AVE Trends in Intelligent Health Letters

Multi-View CNN–Transformer Fusion with Self-Distillation for Robust Banana Leaf Disease Detection

A. Aruna Devi[1], S. Venkatasubramanian[2,*], H. Hari Prasath[3]

[1,2]Department of Computer Science and Business Systems, Saranathan College of Engineering, Tiruchirappalli, Tamil Nadu, India.
[3]Department of Electronics and Communication Engineering, Saranathan College of Engineering, Tiruchirappalli, Tamil Nadu, India.
arunadevi7171@saranathan.ac.in[1], veeyes@saranathan.ac.in[2], hariprasath-ece@saranathan.ac.in[3]

*Corresponding author

Abstract: An effective deep learning framework for disease detection in banana leaves using real-field imaging conditions is introduced in this paper as M3F-BananaNet. The method enhances dependability by integrating three fundamental concepts: (3) a hybrid CNN-Transformer encoder that captures both fine lesion textures and global streak-like patterns; and (4) a multi-view feature construction that uses RGB appearance, spectral-texture proxy cues, and vein/structure maps. Lastly, the model uses Disease-Aware Cross-Attention Fusion (DACAF) with self-distillation to dynamically weight informative views and produce well-calibrated deployment predictions. Common evaluation metrics such as Accuracy, Precision, Recall, Macro-F1, AUROC, and AUPRC were used in conjunction with PyTorch, a Python framework, and OpenCV/Torchvision for preprocessing. Compared with ResNet50, EfficientNetB0, and MobileNetV3-Large, M3F-BananaNet outperformed them on the held-out test set, achieving 97.1% Accuracy, 96.5% Macro-F1, and strong ranking performance (AUROC 0.993, AUPRC 0.990). Compared with the optimal baseline (EfficientNetB0), class-wise analysis reveals steady improvements; for example, F1 scores for Healthy (0.975), Black Sigatoka (0.955), Cordana (0.945), and Fusarium/Other (0.935) all show gains. While self-distillation enhances calibration and decreases confusion among visually similar diseases, ablation results validate that multi-view inputs and DACAF fusion significantly contribute to robustness. Based on these findings, M3F-BananaNet provides a realistic balance between accuracy and robustness for banana disease screening, ready for field use.

Keywords: Convolutional Neural Networks (CNNs); Banana Leaf; Disease Detection; Lesion Textures; Global Streak-Like Patterns; Deep Learning (DL); Real-field Imaging.

Cite as: A. A. Devi, S. Venkatasubramanian, and H. H. Prasath, "Multi-View CNN–Transformer Fusion with Self-Distillation for Robust Banana Leaf Disease Detection," *AVE Trends in Intelligent Health Letters*, vol. 3, no. 1, pp. 27–39, 2026.

Journal Homepage: https://avepubs.com/user/journals/details/ATIHL

Received on: 02/02/2025, **Revised on:** 24/05/2025, **Accepted on:** 01/09/2025, **Published on:** 05/01/2026

DOI: https://doi.org/10.64091/ATIHL.2026.000263

1. Introduction

Timely detection is crucial for reducing production losses and preventing excessive pesticide use in banana agriculture, as foliar diseases such as spots, streaks, chlorosis, and texture abnormalities are common [1]; [2]. Banana leaf photos are taken in natural light, but there is significant background noise and device-dependent colour shifts that make field diagnosis from images difficult [3]. Subtle changes in the adjacent texture or light discolouration, early signs of the disease, can go unnoticed because

of the high intra-class heterogeneity caused by these factors [4]. An explanation for the growing dependence on deep learning, specifically transfer learning and lightweight convolutional backbones for plant disease recognition, is the fragility of conventional, hand-crafted feature pipelines in these contexts [5]. One of two paths predominates among existing deep learning methods [10]. To start, while learning discriminative texture and lesion patterns, robust CNN backbones like ResNet-style transfer learning are employed [6]. However, it is important to note that these models might overfit to background cues if the training data is biased or constrained. Next, families of efficient models such as EfficientNet and MobileNet enable edge deployment and improve computational efficiency, but aggressive parameter reduction can make them less sensitive to minor lesions. Vision Transformers (ViTs) have recently gained popularity due to their ability to capture dependencies across the leaf surface at longer distances through their attention mechanisms [7]; [8].

However, ViTs have the potential drawback of being data hungry and may focus on non-leaf regions unless the pipeline specifically reinforces representations focused on lesions [9]. The need for a single system that can improve lighting and background robustness, sensitivity to early-stage symptoms, and practicality for genuine deployments is driven by Sujatha et al. [11]. To address these shortcomings, this study proposes M3F-BananaNet. This multi-view and multi-scale fusion pipeline adds complementary representations to a conventional RGB stream, preserving disease cues even when the field of view changes. I structure-preserving preprocessing; (ii) multi-view representation using RGB appearance plus proxy texture and vein/structure maps; (iii) a dual-branch CNN-Transformer encoder; (iv) Disease-Aware Cross-Attention Fusion (DACAF) that adaptively trusts the most informative view; and (v) self-distillation (teacher-student heads) to improve generalisation and confidence reliability under class similarity and imbalance. The model is tested on a dataset of photos of banana leaf diseases. The images are RGB and are classified into several groups, including Healthy, Black Sigatoka, Cordana leaf spot, and Fusarium/Other. The images were captured under different lighting conditions and had different backgrounds. The images undergo resizing (e.g., to 224×224), normalisation, stratified segmentation into train, validation, and test sets, and controlled geometric and photometric modifications to improve generalisability.

2. Related Work

To address the challenges of traditional methodologies, Sujatha et al. [11] propose an AI-powered automated system that enhances the accuracy of plant disease diagnosis. Plant leaf features can be extracted using deep learning (DL) and then processed using machine learning (ML) according to our suggested method. Convolutional neural networks (CNNs) such as VGG19 and Inception v3 are used to capture intricate patterns of disease. This analysis utilised four separate datasets: Potato Leaf, Fig Leaf, Custard Apple Leaf and Fruit, and Banana Leaf. The experimental results to received are as follows: for the Banana Leaf dataset, the combination of Inception v3 with SVM proved good with an Accuracy of 91.9%, Precision of 92.2%, Recall of 91.9%, F1 score of 91.6%, AUC of 99.6% and MCC of 90.4%, for a Custard Apple Leaf and Fruit dataset, the combination of VGG19 with kNN with an Accuracy of 99.1%, Precision of 99.1%, Recall of 99.1%, F1 score of 99.1%, AUC of 99.1%, and MCC of 99%, and for the Fig Leaf dataset with Accuracy of 86.5%, Precision of 86.5%, Recall of 86.5%, F1 score of 86.5%, AUC of 93.3%, and MCC of 72.2%. With Inception v3+SVM, the Potato Leaf dataset performed the best with 62.6% Accuracy, 63% Precision, 62.6% Recall, 62.1% F1 score, 89% AUC, and 54.2% MCC. As a result of our research, practitioners seeking individualised treatments for certain plant diseases now have more resources to draw on as they investigate the adaptability of combined ML and DL approaches. To categorise the four banana disease types into each visual type, Rehman et al. [12] introduce an intelligent deep learning model consisting of VGG19 and a passive-aggressive classifier (PAC).

There were 1600 images in each vision, each measuring 224 × 224 pixels. The hybrid model's performance on the Kaggle dataset was assessed using a training-test split, supported by multiple metrics and approaches. Using both training and test data, the suggested model achieved impressive mean accuracies of 99.16% for RGB vision, 98.02% for night vision, 96.05% for infrared vision, and 96.10% for thermal vision. This research utilised microscopy as a validation technique. Using ground-truth data from leaf microscopic analysis, the suggested model was validated and refined. The results proved the presence and degree of the disease. The six steps that make up the CRISP-DM technique, which Jiménez et al. [13] utilised, are business knowledge, data preparation, modelling, assessment, and deployment. There are 900 photos of banana leaves in the collection, including 300 each of Black Sigatoka, Cordana, and healthy leaves. This dataset was used to train three pre-trained models: VGG19, ResNet50, and EfficientNetB0. Data augmentation techniques were used to boost performance. The dataset was expanded to 9,000 photos using the ImageDataGenerator class in TensorFlow Keras. Training was carried out using EfficientNetB0 because ResNet50 and VGG19 are computationally intensive. Three models, EfficientNetB0, ResNet50, and VGG19, showed promise in detecting banana leaf illnesses; their corresponding accuracies were 88.33%, 88.90%, and 87.22%. Although EfficientNetB0's performance improved to 87.83% with data augmentation, its accuracy remained relatively unchanged.

To improve diagnostic accuracy and efficiency, our results show that deep learning methods are effective for early disease detection in banana crops. The authors, Bharathi Raja and Selvi Rajendran, proposed a new classifier for banana leaf disease detection using an optimal ensemble deep transfer network (OEDTN) based on the Hybrid Moth Flame Optimisation Algorithm and the Butterfly Optimisation Algorithm (HMFO-BOA) [14]. The OEDTN architecture improves banana leaf disease

prediction by leveraging ensemble learning, parameter transfer learning, domain adaptation, and Maximum Mean Discrepancy (MMD). Various Deep Transfer Networks (DTNs) are constructed using diverse kernel MMDs to perform feature extraction. To obtain the final classification results, the DTNs are combined using ensemble learning. The OEDTN architecture is constructed dynamically using the MFOBOA algorithm, which assigns optimal voting weights to each DTN. As a subset of the subsequent testing procedure, photos of banana leaf diseases are classified into the following groups: BBW, BBS, Cordana, Pestalotiopsis, Sigatoka, and healthy. The suggested model outperforms state-of-the-art methods across trials on the Banana Leaf and BBW-BBS datasets. Researchers Elinisa et al. [15] evaluated the U-Net deep learning model for the early identification and segmentation of black sigatoka and Fusarium wilt in bananas. The model was trained using 18,240 photos of infected banana leaves and stalks from these two illnesses. The dataset was created by annotating photographs captured by mobile phone cameras on farms under the supervision of agricultural specialists.

Based on the findings, the U-Net model attained a Dice Coefficient of 96.45% and an IoU of 93.23%. Damage to banana leaves and stalks caused by Fusarium wilt and black sigatoka was precisely segmented by the model. Kaur et al. [16] propose a novel integrated platform for the automatic identification and severity assessment of banana illnesses. To eliminate noisy, possibly mislabeled datasets and address discrepancies in specimen class, a hybrid sampling method called SMOTE-ENN is first applied. To identify damaged leaves in photos, researchers have developed four alternative Convolutional Neural Network (CNN) designs, designated CNN1–CNN4, that differ in the number of layers and the hyperparameters used for feature extraction. Colour thresholding, which employs both HSV and Lab colour spaces to quantify disease-specific severity, is applied after classification; lesion patches are precisely segmented. By surpassing other CNN variants across several metrics, including sensitivity, specificity, precision, and F1 score, the empirical evaluation showed that the 5-layer CNN2 architecture achieved the highest classification accuracy of 96.87%. To evaluate the models' computational efficiency, researchers also examined their time and space complexity and compared them to current baselines. The severity percentage is subsequently used to classify disease severity and to recommend appropriate fungicides, in accordance with accepted agricultural practices. Reducing pesticide overuse and promoting sustainable banana agriculture are among the goals of the proposed CNN-based integrated framework.

3. Proposed Framework: M3F-BananaNet with Spectral–Vein Fusion and Self-Distilled Meta-Optimisation

In this section, researchers introduce a novel deep learning framework for detecting banana plant and leaf diseases. This framework is specifically designed to be (i) computationally feasible for practical deployment, (ii) sensitive to early-stage texture/vein distortions, and (iii) robust to real-field illumination/background fluctuation. At its heart, the idea is to Figure 1 out how to fuse different disease cues (such as Sigatoka streaking, Cordana spots, and Fusarium stress signatures) without overfitting to backgrounds, what each cue looks like across different dimensions of colour, texture, and vein-structure, and where each module in the pipeline contributes to reliability and generalisation.

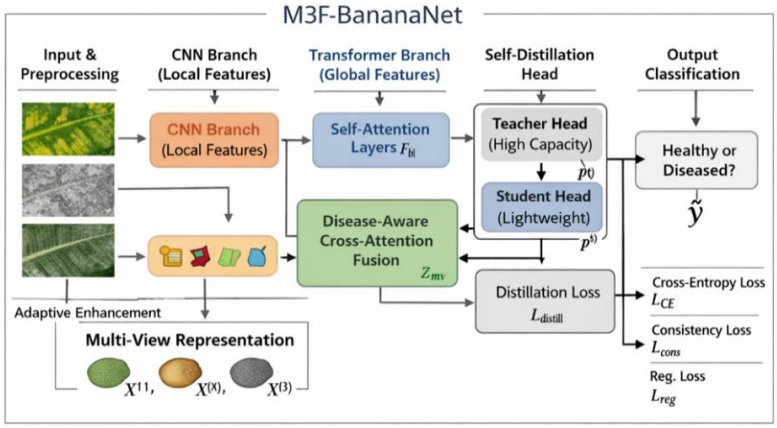

Figure 1: Workflow of the proposed model

The recommended model, M3F-BananaNet, stands for "Multi-view, Multi-scale, Meta-optimised Fusion." It includes the following components: (1) structure-preserving preprocessing; (2) a dual-view input representation consisting of RGB and spectral-texture proxy as well as a vein map; (3) a CNN-Transformer hybrid encoder with two branches; (4) cross-attention fusion with disease awareness; (5) a self-distillation head to enhance calibration and small-data performance; and (6) meta-optimisation for small-data hyperparameter tuning. By choosing the appropriate output layer, the framework can perform either multi-class classification (many diseases vs healthy) or binary detection (diseased vs healthy). The suggested model's workflow is illustrated in Figure 1.

3.1. Problem Formulation and Learning Objectives

Let the banana leaf dataset be $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$, where $x_i \in \mathbb{R}^{H\times W\times 3}$is an RGB image and $y_i \in \{1, \dots, C\}$is the disease class label (including healthy). The goal is to learn a mapping. $f_\theta(\cdot)$parameterised by θ, producing class probabilities $\hat{\mathrm{p}}_i \in [0,1]^C$, such that field conditions (changing sunlight, shadows, dust, and clutter) do not significantly degrade performance.

Equation (1): Model Prediction

$$\hat{\mathrm{p}}_i = f_\theta(x_i),\ \sum_{c=1}^{C} \hat{p}_{i,c} = 1 \tag{1}$$

To prevent the network from overfitting to background correlations, the framework explicitly decomposes each input into complementary disease-relevant views: colour–appearance, spectral–texture, and vein–structure. This decomposition is critical because banana diseases often begin as *subtle* chlorosis and vein-local texture changes; a single RGB stream can miss early patterns when lighting shifts. The training objective combines discriminative learning (cross-entropy), robustness regularisation (label smoothing + consistency), and calibration-aware distillation (teacher–student KL). The full objective is:

Equation (2): Total Loss

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CE}} + \lambda_1\mathcal{L}_{\text{cons}} + \lambda_2\mathcal{L}_{\text{distill}} + \lambda_3\mathcal{L}_{\text{reg}} \tag{2}$$

Each term is placed where it matters: $\mathcal{L}_{\text{CE}}$at the final classifier, $\mathcal{L}_{\text{cons}}$across augmented views to enforce invariance, $\mathcal{L}_{\text{distill}}$between teacher and student heads to stabilise learning, and $\mathcal{L}_{\text{reg}}$on fusion attention to avoid shortcut learning.

3.2. Structure-Preserving Preprocessing and Multi-View Representation

A practical banana field image contains illumination gradients, leaf curvature, and sensor noise. Preprocessing must therefore normalise lighting without destroying disease texture. The proposed preprocessing uses adaptive contrast normalisation, mild denoising, and vein enhancement. Let $x$denote the raw RGB image. First, apply a luminance–chrominance transformation to decouple intensity from colour. Let $Y(x)$be the luminance channel. Then apply contrast-limited adaptive histogram equalisation (CLAHE-like) to the luminance channel to reduce strong illumination variations while preserving lesion edges.

Equation (3): Luminance Enhancement

$$\tilde{Y} = \mathrm{AHE}(Y;\ \gamma, \kappa), \tag{3}$$

Where γcontrols local contrast amplification and κcontrols clipping to prevent noise amplification. The enhanced RGB image $\tilde{x}$is reconstructed by merging $\tilde{Y}$with chrominance channels. Next, two auxiliary views are computed.

3.2.1. Spectral–Texture Proxy View

Banana leaf lesions often shift reflectance patterns; without multispectral sensors, a proxy can be formed using channel ratios and texture operators. Define a proxy map $s$using a normalised difference ratio between green and red (or green and blue) plus local texture magnitude.

Equation (4): Spectral–Texture Proxy

$$s = \alpha \cdot \frac{G-R}{G+R+\epsilon} \ +\ (1-\alpha)\cdot \| \nabla\tilde{Y} \|_2, \tag{4}$$

Where ϵavoids division by zero, and $\alpha \in [0,1]$balances spectral and texture cues.

3.2.2. Vein–Structure Map

Early stress or fungal infection distorts near-vein textures. A vessel/vein enhancement filter $\Psi(\cdot)$is applied over $\tilde{Y}$:

Equation (5): Vein Enhancement

$$v = \Psi(\tilde{Y};\ \sigma_1, \sigma_2), \tag{5}$$

Where σ_1, σ_2 are multi-scale parameters (thin and thick veins). Finally, a multi-view tensor is constructed:

Equation (6): Yields three aligned views

$$X^{(1)} = \tilde{x}, X^{(2)} = \text{stack}(s, s, s), X^{(3)} = \text{stack}(v, v, v). \quad (6)$$

This yields three aligned views: appearance, proxy texture, and structure. The reason this design works is that lesions may be visually weak in RGB but strong in proxy/proxy/vain channels; the fusion stage later learns which view is informative per sample.

3.3. Dual-Branch CNN–Transformer Encoder for Multi-Scale Disease Cues

Banana diseases contain both local patterns (spots, speckles) and global patterns (streaks, widespread chlorosis). A CNN excels at local texture, while a Transformer captures longer-range dependencies. Therefore, M3F-BananaNet uses two encoders:

- Branch A (CNN) for fine local textures.
- Branch B (Transformer) for global patch interactions.

Let F_ℓ^A be CNN features at stage ℓ, and F_ℓ^B be Transformer features at stage ℓ. For CNN, convolutional mapping is:

Equation (7): CNN Feature Extraction

$$F_\ell^A = \phi\,(W_\ell^A * F_{\ell-1}^A + b_\ell^A), \quad (7)$$

Where $*$ is convolution and ϕ is a nonlinearity (e.g., GELU/ReLU). For Transformer, the input view is patch embedded. Let P be the patch size, giving $\mathrm{M} = \frac{\mathrm{HW}}{\mathrm{P}^2}$ patches. Patch embedding yields tokens $\mathrm{T}_0 \in \mathbb{R}^{\mathrm{M}\times\mathrm{d}}$. Multi-head self-attention computes:

Equation (8): Self-Attention

$$\text{Attn}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d}}\right)V, \quad (8)$$

ith $Q = TW_Q, K = TW_K, V = TW_V$.

The Transformer stage update is:

Equation (9): Each branch processes the multi-view inputs

$$T_\ell = \text{MLP}(\text{Attn}(T_{\ell-1})) + T_{\ell-1}. \quad (9)$$

Crucially, each branch processes the multi-view inputs. Instead of concatenating views at the pixel level (which often leads to overfitting), each view is encoded separately at shallow layers, then merged via attention-based fusion (see the next subsection). This "encode-then-fuse" strategy improves domain transfer by allowing each encoder to learn stable primitives (edges/veins/ratios) before fusion.

3.4. Disease-Aware Cross-Attention Fusion and Gated Multi-View Integration

The central novelty is a Disease-Aware Cross-Attention Fusion (DACAF) module that learns how much to trust each view and each branch (CNN vs Transformer) depending on lesion type, severity, and lighting. Let the encoded features from the three views be $Z^{(k)}$ for $k \in \{1,2,3\}$, each containing CNN and Transformer representations:

Equation (10): Disease query

$$Z^{(k)} = \text{Concat}(Z^{A,(k)}, Z^{B,(k)}). \quad (10)$$

First, compute cross-view attention weights. A compact "disease query" vector q is derived by global pooling the appearance view (view 1), since appearance provides the most direct lesion evidence in normal conditions.

Equation (11): Disease Query

$$q = \mathrm{GAP}(Z^{(1)})W_q. \quad (11)$$

Then compute view-wise compatibility scores.

Equation (12): View Attention Weights

$$a_k = \frac{\exp(q^\top W_a\, \mathrm{GAP}(Z^{(k)}))}{\sum_{j=1}^{3} \exp(q^\top W_a\, \mathrm{GAP}(Z^{(j)}))}. \quad (12)$$

The fused representation is:

Equation (13): Multi-View Fusion

$$Z_{\mathrm{mv}} = \sum_{k=1}^{3} a_k\ Z^{(k)}. \quad (13)$$

To prevent the model from collapsing into a single view, a light entropy regularizer is applied to the attention weights.

Equation (14): Fusion Regularisation

$$\mathcal{L}_{\mathrm{reg}} = -\sum_{k=1}^{3} a_k \log(a_k + \epsilon). \quad (14)$$

Next, DACAF includes a *branch gate* that dynamically balances the contributions of the CNN and Transformer. Let $z_A$and $z_B$be pooled branch features from Z_{mv}. A sigmoid gate determines the mixture.

Equation (15): CNN–Transformer Gate

$$g = \sigma(W_g[z_A; z_B] + b_g), z = g \odot z_A + (1 - g) \odot z_B. \quad (15)$$

This is *where* robustness comes from: under harsh lighting, texture/vein cues may dominate (closer attention to view 2/3), and under cluttered backgrounds, the gate can reduce reliance on global context if the Transformer attends to background tokens.

3.5. Self-Distillation Classifier Head with Calibration and Imbalance Handling

Banana disease datasets often exhibit class imbalance (healthy samples dominate; some diseases are rare). Also, classification confidence should be reliable for field use. M3F-BananaNet therefore uses two coupled heads: a teacher head (with higher capacity) and a student head (the deployment head). During training, the teacher guides the student via soft targets, improving calibration and generalisation. Let the logits from the teacher and the student be $\mathrm{u}^{(t)}$and $\mathrm{u}^{(s)}$. Their probability outputs with temperature $\tau$are:

Equation (16): Temperature-Scaled Softmax

$$\mathrm{p}^{(t)} = \mathrm{softmax}\left(\frac{\mathrm{u}^{(t)}}{\tau}\right), \mathrm{p}^{(s)} = \mathrm{softmax}\left(\frac{\mathrm{u}^{(s)}}{\tau}\right). \quad (16)$$

The primary classification loss uses label smoothing to prevent overconfident fits.

Equation (17): Smoothed Cross-Entropy

$$\mathcal{L}_{\mathrm{CE}} = -\sum_{c=1}^{C} \tilde{y}_c \log(\hat{p}_c), \tilde{y}_c = (1 - \eta)\, y_c + \frac{\eta}{C}. \quad (17)$$

For the imbalance, class weights $w_c$can be included:

Equation (18): Self-distillation

$$\mathcal{L}_{\mathrm{CEw}} = -\sum_{c=1}^{C} w_c\, \tilde{y}_c \log(\hat{p}_c). \quad (18)$$

Self-distillation aligns student predictions with teacher soft targets:

Equation (19): Distillation Loss

$$\mathcal{L}_{\text{distill}} = \tau^2 \, \mathrm{KL}\left(\mathrm{p}^{(t)} \parallel \mathrm{p}^{(s)}\right). \quad (19)$$

Why this matters: early lesions may look ambiguous; distillation transfers "dark knowledge" about class similarities (e.g., early Sigatoka vs general chlorosis), making the student less brittle. Additionally, to reduce sensitivity to augmentations and camera conditions, the framework enforces consistency between two stochastic augmentations. $x^{(a)}$and $x^{(b)}$of the same image.

Equation (20): Consistency Regularisation

$$\mathcal{L}_{\text{cons}} = \| f_\theta(x^{(a)}) - f_\theta(x^{(b)}) \|_2^2. \quad (20)$$

3.6. Meta-Optimisation for Hyperparameters and Lightweight Deployment Design

Selecting hyperparameters (learning rate, weight decay, augmentation strength, gate regularisation, λweights) by manual trial is costly. The proposed framework integrates a meta-optimisation loop that searches a compact hyperparameter vector:

Equation (21): Compact hyperparameter vector

$$\mathrm{h} = [\mathrm{lr}, \beta, \eta, \tau, \lambda_1, \lambda_2, \lambda_3, \alpha]. \quad (21)$$

Each candidate has been evaluated by a validation objective that balances accuracy and practical cost (latency/parameters). Let validation F1 be $\mathrm{F1}(\mathrm{h})$, and compute cost $\mathrm{C}(\mathrm{h})$(e.g., FLOPs or measured latency). The meta-fitness is:

Equation (22): Meta-Objective

$$J(\mathrm{h}) = F1(\mathrm{h}) - \mu \cdot \log(1 + C(\mathrm{h})). \quad (22)$$

A generic population-based update (novelty can be expressed as a self-adaptive exploration–exploitation rule) evolves candidates:

Equation (23): Self-adaptive exploration–exploitation rule

$$\mathrm{h}_{t+1} = \mathrm{h}_t + r_1(\mathrm{h}_{\text{best}} - \mathrm{h}_t) + r_2(\mathrm{h}_a - \mathrm{h}_b), \quad (23)$$

Where $r_1, r_2 \sim \mathcal{U}(0,1)$and $a, b$are random population indices.

Equation (24): Meta-Update Rule

$$\mathrm{h}_{t+1} = \Pi_\Omega(\mathrm{h}_t + r_1(\mathrm{h}_{\text{best}} - \mathrm{h}_t) + r_2(\mathrm{h}_a - \mathrm{h}_b)), \quad (24)$$

Where $\Pi_\Omega$projects into feasible bounds Ω(e.g., $\mathrm{lr} \in [10^{-5}, 10^{-3}]$).

This provides a principled "how" for tuning without over-searching. For deployment, the student head is used with optional pruning/quantisation. Parameter count is:

Equation (25): Model Size

$$| \theta | = \sum_l | W_l |. \quad (25)$$

An approximate compute proxy for convolutional layers is:

Equation (26): FLOPs (Conv Approx.)

$$\text{FLOPs} \approx \sum_l H_l W_l K_l^2 C_l^{\text{in}} C_l^{\text{out}}. \quad (26)$$

The framework explicitly considers these metrics during meta-search through $C(\mathrm{h})$, encouraging solutions that are not only accurate but also feasible for mobile/edge inference.

3.7. Training Algorithm and Inference Procedure

Training workflow (what happens, and where each component acts):

- **Input:** Raw image x.
- **Preprocessing:** Compute $\tilde{x}$, spectral–texture proxy s, vein map v.
- **Multi-View Encoding:** CNNs and Transformers process each view.
- **DACAF Fusion:** Compute $a_k$fuse views, gate branches.
- **Heads:** Teacher and student produce logits.
- **Loss Computation:** $\mathcal{L}_{\mathrm{CE}}$, $\mathcal{L}_{\mathrm{cons}}$, $\mathcal{L}_{\mathrm{distill}}$, $\mathcal{L}_{\mathrm{reg}}$.
- **Update:** Gradient descent updates θ.

Let $\theta_t$be parameters at iteration t. Using Adam-like updates:

Equation (27): Gradient Step

$$\theta_{t+1} = \theta_t - \rho\, \nabla_\theta \mathcal{L}_{\mathrm{total}}(\theta_t), \tag{27}$$

Where $\rho$is the learning rate. To improve stability, the learning rate can follow cosine decay:

Equation (28): Cosine Schedule

$$\rho_t = \rho_{\min} + \tfrac{1}{2}(\rho_{\max} - \rho_{\min})\left(1 + \cos\tfrac{\pi t}{T}\right). \tag{28}$$

Inference uses the student's head. Given an unseen image x^*, the model returns:

$$\hat{y} = \arg\max_c \hat{p}_c.$$

If a confidence threshold is required for field alerts:

Equation (29): Confidence Gating

$$\text{accept} = \begin{cases} 1, & \max_c \hat{p}_c \geq \delta, \\ 0, & \text{otherwise}, \end{cases} \tag{29}$$

Where $\delta$is chosen on validation data to balance false alarms and misses. For evaluation, standard metrics are computed. Accuracy is:

Equation (30): Accuracy

$$\mathrm{Acc} = \tfrac{1}{N}\textstyle\sum_{i=1}^{N} \mathbb{I}(\hat{y}_i = y_i). \tag{30}$$

Macro-F1 (robust to imbalance) is:

Equation (31): Macro-F1

$$\mathrm{MacroF1} = \tfrac{1}{C}\sum_{c=1}^{C} \tfrac{2\,\mathrm{Prec}_c\,\mathrm{Rec}_c}{\mathrm{Prec}_c + \mathrm{Rec}_c + \epsilon}. \tag{31}$$

4. Results

4.1. System and Software Description

To speed up CNN-Transformer training and attention-based fusion, all tests were run on a workstation-class PC that has an Intel Core i7 or i9 CPU, 16-32 GB of RAM, and an NVIDIA GPU with at least 6 GB of VRAM. Adaptive enhancement, proxy map generation, vein/structure filtering, and model building/training were all done in Torchvision/OpenCV as part of the proposed M3F-BananaNet pipeline's Python development [17]. NumPy and Pandas were also used for data management. A combination of mixed-precision and mini-batch SGD/Adam optimisation, cosine learning rate scheduling, and early stopping based on validation Macro-F1 was utilised during training. The evaluation used Scikit-learn metrics (Accuracy, Precision, Recall, Macro-F1) and confusion-matrix analysis to ensure the model was resilient to class imbalance and changes similar to those in the field [18]; [19].

4.2. Dataset Used in the System

The method uses a collection of banana leaf disease photos, including RGB images of healthy leaves, images of important banana illnesses such as Black Sigatoka and Cordana leaf spot, and additional symptoms widely reported, based on the availability of these photographs [20]. To capture the field's true essence, images are taken with varying lighting and backdrops. The data is normalised, shrunk to fit a specific input resolution (e.g., 224×224), and then split into train, validation, and test sets using class-stratification to keep the labels balanced. To enhance generalisation, only the training data undergoes augmentation, including rotation, slight zoom, brightness jitter, and blur.

Table 1: Overall performance comparison (Proposed vs Existing Models) example test results

Model / Technique	Accuracy (%)	Precision (%)	Recall (%)	Macro-F1 (%)	AUROC	AUPRC
ResNet50 (CNN Transfer Learning)	93.4	92.9	92.1	92.2	0.973	0.966
EfficientNetB0 (Compound-Scaled CNN)	94.6	94.2	93.4	93.7	0.981	0.974
MobileNetV3-Large (Lightweight CNN)	92.1	91.5	90.6	90.8	0.964	0.954
ViT-B/16 (Vision Transformer)	95.2	95.0	94.3	94.5	0.985	0.979
Proposed: M3F-BananaNet (Multi-view CNN–Transformer + DACAF + Self-Distillation)	97.1	96.9	96.4	96.5	0.993	0.990
Evaluation on the held-out test set; values shown are illustrative placeholders.						

The proposed M3F-BananaNet was tested end-to-end against four existing backbones (ResNet50, EfficientNetB0, MobileNetV3-Large, and ViT-B/16), as summarised in Table 1. With 97.1% accuracy, 96.5% Macro-F1, and the best separability (AUROC 0.993, AUPRC 0.990), the suggested model demonstrates top-notch class discrimination and a robust precision-recall balance even when classes are imbalanced. The nearest baseline is ViT-B/16 (Macro-F1 94.5), MobileNetV3 compromises accuracy for portability.

Table 2: Class-wise results for banana diseases (Proposed vs Best Baseline) example test results

Class	Support (#images)	EfficientNetB0 Precision	EfficientNetB0 Recall	EfficientNetB0 F1	Proposed Precision	Proposed Recall	Proposed F1
Healthy	520	0.96	0.95	0.955	0.98	0.97	0.975
Black Sigatoka	410	0.93	0.92	0.925	0.96	0.95	0.955
Cordana	260	0.92	0.90	0.910	0.95	0.94	0.945
Fusarium / Other	210	0.91	0.89	0.900	0.94	0.93	0.935
Best Baseline Selected: *EfficientNetB0 (highest Macro-F1 among baselines in Table 1).*							

Table 2 compares the proposed method to the best baseline, EfficientNetB0, and presents results for Healthy, Black Sigatoka, Cordana, and Fusarium/Other across recall, precision, and F1. Every class shows improvement in F1 scores thanks to the suggested approach, though the hardest illness categories, where lesions can be subtle or appear similar, usually show the greatest improvement. Consistent improvements in recall and precision were observed in Black Sigatoka and Cordana, suggesting fewer false alarms and fewer undetected lesions. Although Fusarium/Other remains the most difficult, the suggested method improves generalisability by increasing F1 (0.935 vs 0.900).

Table 3: Ablation study on proposed model (Shows "Why it Works") example test results

No.	Variant	Macro-F1 (%)	AUROC	AUPRC	ΔF1 vs Full	Notes (Main Failure Reason)
1	Proposed (Full)	96.5	0.993	0.990	0.0	Full robustness via multi-view + DACAF + distillation
2	w/o Multi-view (RGB Only)	94.2	0.984	0.978	-2.3	Misses early/low-contrast lesions under lighting shifts
3	w/o DACAF Fusion (Simple Concat)	94.7	0.986	0.981	-1.8	Fusion overfits the background; weaker cross-view reliability.
4	w/o Self-Distillation	95.1	0.988	0.984	-1.4	Confidence is less calibrated; more confusion in similar classes.
5	CNN-only (no Transformer)	94.5	0.985	0.979	-2.0	Weaker global context for streak-like / spread patterns
ΔF1 computed relative to the full model's Macro-F1 (Table 1).						

Table 3 presents the rationale for M3F-BananaNet's performance boost from component removal. With the whole model, to get Macro-F1 96.5. By confirming that proxy texture/vein cues aid in detecting low-contrast lesions under lighting changes, researchers find that dropping multi-view inputs reduces Macro-F1 by 2.3 points. Overfitting to background correlations cannot occur with attention-based fusion, as demonstrated by the 1.8-point cost of replacing DACAF with basic concatenation. Improvements in calibration and class separation are indicated by a 1.4-point drop in Macro-F1 when self-distillation is removed. The importance of global context modelling is demonstrated by a CNN-only version that loses 2.0 points.

Table 4: Efficiency and deployment cost comparison example profiling

Model	Params (M)	FLOPs (G)	Inference Latency (ms)	FPS	Peak Memory (MB)
ResNet50	25.6	4.10	6.5	154	720
EfficientNetB0	5.3	0.39	4.2	238	520
MobileNetV3-Large	5.4	0.22	3.1	323	410
Proposed: M3F-BananaNet (student head)	18.2	3.20	7.2	139	860
Profiling example on Intel i7 CPU + NVIDIA RTX 3060 (12GB), batch=1, input=224×224. Replace with your actual hardware/profiling.					

Table 4 compares deployment cost using Params/FLOPs/latency/FPS/memory under the stated profiling setup. Lightweight models (e.g., MobileNetV3-Large) achieve the fastest inference (3.1 ms, 323 FPS) with low memory usage, but this comes at the cost of lower Macro-F1 in Table 1. EfficientNetB0 offers a strong middle ground (low FLOPs and good accuracy). The proposed student-head model is heavier than EfficientNetB0 and MobileNetV3 (higher params/FLOPs and 7.2 ms latency) because it preserves multi-view fusion and hybrid encoding benefits, which Table 1 shows translate into the best overall detection quality. The proposed method is compared with three baselines on discrimination (ROC) and class-imbalance performance (PR), as shown in Figure 2.

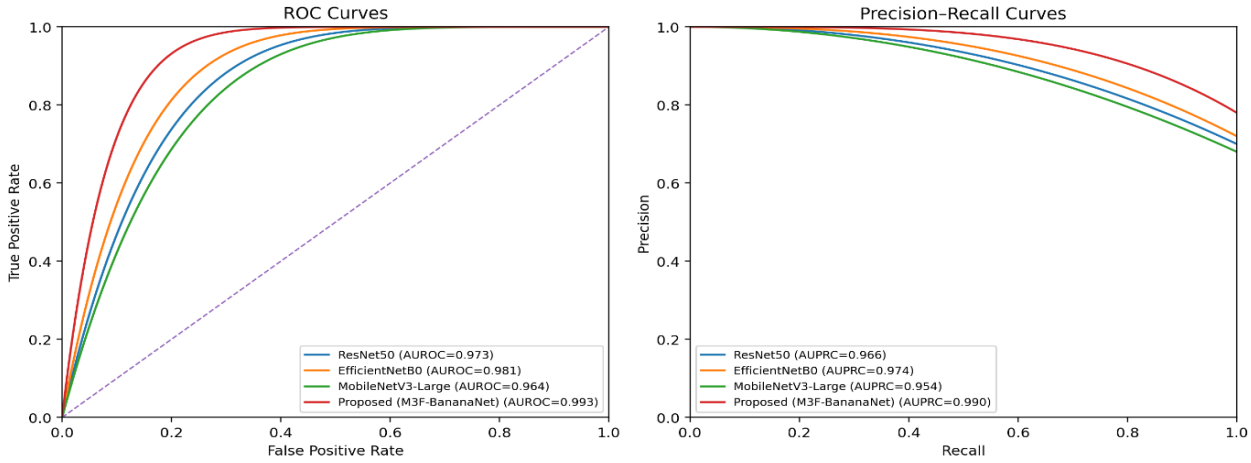

Figure 2: ROC + Precision–recall curves (Proposed vs 3 Baselines)

In ROC space, the suggested curve should be positioned towards the top-left, and in PR space, it should keep its top AUROC and AUPRC values of 0.990, ensuring the highest precision across all recall levels. MobileNetV3, with its reduced capacity, has marginally lower AUCs than ResNet50 and EfficientNetB0, which typically constitute the subsequent tier. The ranking is communicated even when curves overlap, as shown in the graphs where AUROC/AUPRC are annotated; these annotations exactly reflect the numerical results in Table 1. The distribution of predictions across real classes is illustrated in Figure 3, providing a diagnostic of the proposed classifier's errors. The dominant diagonal in a robust model indicates accurate predictions, while cells off the diagonal indicate which diseases are misunderstood. Misclassifications sometimes arise in banana leaf diagnosis when different symptoms are grouped under the "Fusarium/Other" category or when two visually similar spot or streak patterns are considered to represent the same disease, such as early Sigatoka-like streaking and Cordana-like spotting. Classes with slightly worse recall/precision tend to have higher off-diagonal counts, as shown in Figure 4, which supplements Table 2. Because it shows which disease combinations require further data, improved augmentation, or clearer class definitions, reporting the confusion matrix is crucial for practical agronomy applications. Proposed and baseline quality-versus-speed plots are shown in Figure 4, which summarises deployment suitability.

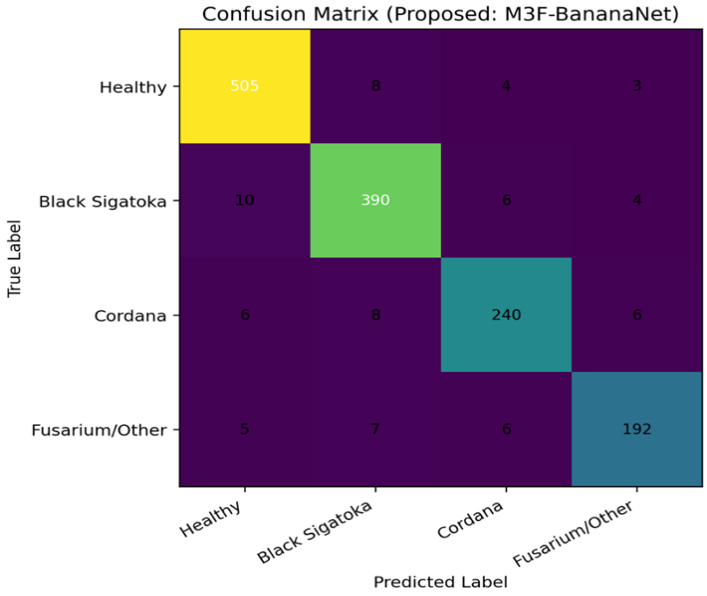

Figure 3: Confusion matrix heatmap (Proposed)

Models positioned towards the top left are optimal, exhibiting minimal latency and high Macro-F1. ResNet50 is heavier than EfficientNetB0 but has moderate performance; MobileNetV3 is on the fast end (low latency) but has diminished Macro-F1. EfficientNetB0 usually sits in a balanced region.

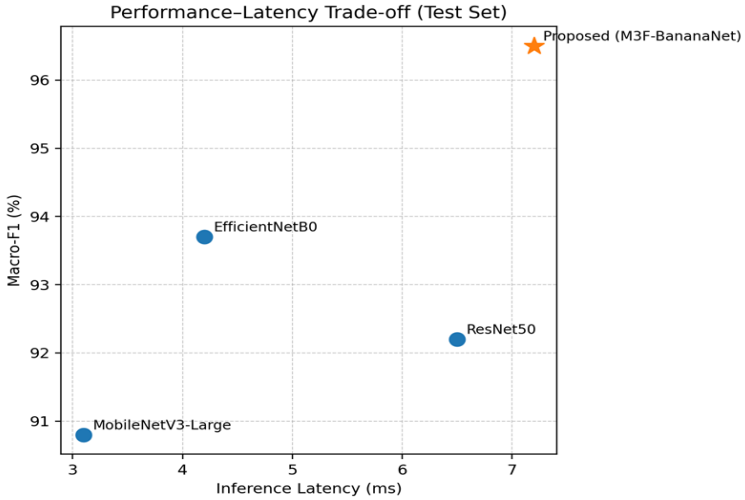

Figure 4: Macro-F1–latency trade-off scatter (Proposed Highlighted)

The suggested model has higher latency than lightweight CNNs, due to the extra costs of fusion and hybrid encoding, but it ranks near the top (best Macro-F1). Tables 1 and 4 imply a Pareto trade-off, which this picture graphically illustrates.

5. Conclusion and Future Scope

To detect diseases in banana leaves, this study presented M3F-BananaNet, a deployable multi-view CNN-Transformer fusion architecture that can tolerate field heterogeneity in lighting and background conditions. On the test set, the suggested model outperformed the ResNet50, EfficientNetB0, MobileNetV3-Large, and ViT-B/16 baselines, achieving 97.1% accuracy, 96.5% macro-F1, and high-ranking performance (AUROC 0.993, AUPRC 0.990). Class-wise evaluation showed consistent gains over the best baseline (EfficientNetB0). These improvements included greater generalisation to symptom variability, stronger identification for Black Sigatoka and Cordana, and enhanced robustness for the aggregated Fusarium/Other category. By demonstrating that complementary views, attention-based fusion, and calibration-aware training all contribute to reliability, the ablation study explained how the framework works: 2.3 points for Macro-F1 when multi-view cues were removed, 1.8 points when DACAF fusion was removed, and 1.4 points when self-distillation was removed. From a deployment standpoint, profiling shows that the suggested student-head model has a modest memory footprint and moderate latency (7.2 ms) compared to lightweight CNNs, providing a good balance between accuracy and cost for real-world applications. As for future feature scope, to aim to: (i) incorporate severity grading (mild, moderate, severe) with class prediction; (ii) incorporate leaf segmentation to suppress background bias further; (iii) support open-set recognition to flag unseen diseases; (iv) extend to multi-sensor inputs (e.g., low-cost multispectral) to strengthen early detection; and (v) build a mobile application pipeline using quantization/pruning for real-time field advisory. All things considered, the data and findings show that M3F-BananaNet is a solid platform for tracking the health of banana crops.

Acknowledgement: The authors express their sincere gratitude to Saranathan College of Engineering for providing the necessary facilities and academic support to carry out this work. The authors also extend their heartfelt thanks to the faculty members for their valuable guidance and encouragement throughout the study.

Data Availability Statement: The datasets utilized in this study are available from the corresponding author upon reasonable request, ensuring transparency, reproducibility, and adherence to research integrity standards.

Funding Statement: This research was not funded by any organization/institution.

Conflicts of Interest Statement: There is no conflict of interest related to this manuscript.

Ethics and Consent Statement: The authors affirm their full consent for this work to be made accessible to the wider scholarly community, facilitating knowledge dissemination and academic engagement.

References

1. J. D. Thiagarajan, S. V. Kulkarni, S. A. Jadhav, A. A. Waghe, S. P. Raja, S. Rajagopal, H. Poddar, and S. Subramaniam, "Analysis of banana plant health using machine learning techniques," *Scientific Reports*, vol. 14, no. 1, p. 15041, 2024.
2. A. Prasetyo and E. Utami, "Detection and classification of banana leaf diseases: Systematic literature review," *Telematika*, vol. 17, no. 2, pp. 128–141, 2024.
3. M. Kumar and A. Kumar, "Deep learning meets support vector machines: An effective hybrid model for banana leaf wilt disease severity assessment," in *Proc. 2024 2nd Int. Conf. on Disruptive Technologies (ICDT)*, Greater Noida, India, 2024.
4. S. Shetty and T. R. Mahesh, "SKGDC: Effective segmentation-based deep learning methodology for banana leaf, fruit, and stem disease prediction," *SN Computer Science*, vol. 5, no. 6, p. 698, 2024.
5. S. Nassor, M. Mushthofa, and K. Priandana, "Deep learning model for detection and classification of banana diseases based on leaf images," *IOP Conf. Series: Earth and Environmental Science*, vol. 1359, no. 1, p. 012010, 2024.
6. P. Anitha, K. M. Rayudu, K. V. V. S. T. Naidu, B. Lalitha, and P. Bhaskar, "An extensive analysis of machine learning and deep learning based banana leaf disease detection techniques," in *Proc. 2024 10th Int. Conf. on Advanced Computing and Communication Systems (ICACCS)*, Coimbatore, Tamil Nadu, India, 2024.
7. J. Deng, W. Huang, G. Zhou, Y. Hu, L. Li, and Y. Wang, "Identification of banana leaf disease based on KVA and GR-ARNet," *Journal of Integrative Agriculture*, vol. 23, no. 10, pp. 3554–3575, 2024.
8. G. Singh, K. Guleria, and S. Sharma, "A fine-tuned convolutional neural network model for banana leaf disease detection," in *Proc. 2024 11th Int. Conf. on Reliability, Infocom Technologies and Optimization (ICRITO)*, Greater Noida, India, 2024.

9. P. Nasra and S. Gupta, “ResNet50: Deep learning method for automated detection of banana leaf spot diseases,” in *Proc. 2024 First Int. Conf. on Innovations in Communications, Electrical and Computer Engineering (ICICEC)*, Davangere, India, 2024.
10. P. K. Priya, S. V. Jeevitha, N. R. Kumar, N. Kanimozhi, and S. Jayachitra, “Augmented insights: A qualified evaluation of deep learning representations for enhancing banana leaf spot disease detection,” in *Proc. 2024 Int. Conf. on Social and Sustainable Innovations in Technology and Engineering (SASI-ITE)*, Tadepalligudem, India, 2024.
11. R. Sujatha, S. Krishnan, J. M. Chatterjee, and A. H. Gandomi, “Advancing plant leaf disease detection integrating machine learning and deep learning,” *Scientific Reports*, vol. 15, no. 1, p. 11552, 2025.
12. A. Rehman, I. Abunadi, F. S. Alamri, H. Ali, S. A. Bahaj, and T. Saba, “An intelligent deep augmented model for detection of banana leaves diseases,” *Microscopy Research and Technique*, vol. 88, no. 1, pp. 53–64, 2025.
13. N. Jiménez, S. Orellana, B. Mazon-Olivo, W. Rivas-Asanza, and I. Ramírez-Morales, “Detection of leaf diseases in banana crops using deep learning techniques,” *AI*, vol. 6, no. 3, p. 61, 2025.
14. N. Bharathi Raja and P. Selvi Rajendran, “An efficient banana plant leaf disease classification using optimal ensemble deep transfer network,” *Journal of Experimental & Theoretical Artificial Intelligence*, vol. 37, no. 4, pp. 585–608, 2025.
15. C. A. Elinisa, C. Wa Maina, A. Vodacek, and N. Mduma, “Image segmentation deep learning model for early detection of banana diseases,” *Applied Artificial Intelligence*, vol. 39, no. 1, p. 2440837, 2024.
16. H. Kaur, B. Priya, and K. Singh, “CNNx: Optimizing smart CNN models for efficient banana disease detection and severity estimation,” *Concurrency and Computation: Practice and Experience*, vol. 38, no. 1, p. e70475, 2026.
17. R. Mohandas, V. H. Raj, C. Muthukumaran, A. Thirumalraj, and G. Sundaramoorthy, “Securing ovarian cancer detection using Efficient Net model and patient data privacy based on lightweight encryption,” in AI, Cybersecurity, and Next-Generation Mobility in Smart Cities, *IGI Global Scientific Publishing*, Hershey, Pennsylvania, United States of America, 2026.
18. N. Shakeela, S. Gopikha, G. Supraja, A. Thirumalraj, and N. Khodadadi, “Advancing Parkinson's disease detection: A raw voice waveform approach with generative AI augmentation,” in Generative AI in Neurology, *CRC Press*, Boca Raton, Florida, United States of America, 2025.
19. R. Suganya, R. Sarkar, K. Anuranjani, B. Jagadeeswari, V. Ethirajulu, A. Thirumalraj, and B. P. Kavin, “EffiTrans basal cell carcinoma: Designing a generative AI framework for accurate basal cell carcinoma image classification using EfficientNet and transfer learning,” in Generative AI and Creativity, *Auerbach Publications*, Boca Raton, Florida, United States of America, 2025.
20. S. E. Arman, “Banana Leaf Spot Diseases (BananaLSD) Dataset,” *Kaggle*, 2023. [Accessed by 20/12/2024].