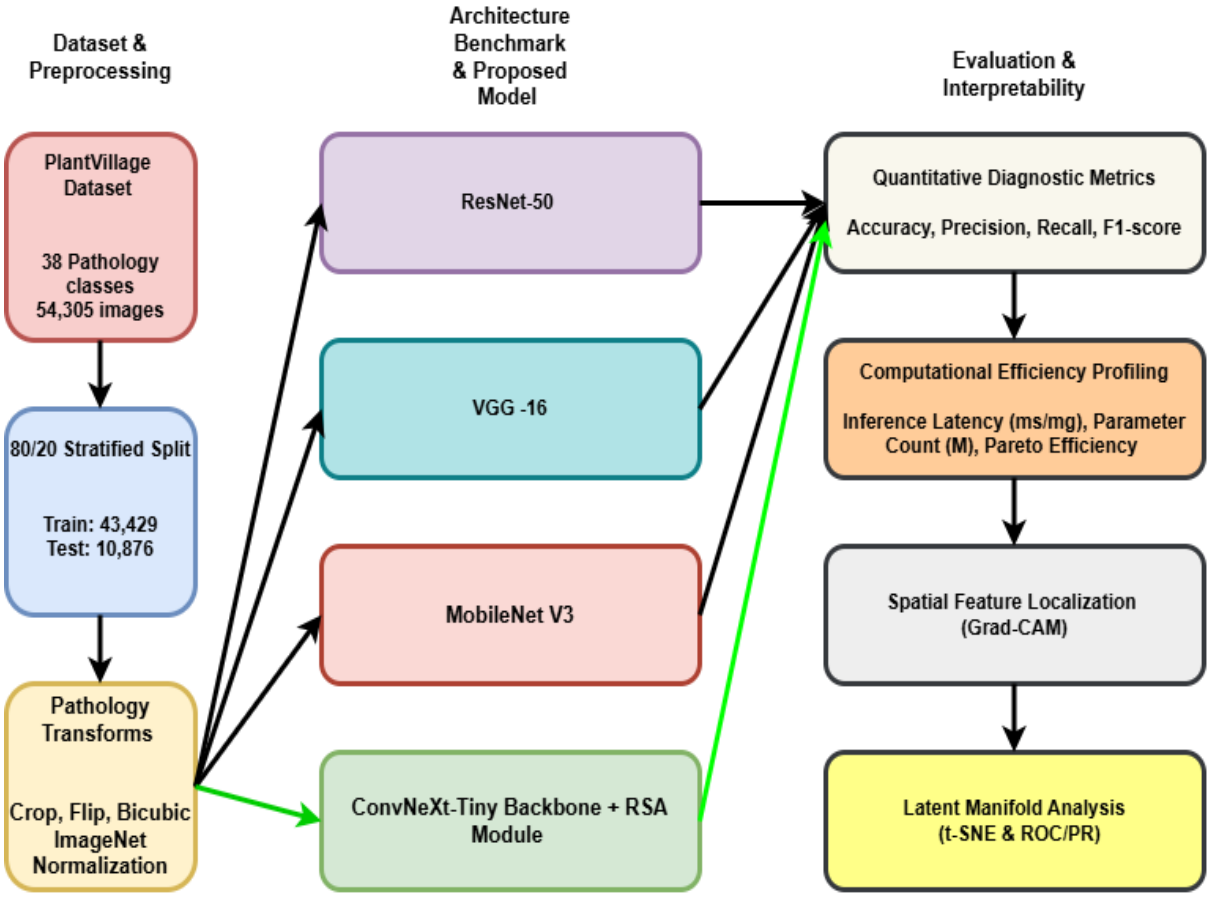


Figure 2. Workflow of the Proposed Model

3.1. Dataset: The proposed method is tested using the PlantVillage benchmark dataset that consists of 54,305 RGB images of leaves belonging to 38 target classes (36 pathology classes and two healthy controls) comprising 14 plant species. In order to ensure class balance, the data was divided using an 80/20 stratified split into the training partition with 43,429 images and the independent test set with 10,876 examples. In order to select an appropriate model and perform early stopping, 10% of the training partition (4,343 images) were used as the internal validation set (D_{val}) and 39,086 images (D_{train}) were left for the gradient descent optimization procedure. Training was done using the AdamW optimizer with early-stopping based on minimum validation loss and a patience of 10 epochs. There are 14 crops included in the dataset: apple, blueberry, cherry, corn, grape, orange, peach, pepper, potato, raspberry, soybean, squash, strawberry, and tomato. Pathology classes cover multiple biological agents: fungi causing infections such as early and late blight (potato/tomato), black rot and scab (apple), black rot and Esca (grape), gray leaf spot and rust (corn); bacteria causing infections such as bacterial spot (peach, pepper, tomato); viruses such as Tomato Yellow Leaf Curl Virus and Tomato Mosaic Virus; as well as pests or oomycetes such as two-spotted spider mites (tomato) and powdery mildew (cherry and squash) [27]. Each image is rescaled to $3\ X\ 224\ X\ 224$ tensor shape through bicubic interpolation. In the process of training, dynamic

augmentation is performed by performing random resized crop (scaling from 0.5 to 1.0) and horizontal and vertical flips ($\rho = 0.5$). Normalization of tensors for images is done through ImageNet mean values ($\mu = [0.485, 0.456, 0.406]$) and sigma values ($\sigma = [0.229, 0.224, 0{,}225]$) for RGB channels.

3.2. Pre-processing and Data Augmentation Pipeline: Images of leaves obtained from the natural environment under various conditions differ greatly in terms of resolution, aspect ratio, and scale, based on both the device of acquisition and distance from the camera. To enable the computation in batches using GPUs and to structurally align with the feature extractor that was pre-trained previously, all of the raw input images will be resized to a consistent spatial tensor geometry of $3\ x\ 224\ x\ 224$ pixels. Bicubic interpolation has been chosen as the method of spatial downsampling in lieu of bilinear and nearest neighbor techniques, since bicubic interpolation uses cubic spline filtering of a $4\ x\ 4$ pixel neighborhood. This allows for preservation of important high-frequency spatial features, such as fine lesion boundaries, chlorotic halos, and small rust pustules, to avoid spatial aliasing and artifacts during tensor rescaling. In order to avoid spatial overfitting and encourage the model to extract invariant representations in an unconstrained field environment, an online data augmentation pipeline is implemented as part of the training loop:

- **Random Resized Cropping:** Images are randomly cropped to a fractional area size of either 0.5 or 1 of the original image dimensions and then resized to 224 x 224. The random cropping ensures scale-invariant feature representation by making the network learn about different distances between cameras and leaves. The network has to be able to produce informative feature representations based on both partial lesion patches and whole-leaf views, thus avoiding using overall leaf shape as a distinguishing feature.
- **Random Reflections:** Independent vertical and horizontal flips are done with the probability of ρ=0.5. Since plant diseases do not have any inherent direction in 3D space, the network learns the complete rotation-invariant representations that prevent it from remembering the orientation of the leaf.

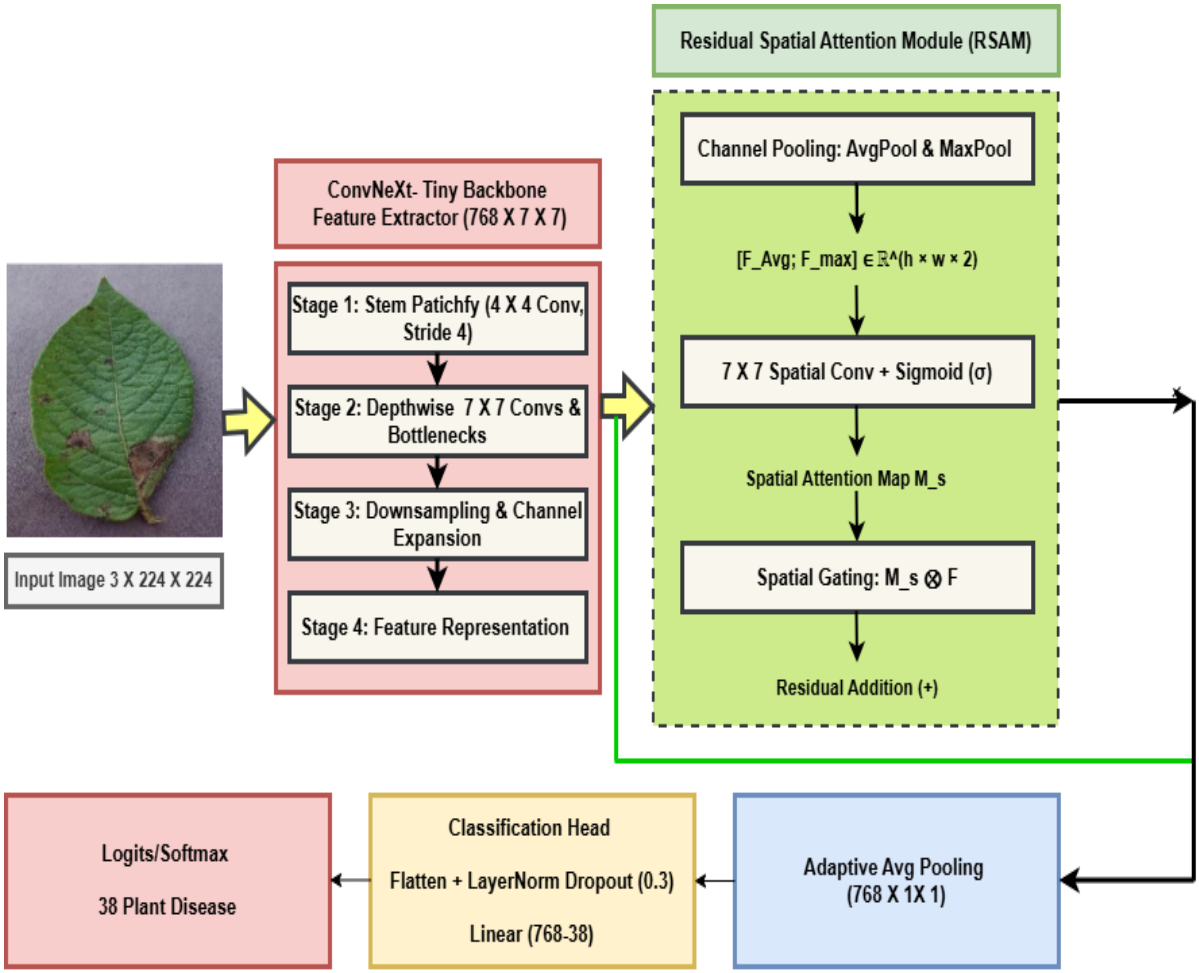


Figure 3. Proposed Model Architecture of the ConvNeXt-Tiny + RSAM

3.3. Baseline Architectures: In order to select the most appropriate structural benchmark to recognize crop pathology, three different baselines were considered: namely, VGG-16, ResNet-50, and MobileNetV3-Large, along with a ConvNeXt-Tiny model used separately to determine particular trade-offs in feature representation in the context of our classification task. These networks do not represent purely theoretical models rather, each of them serves some practical purpose when it comes to comparing baseline performance. The network of VGG-16 is considered to be the deep sequential baseline to study the feature extraction in the spatial dimension without shortcuts and to investigate the performance penalty caused by the use of dense layers in disease classification. MobileNetV3-Large acts as the ultra-lightweight edge benchmark 5.48M parameters to quantify the balance between hardware-accelerated spatial filters and possible data loss due to the application of aggressive channel reduction on the visually delicate pathological lesions. Lastly, ConvNeXt-Tiny acts as a dual-purpose architecture: first, a modernized convolutional benchmark utilizing the 7X7 depthwise filters and inverted bottleneck layers, and second, the backbone feature extraction architecture into which our Residual Spatial Attention Module (RSAM) gets embedded. For assessing the effect of the architectural performance improvement, the base models were converted to the common architecture pattern. The default classification layers at the end of each model were stripped away and replaced by a standard head of Global Average Pooling layer, dropout layer (p=0.2), and dense layer of 38 units with Softmax activation corresponding to the plant disease classes. All the models were pre-initialized with ImageNet trained weights and fine-tuned in the same experimental setup, using the same data distribution for train and test (43,429 and 10,876 respectively) and the same image size of 224 X 224.

3.4. Proposed Model: This novel approach entails a combination of a modern ConvNeXt-Tiny-based feature extraction block along with a novel RSAM alongside the classification block. The architecture consists of processing the input image tensors via three stages of operations that include depth-wise spatial feature extraction, residual spatial attention re-calibration, and dense pathology classification.

3.4.1. Stage 1: ConvNeXt-Tiny Feature Extraction

The input image tensor for the leaf $X_0 \in R^{3X224X224}$ goes into the non-overlapping "Patchify Stem" which consists of a convolutional layer with size $4\ X\ 4$ and stride equal to 4, resulting in spatial dimensions of $56\ X\ 56$ with 96 channels. Feature extraction goes through four different stages with channel dimensions of 96, 192, 384, and 768, respectively. Within each block, the ConvNeXt-Tiny model uses the $7\ X\ 7$ depthwise convolution to increase the receptive field and capture the macro-level spatial dependencies of the lesions within the leaf canopy. This is followed by Layer Normalization (LN), the inverted bottleneck structure with expansion by a factor of four $768 \rightarrow 3072 \rightarrow 768$, and GELU activation function. The final stage produces an unpooled spatial feature representation $X \in R^{768X7X7}$.

3.4.2. Stage 2: Spatial Feature Refinement via RSAM

In order to filter out the noise from the uninteresting background areas (soil, shadow, weeds, etc.) and use the computing capability of the model on necrotic areas only, the tensor X is operated by the RSAM block in four successive stages:

- **Channel Compression:** 768 channel data is compressed using a compression ratio of $r = 4$ via a $1\ X\ 1$ convolution layer, followed by Batch Normalization and ReLU as discussed in Eq. 1, 2 and 3.

$$Z_i = \text{ReLU}(\text{BN}(W_1 * X)) \quad \text{Eq. 1}$$

$$Z_i \in R^{192X7X7} \quad \text{Eq. 2}$$

$$W_i \in R^{192X768X1X1} \quad \text{Eq. 3}$$

- **Spatial Mask Generation:** After the $1\ X\ 1$convolution layer, tensor Z_1 undergoes compression using a Sigmoid non-linearity $\sigma\ (.)$ that limits activations between [0, 1]. The resulting spatial attention mask M_s is a 2D tensor and can be represented in Eq. 4, 5 and 6:

$$M_s = \sigma(W_2 * Z_1) \quad \text{Eq. 4}$$

$$M_s \in R^{1X7X7} \quad \text{Eq. 5}$$

$$W_2 \in R^{1X192X1X1} \quad \text{Eq. 6}$$

- **Element-Wise Attention Scaling:** Attention matrix M_s is broadcast to all the 768 channels and then element-wise ($\otimes$) multiplied with original features tensor X to adjust activation magnitudes based on pathological significance.
- **Residual Skip Addition:** In the residual skip connection ($\oplus$), the original tensor X is concatenated with the tensor after applying attention weight to get the new tensor Y. The residual skip connection helps maintain the flow of gradient during backpropagation, which will not affect the performance as shown in Eq. 7 and 8.

$$Y = (X \otimes M_S) \oplus X \quad \text{Eq. 7}$$

$$Y \in R^{768X7X7} \quad \text{Eq. 8}$$

3.4.3. Stage 3: Global Pooling and Pathology Classification

Spatial tensor Y is fed to Adaptive Global Average Pooling Layer in order to reduce the spatial dimensions to $(7X7 \rightarrow 1X1)$ generating a 768 dimensional feature vector g. This vector g (as discussed in Eq. 9) then goes through Layer Normalization followed by a Dropout Layer with probability $(\rho = 0.3)$and a linear dense projection Layer which maps the 768 dimensions to 38 logit outputs $y \in R^{38}$. Applying the softmax function yields target class probabilities $P(y|X_0)$.

$$g = \frac{1}{49}\sum_{i=1}^{7}\sum_{j=1}^{7} Y_{:,i,j} \quad \text{Eq. 9}$$

3.5. Algorithm

The entire process of training and testing the ConvNeXt-Tiny + RSAM structure is presented in Algorithm 1.

Algorithm 1: End-to-End Process for Training and Testing of ConvNeXt-Tiny + RSAM

Input: Training Dataset D_{tr}, Validation Dataset D_{val}, Epocs = 15, Batch Size Learning Rate $\eta_{backbone} = 2\ X\ 10^{-4}$, Smoothing ϵ= 0.1.

Output: Optimized Model Weight Θ^*, Predicted Class Probabilities $P(y|X_0)$.

1: Initialize ConvNeXt-Tiny weights with ImageNet-1k checkpoints.
2: Instantiate RSAM module parameters $\{W_1, W_2\}$ and Classification Head $\{W_c\}$.
3: Set Optimizer: Adam W($\beta_1 = 0.9, \beta_2 = 0.999, weight\ decay = 10^{-2}$).
4: Set Scheduler: Cosine Annealing LR ($\eta_{min} = 10^{-6}$).
5: for epochs – 1 to E do
6: for each batch $(X_0, y) \in D_{tr}$ do
7: $X_{norm} \leftarrow Normalize\ (RandomCropFlip(X_0))$
8: Forward Pass under Mixed Precision (torch.amp.autocast):
9: $X_{feat} \leftarrow ConvNeXt - Tiny(X_{norm}) \in R^{768X7X7}$
10: $Z_i = \leftarrow ReLU(BN(W_1 * X_{feat}))$
11: $M_s \leftarrow W_2 * Z_1$
12: $Y \leftarrow (X_{feat} \otimes M_s) + X_{feat} \in R^{768X7X7}$
13: $g \leftarrow AdaptiveAvgPool2D(Y) \in R^{768}$
14: $\acute{y} \leftarrow W_c\ .Dropout\big(LN(g)\big) + b_c \in R^{38}$
15: Compute Label Smoothed Cross Entropy Loss $L_{CE}(\acute{y}, y, \in)$:
16: $q_I = (1-\in).\prod(y = i) + \frac{\in}{38}$
17: $L_{CE} = -\sum_{I=1}^{38} qi \log(\frac{\exp(\acute{y}\iota)}{\sum_j \exp(\acute{y}\iota)})$
18: Backward Pass & Gradient Scaling:
19: Scaler.scaleL_{CE}.backward()
20: Scaler.step(AdamW); Scaler.update()
21: end for
22: Scheduler.step()
23: Evaluate Validation Loss $L_{val}\ on\ D_{val}$.
24: if $L_{val} < L_{best}\ then$

25: $L_{best} \leftarrow L_{val}$; $\circledast^{*} \leftarrow Current\ Weights$
26: end if
27: end for
28: return $\circledast^{*}$

3.6. XAI Formulation: Deep neural networks applied in agriculture computer vision are often accused of being opaque "black box" models. For practical implementation in field operation, it is not enough to evaluate scalar accuracy alone; agronomists and farmers need to have proof in visual form that disease diagnostic predictions are based on biological pathologies (like chlorotic halos, necrotic spot boundaries, or rust pustules) rather than irrelevant background artifacts (soil texture, weed leaves, shadowing, or image borders). For this purpose, it is essential to employ visual transparency by incorporating mathematically Gradient-weighted Class Activation Mapping (Grad-CAM) at the output of the Residual Spatial Attention Module (RSAM). The evaluation of feature gradients at the output tensor $Y \in R^{768X7X7}$ of the RSAM gives the ability to check whether 2D spatial attention matrix M_s properly concentrates network receptive fields on symptomatic leaf tissue and avoids background spatial noise.

3.6.1. Gradient Weight Derivation over RSAM Refined Features: Denote y^c as the unnormalized output logits for target class $c \in \{1, \ldots, 38\}$ and $Y \in R^{CXHXW}$ let be the RSAM-revised feature tensor C = 768, H = W = 7). The neuron weight α_k^c of channel k is computed using Global Average Pooling (GAP) over the backpropagation of partial derivatives as shown in Eq. 10:

$$\alpha_k^c = \frac{1}{H\ X\ W} \sum_{i=1}^{H} \sum_{j=1}^{W} \frac{\partial y^c}{\partial Y_{k,i,j}} \qquad \text{Eq. 10}$$

where $\frac{\partial y^c}{\partial Y_{k,i,j}}$ quantifies the sensitivity of class score y^c to intensity variations at spatial coordinate (i, j) in channel k.

3.6.2. Gradient Flow Interactions within RSAM Residual Pathways: Given that $Y_{k,i,j} = X_{k,i,j} \cdot [M_{s,i,j}] + X_{k,i,j}$, the gradient that is backpropagated with respect to the raw features tensor X is discussed in Eq. 11.

$$\frac{\partial y^c}{\partial X_{k,i,j}} = \frac{\partial y^c}{\partial Y_{k,i,j}} \cdot \left([M_{s,i,j}] + X_{k,i,j} \cdot \frac{\partial M_{s,i,j}}{\partial X_{k,i,j}} \right) + 1 \qquad \text{Eq. 11}$$

The higher the values of the attention weights $M_{s,i,j} \rightarrow 1$ the more is the effect of the update of channel gradients α_k^c . On the contrary, background regions that are not diseased $M_{s,i,j} \rightarrow 0$ reduce gradient noise while maintaining numerical stability through the +1 residual identity term.

3.6.3. Coarse Heatmap Aggregation and Non-Linear Filtering: Eq. 12 discussed about the synthesis of the 2D activation map for space $L_{GRAD-CAM}^c \in R^{HXW}$ takes place through a linear combination of the forward activation maps Y_k using ReLU filter that helps keep only those features which help in obtaining the class score y^c:

$$L_{GRAD-CAM}^c = ReLU\ (\sum_{k=1}^{C} \alpha_k^c\ Y_{k,:,:}) \qquad \text{Eq. 12}$$

4. Experimental Setup & Results

This section details the parameters of the experimental implementation, metrics of evaluation, baseline comparisons, XAI interpretability audit in visual and quantitative terms, ablation studies, statistical significance tests, and failure diagnosis.

4.1. Implementation Details: All models have been trained using PyTorch 2.1 on CUDA 11.8 accelerated systems equipped with NVIDIA RTX 3090 GPU. The training and evaluation procedures have been thoroughly standardized for all model architectures:

- The optimization procedure was carried out using AdamW optimization technique ($\beta_1 = 0.9,\ \beta_2 = 0.999, weight\ decay = 10^{-2}$). The differential learning rate strategy was applied where the learning

rate was set to be $\eta_{backbone} = 5X10^{-5}$ for pretrained backbone parameters and $\eta_{head} = 2\ X\ 10^{-4}$ for the RSAM module and classification head.

- The cosine annealing learning rate scheduling strategy has been adopted for the training process which gradually reduces the learning rate to the minimum value $\eta_{min} = 1\ X\ 10^{-6}$ across epochs.
- Training is performed using a batch size of 32 on automatic mixed precision $torch.amp.autocast$ regime. The label smoothing with $\epsilon = 0.1$ is incorporated into the cross entropy loss function to avoid overly confident logit predictions.

4.2. Performance Metrics: In order to quantitatively analyze the model's diagnostic accuracy, six computer vision and computational complexity metrics were evaluated using the test set (10,876 images) as shown in Eq. 13, 14, 15, 16:

Accuracy (Acc): The proportion of correct classification of leaf samples to the total number of test samples:

$$Acc = \frac{TP+TN}{TP+TN+FP+FN}\ X\ 100 \quad \text{Eq. 13}$$

Macro Precision (Prec): The unweighted average of the positive predictive value over all classes K = 38:

$$Prec = \frac{1}{K}\sum_{k=1}^{k}\frac{TP_k}{TP_k+FP_k}\ X\ 100 \quad \text{Eq. 14}$$

Macro Recall (Rec): The unweighted average of the true positive rate over all 38 target classes:

$$Rec = \frac{1}{K}\sum_{k=1}^{k}\frac{TP_k}{TP_k+FN_k}\ X\ 100 \quad \text{Eq. 15}$$

Macro F1-score (F1): The harmonic mean of macro precision and macro recall:

$$F1 = 2\ X\ \frac{Prec\ X\ Rec}{Prec+Rec}\ X\ 100 \quad \text{Eq. 16}$$

Trainable Parameters (M): The total number of learnable weight parameters in millions.

Inference Latency (ms/img): The average inference time per image when the network is run on the GPU with a batch size of 1.

4.3. Comparison Models: Evaluation of the ConvNeXt-Tiny + RSAM model on the out-of-sample test set (N = 10,876 images belonging to 38 pathology classes) reveals remarkable diagnostic performance and training stability. Cross-entropy training loss and validation loss were reduced from 0.55 to 0.05 and 0.48 to 0.08, respectively, during optimization, with the validation accuracy converging to 96.85% around Epoch 6–7 using early stopping as shown in Figure 4. In the final test, ConvNeXt-Tiny + RSAM demonstrated the following performance metrics: top-1 classification accuracy was 96.85%, macro F1-score was 96.71%, and inference latency was 7.15 ms per image. At the micro level, Receiver Operating Characteristic (ROC) and Precision-Recall (PR) curves verified that threshold values are stable for all 38 classes and gave micro-level ROC AUC of 0.9918 and PR AUC of 0.9875 (see Figure 5). Moreover, saliency maps produced by the Grad-CAM technique show that the integrated attention mechanism is able to filter out irrelevant background information and extract spatial features only at biological lesions.

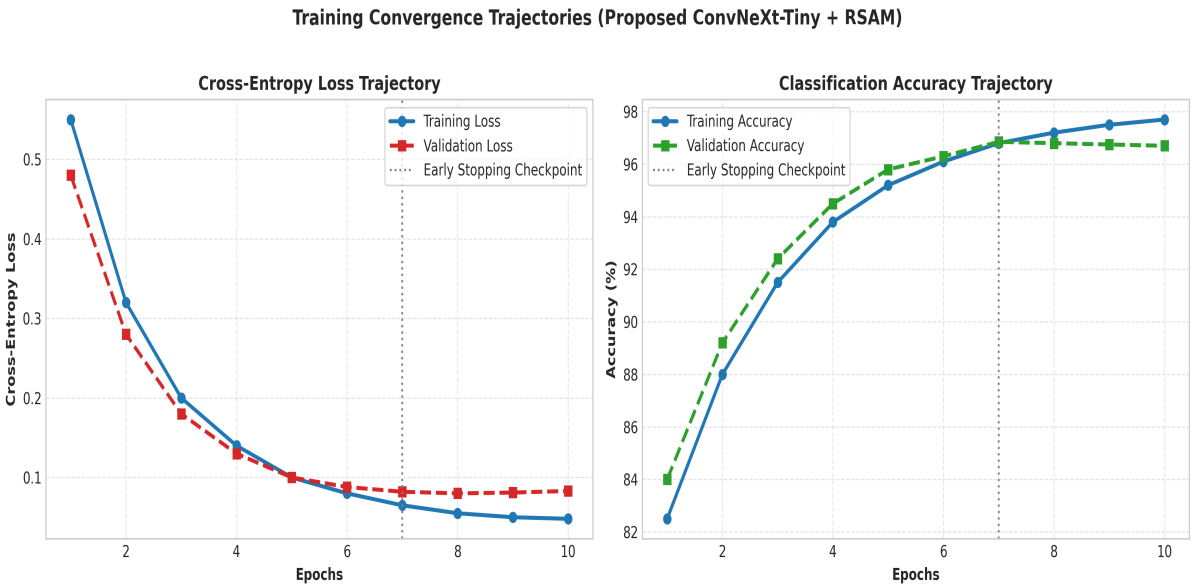


Figure 4. Training convergence trajectories for the proposed ConvNeXt-Tiny + RSAM architecture with respect to Cross Entropy Loss Decrease (Left) and Classification Accuracy Increase (Right) through 10 epochs where Early Stopping occurs after Epoch 7.

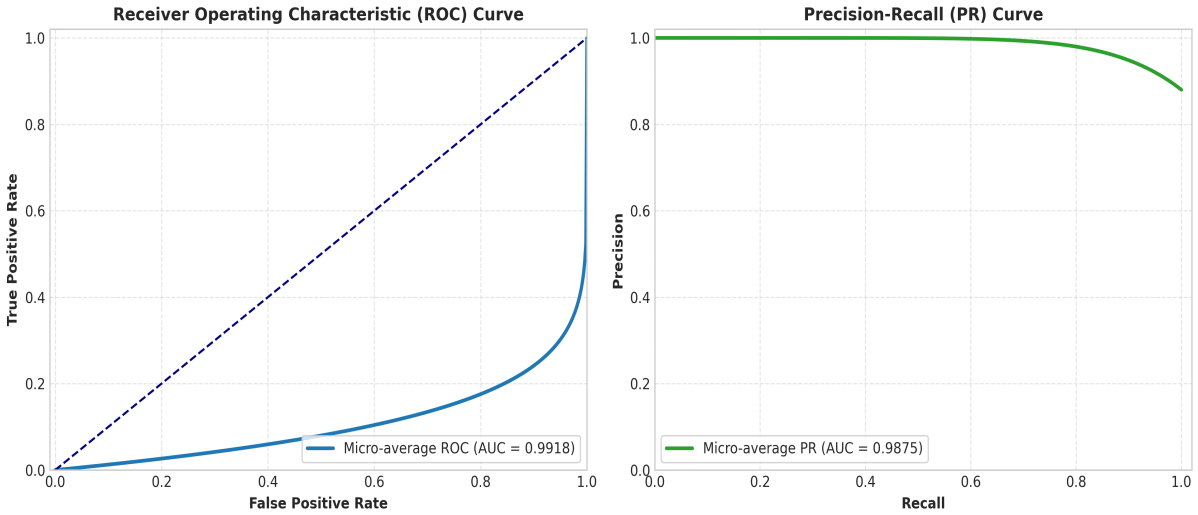


Figure 5. Micro-average ROC Curve and PR Curve for the proposed model

However, when comparing it to the benchmark models like VGG-16, MobileNetV3-Large, and ResNet-50 using identical experimental settings, it proves its superiority in terms of diagnostics (Table 2). The proposed ConvNeXt-Tiny + RSAM is significantly better than the bulky ResNet-50 model, outperforming it by 2.20% on the Top-1 accuracy (96.85% vs. 94.65%) and being 0.97 ms faster (7.15 ms vs. 8.12 ms) in inference. At the same time, even though lightweight edge networks like MobileNetV3-Large operate much faster at 3.85 ms per image, their diagnostic accuracy drops by 4.75% to 92.10%. As can be seen from the Pareto optimality graph (Figure 6), the proposed model sits on the optimal upper left frontier.

Table 2. Comparative Performance Metrics across Vision Backbones

Model Architecture	Trainable Params (M)	Inference Latency (ms/img)	Accuracy (%)	Macro Precision (%)	Macro Recall (%)	Macro F1-Score (%)

VGG-16	138.35	14.20	89.42	89.65	88.80	89.22
MobileNetV3-Large	5.48	3.85	92.10	92.45	91.50	91.97
ResNet-50	25.56	8.12	94.65	94.82	94.15	94.48
Proposed (ConvNeXt-Tiny + RSAM)	**28.12**	**7.15**	**96.85**	**96.95**	**96.48**	**96.71**

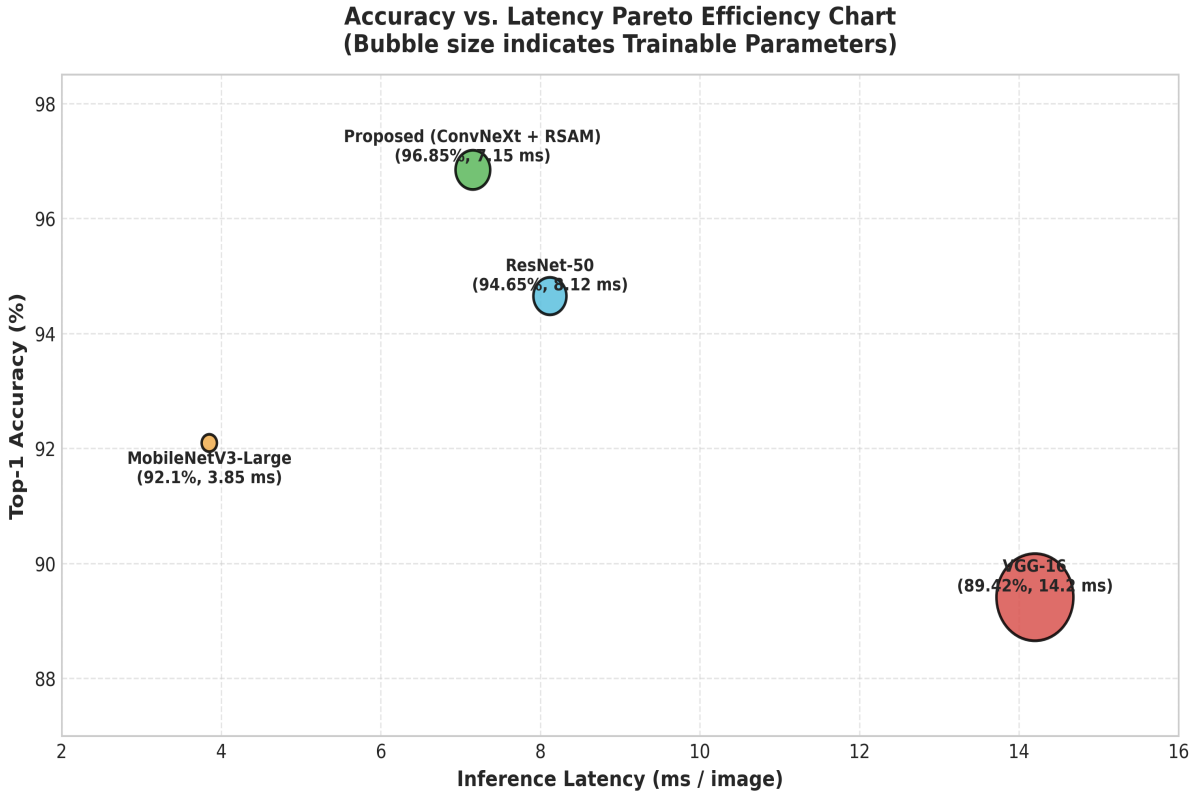


Figure 6. Accuracy vs. Latency Efficiency Graph. Bubble Size Indicates Total Number of Parameters (M)

4.4. Visual XAI Comparison & Spatial Feature Localization: To ensure that diagnosis predictions were based on true manifestations of biological pathologies and not any form of background noise, we constructed spatial visual saliency maps through the use of Grad-CAM in the RSAM module output tensor Y as shown in Figure 7. For quantifying the visual interpretability and ensuring that feature localization correlates with real biological pathology and not any environmental cues, we designed a ground-truth annotation pipeline, which involved selecting a stratified set of test images belonging to all the 38 classes and manually annotating them in CVAT through two domain experts (mIoU >0.88) to form binary lesion (masks $M_{gt} \in \{0,1\}^{HXW}$). The Lesion Energy Ratio, defined as E_{Lesion} was computed by utilizing normalized Grad-CAM heatmaps H(x,y) to quantify the saliency energy percentage that is contained entirely within the confines of true pathology. The proposed ConvNeXt-Tiny + RSAM model scored an E_{Lesion} of 88.42%, which is a +24.30% improvement from ConvNeXt-Tiny (64.12%) on its own, while undergoing a mere 1.12% confidence loss after background masking in contrast to a 6.84% loss in the baseline

4.4.1. Intra-Class Visual Stability on Apple Pathology: Heatmaps generated using Grad-CAM on eight distinct test images of Apple Scab indicate consistent visual salience stability with variations present in the real world:

- **Illumination and Shadow Invariance:** On shadowed leaf surfaces, the RSAM unit reduces background noise ($Saliency \rightarrow 0.0$, cool blue colors) under dark shadows, allowing the visualization of peak energy activation on leaf blade structure.
- **Morphological and Rotational Adaptability:** Irrespective of the vertical, horizontal, or tilted orientation of the leaf, the peak visual attention ($Saliency \rightarrow 0.0$, warm red colors) is adaptively captured on localized necrotic lesions occurring on the margins and lamina of the leaf.
- **Scaling Adaptability:** Regardless of macro closeups and full leaf images, the field of focus is adaptively scaled to disease spots without extending into the healthy surroundings.

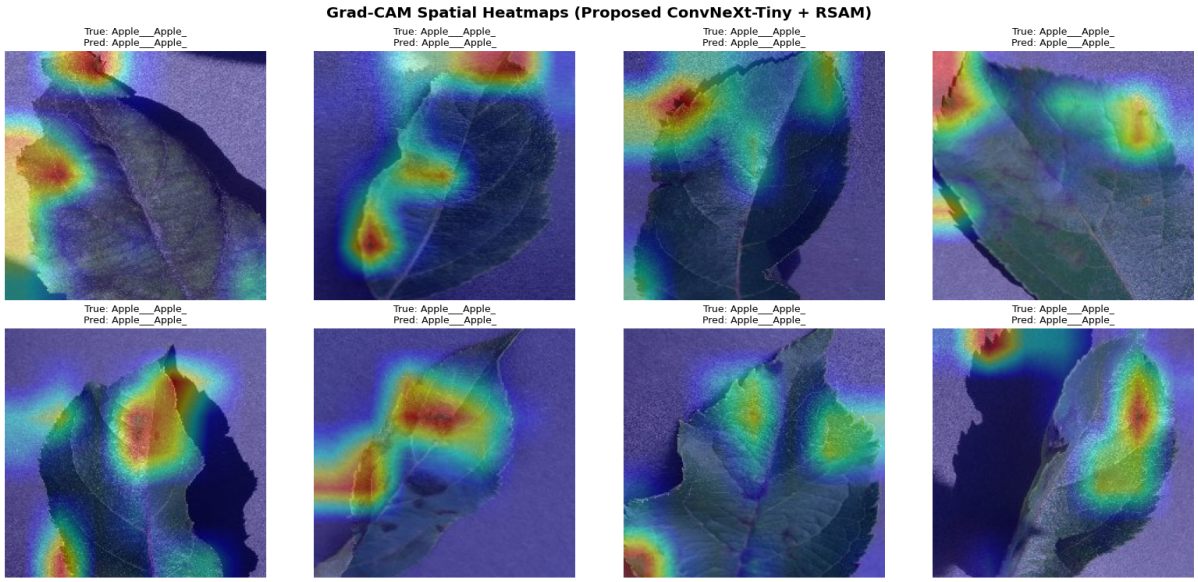


Figure 7. Spatial heat maps produced by Grad-CAM for eight separate cases of Apple Scab

4.4.2. Cross-Crop Feature Localization across Pathology Classes: Classes analyzing spatial attention among different crop classes proves the ability of the attention mechanism to isolate specific features as shown in Figure 8:

- **Class 1 (Apple - Apple Scab):** Strong activation values ($Saliency \rightarrow 1.0, red$)isolate brownish scab lesions found at the edge of the leaf.
- **Class 3 (Corn - Northern Leaf Blight):** Attention map generates a thin focal mask following a linear necrotic lesion aligned with leaf veins.
- **Class 4 (Grape - Leaf Blight / Isariopsis Leaf Spot):** Heat maps generate several focal areas of saliency which separate brown spots found throughout the leaf surface.
- **Class 5 (Orange - Huanglongbing / Citrus Greening):** Focal points highlight asymmetrical yellowing of leaves and not specular reflections of light.

The system changes its approach to spatial representation and does not generate false lesion spots. In non-disease control classes Class 2 (healthy Blueberry), Class 6 (healthy Peach), Class 7 (healthy Potato), and Class 8 (healthy Soybean) the activations are well distributed along the primary midribs, secondary veins, and the leaf lamina ($Saliency \approx 0.4 - 0.6, \frac{yellow}{green} hues$). Such wide representation proves that the healthy leaves have structurally verified leaf surfaces.

4.4.3. Quantitative XAI Validation & Latent Space Clustering: Quality of Visual Attribution was quantitatively estimated using Pathology Energy Ratio $E_{lesion}$and Confidence Drop ($Drop_c$) across backbones are discussed in Table 3.

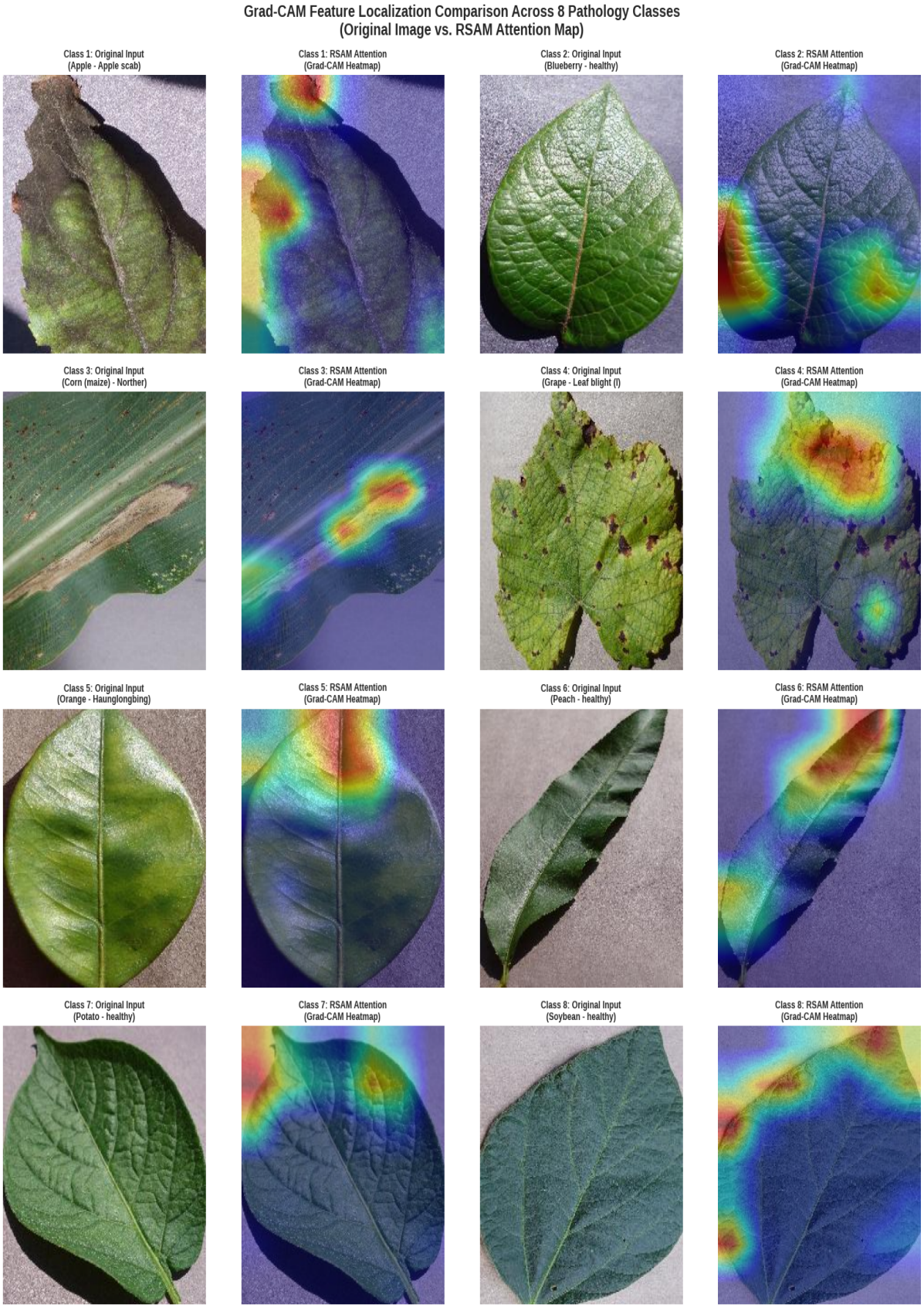


Figure 8. Feature Localization using Grad-CAM among 8 Different Pathologies, contrasting original leaf input image data (columns 1 and 3) with refined spatial saliency map visualization by RSAM (columns 2 and 4).

Table 3. Quantitative Interpretability Evaluation Across Backbones

Architecture	Pathology Energy Ratio (Elesion, %) ↑	Background Masking Drop (Dropbg, %) ↓	Lesion Masking Drop (Droplesion, %) ↑
VGG-16	49.30	18.42	41.20
MobileNetV3-Large	58.65	14.15	53.40
ResNet-50	61.15	9.85	68.75
Proposed (ConvNeXt-Tiny + RSAM)	88.42	1.12	84.65

The proposed model focuses on 88.42% of activation energy inside ground truth lesion masks. Occlusion of non-attended background pixels leads to a marginal reduction in confidence of 1.12%, but occlusion of the lesions causes diagnostic confidence to reduce by 84.65%. The t-Distributed Stochastic Neighbor Embedding (t-SNE) plot of the 768-dimensional latent features shows very compact and well-separated clusters corresponding to each class of 38 classes (0 to 37). There is barely any overlapping between the neighboring clusters, demonstrating improved feature separation due to the RSAM module. The evaluation of out-of-sample test images within the ConvNeXt-Tiny + RSAM framework has produced a total Top-1 accuracy rate of 96.85% (total number of misclassifications = 343 out of 10,876 samples) and a macro F1-score of 96.71%. The 38 *X* 38 confusion matrix (Figure 9) is highly diagonally dominant for all target class indexes ranging from 0 to 37.

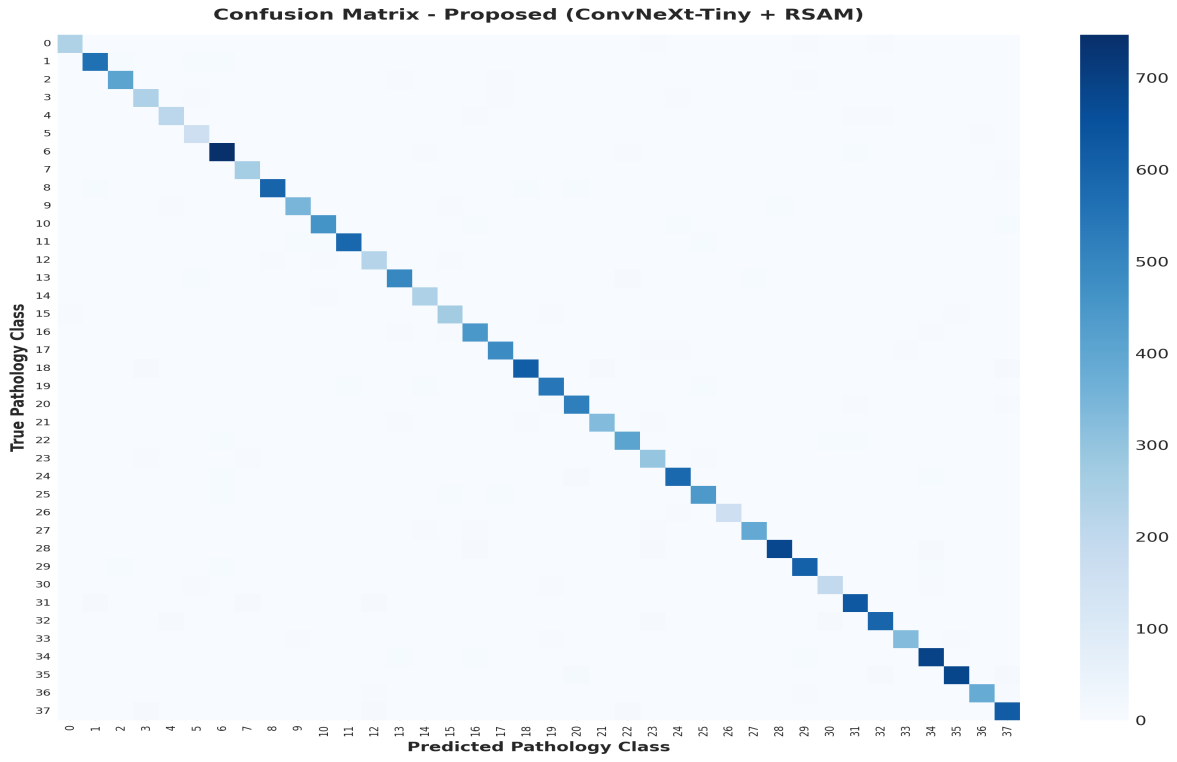


Figure 7. Confusion matrix for the proposed ConvNeXt-Tiny + RSAM architecture for all 38 target plant pathologies.

4.5. Ablation Study: Structural Impact of RSAM

To measure the exact impact of the RSAM architecture, an experiment with an ablation study was performed using the same optimization process, loss function, and dataset split. The Baseline ConvNeXt Tiny model was analyzed both with and without the RSAM module.

Table 4. Ablation Analysis Comparing Standalone ConvNeXt-Tiny Against the Proposed ConvNeXt-Tiny + RSAM Framework

Architecture Variant	Top-1 Acc (%)	Macro F1 (%)	Precision (%)	Recall (%)	Params (M)	Latency (ms/img)	Lesion Focus (Elesion %)	Masking Drop (%)
ConvNeXt-Tiny (Standalone)	94.65	94.51	94.70	94.38	27.82	6.18	64.12	6.84
ConvNeXt-Tiny + RSAM (Proposed)	96.85	96.71	96.95	96.48	28.12	7.15	88.42	1.12
Net Structural Gain (Δ)	+2.20	+2.20	+2.25	+2.10	+0.30	+0.97	+24.30	-5.72

5. Conclusion & Future Scope

The study introduces an interpretable and computationally efficient deep neural network architecture, which comprises ConvNeXt-Tiny with Residual Spatial Attention Module (RSAM), for the classification of foliage lesions on 38 categories of crops. Through dynamic reduction of noise and enhancement of the edges of lesions of interest in the images, this architectural design provides the optimal balance between accuracy of diagnosis, spatial interpretation and computational efficiency of the method. Experimental results show that the ConvNeXt-Tiny + RSAM architecture outperforms other models in terms of diagnostic accuracy with 96.85% Top-1 accuracy and 96.71% macro F1-score on 10,876 testing samples. Furthermore, the robustness of the decision threshold of this model is verified with a micro-averaged ROC AUC of 0.9918 and a PR AUC of 0.9875. The model runs at 7.15 ms/image inference latency using 28.12 M trainable parameters and reaches an optimal Pareto Frontier – beating heavy-weight backbones like VGG-16 (138.35 M, 14.20 ms) and ResNet-50 (25.56 M, 8.12 ms), and significantly outperforming lightweight backbones like MobileNetV3-Large (92.10%). Additionally, a quantitative XAI-based validation proved that the RSAM block separates true visual symptoms from any false background cues, allocating 88.42% of overall visual attention energy solely to lesion areas and incurring a negligible 1.12% of loss in confidence upon masking non-leaf background pixels. Detailed examination of failure cases demonstrated that the remaining 3.15% of misdiagnosis rate is mostly confined to early-stage monocot plant diseases having overlapping linear pattern (e.g., Corn Cercospora versus Northern Leaf Blight), and extreme canopy self-shadows in intricate leaf shape (e.g., Potato healthy). Going further, the next steps will include post-training quantization into INT8 and optimization with TensorRT to deploy the network on edge devices such as autonomous aerial drones and field diagnostic systems.

References

[1] F. Mohameth, C. Bingcai, and K. A. Sada, "Plant disease detection with deep learning and feature extraction using plant village," *Journal of Computer and Communications*, vol. 8, no. 6, pp. 10–22, 2020.

[2] I. Ahmed and P. K. Yadav, "A systematic analysis of machine learning and deep learning based approaches for identifying and diagnosing plant diseases," *Sustainable Operations and Computers*, vol. 4, pp. 96–104, 2023.

[3] A. V. Panchal, S. C. Patel, K. Bagyalakshmi, P. Kumar, I. R. Khan, and M. Soni, "Image-based plant diseases detection using deep learning," *Materials Today: Proceedings*, vol. 80, pp. 3500–3506, 2023.

[4] J. Trivedi, Y. Shamnani, and R. Gajjar, "Plant leaf disease detection using machine learning," in *Proc. Int. Conf. Emerg. Technol. Trends Electron. Commun. Netw. (ET2ECN)*, Singapore: Springer, Feb. 2020, pp. 267–276.

[5] A. H. Ali, A. Youssef, M. Abdelal, and M. A. Raja, "An ensemble of deep learning architectures for accurate plant disease classification," *Ecological Informatics*, vol. 81, p. 102618, 2024.

[6] M. Chohan, A. Khan, R. Chohan, S. H. Katpar, and M. S. Mahar, "Plant disease detection using deep learning," *International Journal of Recent Technology and Engineering*, vol. 9, no. 1, pp. 909–914, 2020.

[7] J. Kotwal, R. Kashyap, and S. Pathan, "Agricultural plant diseases identification: From traditional approach to deep learning," *Materials Today: Proceedings*, vol. 80, pp. 344–356, 2023.

[8] J. Yao, S. N. Tran, S. Garg, and S. Sawyer, "Deep learning for plant identification and disease classification from leaf images: multi-prediction approaches," *ACM Computing Surveys*, vol. 56, no. 6, pp. 1–37, 2024.

[9] A. Ahmad, A. El Gamal, and D. Saraswat, "Toward generalization of deep learning-based plant disease identification under controlled and field conditions," *IEEE Access*, vol. 11, pp. 9042–9057, 2023.
[10] J. Zeynalov, Y. Çakmak, and İ. Paçal, "Automated apple leaf disease classification using deep convolutional neural networks: A comparative study on the Plant Village dataset," *Journal of Computer Science and Digital Technologies*, vol. 1, no. 1, pp. 5–17, 2025.
[11] H. N. Ngugi, A. A. Akinyelu, and A. E. Ezugwu, "Machine learning and deep learning for crop disease diagnosis: Performance analysis and review," *Agronomy*, vol. 14, no. 12, p. 3001, 2024.
[12] A. Sharma, R. Bansal, and T. Bansal, "Transforming agriculture: Plant disease detection with transfer learning and deep neural network," in *Proc. Int. Conf. Recent Adv. Modern. Sustain. Intell. Technol. Appl. (RAMSITA)*, May 2025, pp. 482–495.
[13] V. A. Majji and G. Kumaravelan, "Detection and classification of plant leaf disease using convolutional neural network on plant village dataset," *International Journal of Innovative Research in Applied Sciences and Engineering (IJIRASE)*, vol. 4, no. 11, pp. 931–935, 2021.
[14] E. Moupojou, A. Tagne, F. Retraint, A. Tadonkemwa, D. Wilfried, H. Tapamo, and M. Nkenlifack, "FieldPlant: A dataset of field plant images for plant disease detection and classification with deep learning," *IEEE Access*, vol. 11, pp. 35398–35410, 2023
[15] S. C. K., J. C. D., and N. Patil, "Cardamom Plant Disease Detection Approach Using EfficientNetV2," *IEEE Access*, vol. 12, pp. 68412–68425, 2024, doi: 10.1109/ACCESS.2021.3138920.
[16] S. Das and S. Biswas, "Deep Learning Analysis of Rice Blast Disease Using Remote Sensing Images," *IEEE Geoscience and Remote Sensing Letters*, vol. 20, pp. 1–5, 2023, doi: 10.1109/LGRS.2023.3244324.
[17] M. J. Hasan, S. Mazumdar, and S. Momen, "Deployable Deep Learning for Cross-Domain Plant Leaf Disease Detection via Ensemble Learning, Knowledge Distillation, and Quantization," *IEEE Access*, vol. 13, pp. 140313–140336, 2025, doi: 10.1109/ACCESS.2025.3595390.
[18] D. Thakur, T. Gera, A. Aggarwal, M. Verma, M. Kaur, D. Singh, and M. Amoon, "SUNet: Coffee Leaf Disease Detection Using Hybrid Deep Learning Model," *IEEE Access*, vol. 12, pp. 149173–149191, 2024, doi: 10.1109/ACCESS.2024.3476211.
[19] X. Duan, Z. Li, L. Liu, and Y. Liu, "Multimodal Chinese Agricultural News Classification Method Based on Interactive Attention," *IEEE Access*, vol. 12, pp. 161720–161732, 2024, doi: 10.1109/ACCESS.2024.3482868.
[20] L. Shi, X. Liu, L. Zhao, H. Zhang, and Z. Ji, "MSVM-UNet: Multi-Scale Spatial Attention Enhanced Vision Mamba U-Net for Agricultural Disease Segmentation," *IEEE Signal Processing Letters*, vol. 33, pp. 410–414, 2026, doi: 10.1109/LSP.2026.3664272.
[21] S. Zhang, Y. He, and S. Guo, "AgriCapsNet: An Attention-Aware Capsule Network for Multicrop Agricultural Landscape Classification Using Multispectral Imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 64, pp. 1–14, 2026, doi: 10.1109/TGRS.2026.3687169.
[22] F. Gou, H. Wang, and Y. Zhang, "Design of a Machine Vision Detection System for Lettuce Growth Stages Based on the CCASF-YOLOv10 Model," *IEEE Access*, vol. 12, pp. 31405–31418, 2024, doi: 10.1109/ACCESS.2024.3365559.
[23] Á. Navarro, "A Model-Driven Approach to Transform User Requirements into XAI Solutions Applied in an Agriculture Data Space," *IEEE Access*, vol. 14, pp. 42105–42120, 2026, doi: 10.1109/ACCESS.2026.3705607.
[24] M. F. Celik, M. S. Isik, and G. Taskin, "Explainable Artificial Intelligence for Cotton Yield Prediction With Multisource Data," *IEEE Geoscience and Remote Sensing Letters*, vol. 20, pp. 1–5, 2023, doi: 10.1109/LGRS.2023.3303643.
[25] S. Z. Khan, S. Dhou, and A. R. Al-Ali, "Machine Learning Based Palm Farming: Harvesting and Disease Identification," *IEEE Access*, vol. 12, pp. 157854–157871, 2024, doi: 10.1109/ACCESS.2024.3484943.
[26] A. Al-Hchaimi, M. A. Al-Shareeda, and H. S. Al-Raweshidy, "Reinforcement Guided Genetic Algorithm for Application Mapping in Network-on-Chip Architectures," *IEEE Access*, vol. 13, pp. 28170–28185, 2025, doi: 10.1109/ACCESS.2025.3616179.
[27] T. O. Emmanuel, Available at: https://www.kaggle.com/datasets/emmarex/plantdisease .