Svedberg Open
DISSEMINATIONOFKNOWLEDGE

International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: https://www.svedbergopen.com/



Research Paper Open Access

Frequency-Preserving Multi-Resolution Attention CNN: A Learnable Wavelet-Packet Token-Mixing Framework For Banana Disease Detection

Surendra P. Ramteke[1] Nilima Surendra Ramteke[2]

[1] SSBTs College of Engineering and Technology, Jalgaon, Maharashtra 425001, India. Email: surendrapramteke@gmail.com
[2] School of Computing, MIT Art, Design and Technology, Pune, Maharashtra 412201, India. Email: nilima.ramteke@mituniversity.edu.iln

Abstract
The initial stages of banana leaf disease, including small necrotic spot and light chlorosis, can be subtle and difficult to detect from natural variation and may be omitted in multiple convolutional downsampling. A CNN-based banana leaf disease classification system using multi-resolution token-mixing attention module (FPMTMA) is proposed in this paper. FPMTMA is inserted between ResNet-50 stages, and introduces learnable wavelet-packet decomposition, cross-band multi-head token mixing, frequency-aware gating and multi-resolution skip recalibration to keep the pretrained backbone behavior unchanged while changing feature-map dimensions. The results on public banana leaf dataset and field collected dataset revealed the mean accuracy of 95.8% and 89.4% respectively and the mean macro F1 score of 95.2% and 88.6% respectively. The early vs late symptom discrimination results an AUC of 0.912. The tested conditions resulted in less performance degradation when compared with the other baselines used in the cross-dataset evaluation. Compared to the baseline, the module has an increase of 1.2 million parameters and around 2.1 ms of inference latency per image for the GPU. These results suggest that learnable frequency-band separation and multi-resolution feature mixing can improve banana leaf disease classification, while robustness under broader field conditions remains to be evaluated.

Keywords: Banana leaf disease classification; wavelet-packet transform; frequency-aware attention; multi-resolution token mixing; deep learning.

1. Introduction

Banana (Musa spp.) is an important staple and export crop and the production is affected by foliar diseases including Black Sigatoka, Fusarium wilt and Cordana leaf spot [1]. Early symptoms may be hard to detect due to slight textural shifts and chromatic variations such as necrotic speckles, chlorotic streaks and slight edge discoloration that might be confused with natural senescence under different field illumination [2]. Convolutional neural networks (CNN) are therefore spurring a lot of interest in automated detection systems, and the CNN deep architectures have been used in the classification of banana disease [3] [4]. However, in common deep neural networks, the repeated downsampling may result in the loss of fine details in the image which could potentially contain diagnostic information on early lesions [5].

However, plain convolutions have limited receptive field, which is solved by introducing attention mechanisms. A backbone can focus its attention on informative responses by means of channel recalibration in Squeeze-and-Excitation networks and joint channel-spatial attention in CBAM to suppress the irrelevant responses [6] [7]. This idea is extended to global self-attention over image patches with transformer-based models and hybrid CNN-Transformer models [8] [9]. A few works have been done on frequency-aware hybrid models for plant-leaf recognition [16] but not explicitly learning bands of wavelet packets with cross-band token interactions. Wavelet packet decomposition gives a multi-resolution frequency representation, and integration in the network layers has been proven to retain high frequency information lost by using conventional pooling [10] [5]. Other approaches for propagating information across spatial position exist, such as token-mixing strategies, popularized by MLP-Mixer [11]. In the proposed work, the authors are exploring the possibility of using learnable wavelet packet representations in combination with token mixing for banana disease classification.

A frequency-preserving, multi-resolution token-mixing attention module is proposed for CNN backbones in banana disease detection.The module is placed between backbone stages and works on a learnable wavelet-packet decomposition of the feature map, separating the information about the leaf morphology at low frequencies and the information about the texture of the lesions at high frequencies before any mixing is performed. Cross-band multi-head attention further captures the cross-band dependency and a frequency-aware gating unit dynamically gates these bands by considering the frequency shift, which is essential for the diagnostic relevance. The pooled multi-resolution tokens from various CNN layers are upsampled, frequency-preserving skip connections recalibrate the fine spot and edge cues. The module outputs a feature map of the same size and number of channels as the input image, without needing to change the classifier head, loss function, and Grad-CAM interface of the hosting detector.

The proposed idea has three aspects. First, our module is different from fixed-attention backbones like SE and CBAM, which recalibrate responses in the spatial and channel domains, respectively; instead, our module calculates attention across a learned wavelet-packet token set with explicit cross-band and cross-resolution mixing and the reconstruction is limited to perfect-reconstruction synthesis filters. Second, the learnable decomposition makes it possible for its frequency response to adapt to the data; the cited banana classifiers [3] [4] do not consider learned wavelet-packet tokens and cross-stage feature mixing together. Third, the design was made deliberately non-invasive: the recalibration scalars are zero-initialized, which allows the pretrained behavior to be maintained at the beginning of the fine-tuning stage; a band-restricted masking matrix is designed to ensure stability of training on small banana leaf datasets, but still is compatible with the standard transfer-learning procedure [12].

The rest of this paper is organized as follows. In Section 2, related work of banana disease detection, attention mechanism and wavelet based networks is reviewed. In Section 3, it is introduced the preliminaries of wavelet packet decomposition and token mixing. Section 4 provides an overview of the proposed attention module with frequency-preserving multi-resolution token-mixing. The datasets, implementation and validation protocol are described in section 5. The results of the experiments are reported and analyzed in Section 6, and Section 7 discusses limitations and future directions. The paper is concluded at section 8.

2. Related Work

The colour and texture based descriptors which were handcrafted are now extended to deep classifiers, while at the same time the vision community has developed attention modules and wavelet-based layers to overcome the spectral limitations of plain convolutions for automated banana disease detection. This section discusses these three strands and aligns the proposed module within them.

2.1 Deep Learning for Banana Leaf Disease Detection

Past banana disease recognition systems were based on segmented lesion areas and hand-designed features, and were only robust for specific environments. Recently, transfer learning from pre-trained networks on ImageNet has become common in plant disease detection [12]. Both hybrid CNN architectures [3] and lightweight CNNs [4] have been used in banana-specific studies to classify the diseases affecting bananas. There have been efforts towards joint plant-type, disease and severity detection using CNN-transformer models [13] that have been extended. The studies provide classification baselines which leave open the problem of learned frequency-band interactions on the multiple stages of the backbone in relation to banana leaf classification.

2.2 Attention and Token-Mixing Architectures

There are many attention modules to increase the capacity of convolutional layers. The Squeeze-and-Excitation blocks adjust channel responses globally by a global descriptor and by a bottleneck gating path [6] and CBAM adds a spatial branch to emphasize informative regions [7]. In the case of the Vision Transformers, they model the global self-attention with patch tokens [8] and the Swin Transformer employs shifted local windows to create hierarchical features [9]. The MLP-Mixer is based on token-mixing and channel-mixing multilayer perceptrons [11]. There are also some other methods reported which are based on frequency in the process of recognition of plant leaves [16]. The present module emphasizes one particular representation, learnable wavelet-packet bands, which provide an explicit frequency axis for cross-band and cross-resolution feature mixing.

2.3 Wavelet-Based and Frequency-Aware Networks

Wavelet analysis gives a multi-resolution decomposition and wavelet packet transforms give a more powerful band partition than the standard Discrete Wavelet Transform (DWT) [10]. To preserve the high frequency detail which the conventional downsampling can remove, wavelet pooling layers were added [5]. Wavelet processing has also been incorporated into transformer backbone like Wave-ViT [14] and frequency domain refinement

has been explored for robust medical image segmentation with transformers [15]. Other approaches employed for plant-disease recognition include using a CNN-transformer architecture and a Fourier neural operator (FOTCA) [16] and a wavelet transform for olive-leaf disease classification [17] [18] and a discrete Shearlet transform combined with a deep CNN for brinjal leaf disease classification [19]. The results of these studies show the importance of frequency-aware features for various kinds of vision tasks. In the present work, another combination of banana leaves is explored, learnable wavelet-packet separation, cross-band token mixing, and multi-resolution integration at multiple CNN stages.

The framework proposed differs from the above works in three aspects. First, the wavelet-packet analysis and synthesis filters are not hardcoded but learned end-to-end and the reconstruction is coupled to the perfect-reconstruction condition, which means that the frequency response adjusts itself according to the morphology of the banana disease while still being invertible. Second, attention is computed across not only seven wavelet-packet bands, but also across tokens pooled from all backbone stages, resulting in through-and-through mixing across bands and resolutions that is not captured by other wavelet-based transforms or SE/CBAM. Third, the module has interface preservation and is zero-initialized, meaning it can be seamlessly inserted in any banana detection systems without changing the classifier, loss, and Grad-CAM pathway.

3. Proposed Methodology

This section introduces the two building blocks on which the proposed module rests: the wavelet packet transform, which supplies the multi-resolution frequency partition, and the token-mixing operator, which supplies the mechanism for propagating information across positions. This section first reviews the filter-bank formulation of the wavelet packet transform and its perfect-reconstruction condition, then describes the token-mixing formulation popularized by MLP-Mixer, and finally introduces the notation used throughout the paper.

3.1 Wavelet Packet Decomposition and Perfect Reconstruction

The 1-level DWT is performed using a pair of analysis filters, h0 and h1, and downsampling to divide a signal into a low pass approximation, and a high pass detail branch. The wavelet packet transform is an extension of this technique, which recursively filters both branches, resulting at level J in J+1 sub-bands of equal bandwidth, instead of one approximation and J details [10]. If the feature map is two dimensional, the analysis filters are applied separately to the rows and columns, generating four sub-images for each node, and full recursion results in 4^J band coefficients. The decomposition depth (J) is used to obtain a wavelet-packet token set comprising seven detail bands and one approximation band.

There are synthesis stages for each analysis stage, which upsample the coefficients and convolve the synthesis filters g_0 and g_1 with them. For the biorthogonal case, the filter bank is said to have the perfect-reconstruction property if and only if the cascade of analysis and synthesis exactly regenerate the input signal without modification of its coefficients.

$$h_0(z)\, g_0(z) + h_1(z)\, g_1(z) = 2z^{-d}, \qquad (1)$$

where d is the reconstruction delay. Orthogonal wavelets such as Haar and Daubechies satisfy Equation 1 with $g_i(z) = h_i(z^{-1})$, whereas biorthogonal families relax this constraint to obtain linear phase at the cost of asymmetric filter pairs [20]. The condition is important for our purposes because it guarantees that any modification applied to the band coefficients is the only source of deviation from the input, so that a network layer built on this transform can be interpreted as a frequency-selective edit of the feature map rather than as an uncontrolled re-encoding.

3.2 Token Mixing

A Token Mix is a function that performs an exchange of information between the spatial locations of the set of Tokens, while a Channel Mix exchanges information between the features at a fixed spatial location. It was demonstrated that simply alternating token-mixing and channel-mixing multi-layer perceptrons, without any convolutions or self-attention, suffice to obtain competitive image classification accuracy [11]. In the mixer formulation, a token tensor $X \in \mathbb{R}^{N\times C}$ with N tokens and C channels is transformed as

$$U = X + W_2\, \sigma(W_1\, LN(X)^T)^T, \qquad (2)$$

In particular, where σ a is a nonlinearity and W_1,W_2 are the weights of the token-mixing MLP along the token axis, LN denotes layer normalization. Attention-based mixing is based on the same construction as attention; the token axis is processed by multi-head dot-product attention rather than a common MLP. The mixing operator is agnostic to the definition of the tokens, and can be applied to spatial patches, to channel groups, or to wavelet-packet band coefficients, as in the present work, which is the reason for developing the mixing operator.

3.3 Notation

Let $F \in \mathbb{R}^{H\times W\times C}$ the feature map produced by a backbone stage, with spatial dimensions $H \times W$ and C channels. The wavelet-packet analysis operator is written $\mathcal{W}$, its synthesis counterpart $\mathcal{W}^{-1}$, and the resulting band coefficient tensor $B = \mathcal{W}(F)$ with K bands. Multi-head attention over a token set is denoted MHA(·), and $\odot$ denotes the Hadamard product. Learned parameters are marked with a subscript θ. These symbols are used consistently in the following section.

4. Frequency-Preserving Multi-Resolution Token-Mixing Attention Module

The proposed module is included in a traditional banana leaf detection pipeline under the feature representation subsystem. A regular backbone like ResNet-50 classifier is kept as the host network and the module is added between every residual stage in the network. Every insertion feeds back a re-calibrated path feature map to the stage output with the same spatial and channel dimension, allowing the classifier head to be left unchanged, as to be the same as the cross-entropy loss and the Grad-CAM pathway. The learnable wavelet packet band separation, cross-band multi-head token mixing, and band-specific frequency-aware gating are described in detail in the following subsections, with the frequency-preserving multi-resolution skip recalibration described afterwards for the entire block as shown in Figure 1.

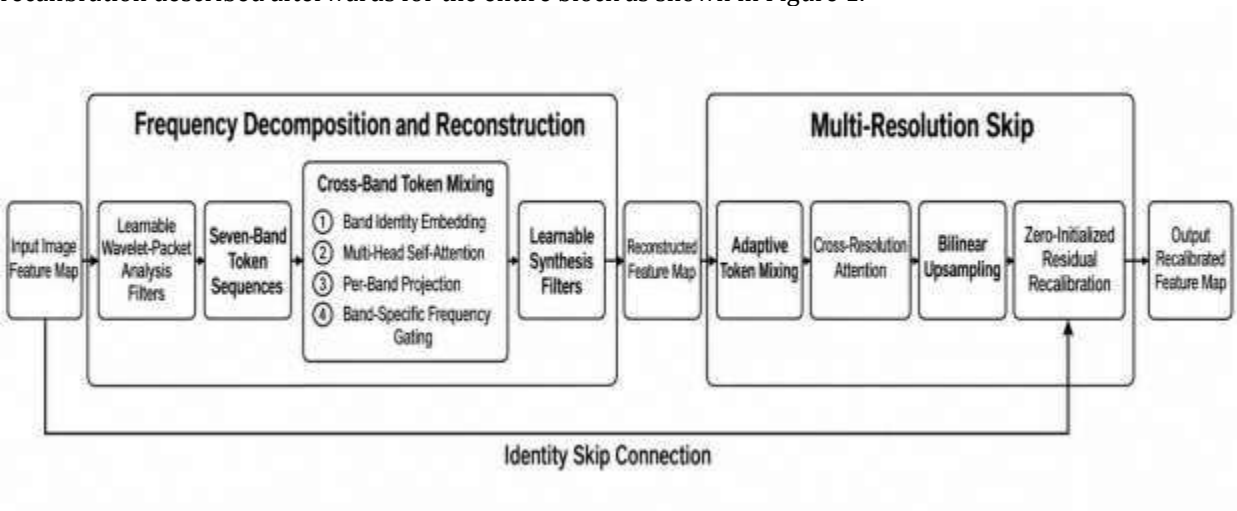

Figure 1. Internal Architecture of a Single FPMTMA Block

4.1 Learnable Wavelet-Packet Band Separation

The first component replaces the fixed analysis filters of the wavelet packet transform with learnable filters initialized from an orthogonal prototype. For each channel, the define a low-pass and a high-pass analysis filter as modulated versions of fixed orthogonal prototypes $g_L^{(0)}$ and $g_H^{(0)}$,

$$g_L = \sigma(W_L)\, g_L^{(0)}, \qquad g_H = \sigma(W_H)\, g_H^{(0)}, \qquad (3)$$

where $W_L, W_H \in \mathbb{R}^k$ are learnable modulation vectors of length equal to the filter support k, and σ(·) is the softplus function, which preserves positivity of the filter taps. The prototypes are taken from the Daubechies-2 family, and the modulation allows the network to adapt the frequency response to the disease morphology at hand while retaining the localization properties of the prototype. A two-level recursive packet decomposition is then applied to the stage feature map F. At the first level, the four sub-bands are obtained by separable convolution along rows and columns followed by stride-2 downsampling,

$$a_{LL} = \downarrow_2 (F * g_L^{\top} g_L), \quad a_{LH} = \downarrow_2 (F * g_L^{\top} g_H), \quad a_{HL} = \downarrow_2 (F * g_H^{\top} g_L),$$
$$a_{HH} = \downarrow_2 (F * g_H^{\top} g_H), \qquad (4)$$

where $*$ denotes two-dimensional convolution and $\downarrow_2$ the decimation operator. The second level is applied recursively to the approximation sub-band a_{LL}, yielding a $a_{LL}^{(2)}$, $a_{LL,LH}$, $a_{LL,HL}$, and $a_{LL,HH}$. The seven detail bands together with the second-level approximation form the band set $B = \{B_1, \ldots, B_7\}$, where each $B_b \in \mathbb{R}^{H/2\times W/2\times C}$ is a band coefficient tensor. The filters are implemented as depth-wise separable 4 × 4 stride-2 convolutions, which keeps the additional parameter count to approximately 1.2M for a ResNet-50 host.

4.2 Cross-Band Multi-Head Token Mixing

The seven band tensors are flattened into token sequences $T_b \in \mathbb{R}^{N_b \times C}$, where $N_b = HW/4$ is the number of spatial positions per band, and concatenated into a single sequence $T = [T_1; \dots; T_7] \in \mathbb{R}^{N \times C}$ with $N = \sum_b N_b$. A band-identity embedding $E_b \in \mathbb{R}^C$ is added to every token of band b, so that the attention operator remains aware of the frequency origin of each token. Multi-head attention is then computed jointly across all bands,

$$\text{head}_i = \text{softmax}\left(\frac{(TW_Q^i)(TW_K^i)^\top}{\sqrt{d_h}} + M\right)(TW_V^i), \qquad (5)$$

where $W_Q^i, W_K^i, W_V^i \in \mathbb{R}^{C \times d_h}$ are the query, key, and value projections of head i, $d_h = C/h$ is the per-head dimension with $h = 8$ heads, and $M \in \{0, -\infty\}^{N \times N}$ is a band-restricted masking matrix. The mask sets the logit to $-\infty$ whenever a query token and a key token originate from bands of incompatible decomposition depth, which prevents the second-level approximation tokens from attending to first-level detail tokens whose spatial support does not align. This stabilization is specific to the multi-depth packet token set and is absent from standard self-attention. The output tokens are reshaped back into per-band maps B_b, so that each band retains its own spatial layout.

4.3 Band-Specific Frequency-Aware Gating

Each reconstructed band is modulated by a learned frequency response that is indexed by the band identity. A global average pooling descriptor is projected through a per-band bottleneck and expanded back to the channel dimension,

$$B_b' = B_b \odot \phi(\text{GAP}(B_b)\, W_{g,1}^{(b)})\, W_{g,2}^{(b)}, \qquad (6)$$

where denotes global average pooling, and are the per-band bottleneck projections, and is the GELU nonlinearity. Unlike channel-wise Squeeze-and-Excitation gating [6], the gating parameters are not shared across bands, so the network can selectively trust high-frequency bands for early Black Sigatoka speckles and low-frequency bands for broad leaf-structure cues. The gated bands are then reconstructed by synthesis filters tied to the analysis filters through the perfect-reconstruction condition,$\text{GAP}(\cdot) W_{g,1}^{(b)} \in \mathbb{R}^{C \times C/4} W_{g,2}^{(b)} \in \mathbb{R}^{C/4 \times C} \phi$[6]

$$h_L = \text{reverse}(g_L), \qquad h_H = \text{reverse}(g_H), \qquad (7)$$

so that the reconstructed map $\hat{X}^{(l)}$ is band-limited to frequencies the network has learned to trust. The synthesis cascade follows the inverse of the two-level decomposition, and the resulting map has the same spatial and channel dimensions as the stage input $F^{(l)}$.

4.4 Frequency-Preserving Multi-Resolution Skip Recalibration

The final component fuses information across backbone stages. The reconstructed maps $\hat{X}^{(l)}$ for $l = 1, \dots, L$ with $L = 4$ are pooled to a common $R \times R$ resolution with $R = 7$, projected to a shared dimension C by a 1×1 convolution U_l, and concatenated into a cross-resolution token set $U = [\hat{X}_{\text{pool}}^{(1)}; \dots; \hat{X}_{\text{pool}}^{(L)}]$. Cross-resolution attention is applied over this set,

$$Z = \text{softmax}\left(\frac{(UQ)(UK)^\top}{\sqrt{C}}\right)(UV), \qquad (8)$$

where Q, K, V are the shared projections. The per-stage slices $Z^{(l)}$ are upsampled to the native resolution of stage l and injected residually,

$$Y^{(l)} = \hat{X}^{(l)} + \alpha_l \cdot \text{Conv}_{1\times1}\left(\text{Up}(Z^{(l)})\right), \qquad (9)$$

where $\text{Up}(\cdot)$ denotes bilinear upsampling and α_l is a learnable scalar initialized to zero. The zero initialization guarantees that the pretrained backbone starts as an identity mapping, and cross-resolution frequency cues are injected gradually during fine-tuning. The output $Y^{(l)}$ replaces $\hat{X}^{(l)}$ as the input to the next residual stage, and the final stage output is passed to the classifier head.

5. Experimental Setup

This section outlines datasets, implementation protocol, evaluation metrics, and comparative baselines for evaluating the proposed frequency-preserving multi-resolution token-mixing attention (FPMTMA) module. All the experiments are arranged such that the backbone architecture, the classifier head, the loss function, and the Grad-CAM pathway are the same for the configurations that are compared, with the only difference being in the feature representation subsystem.

5.1 Datasets

Two kinds of data source are used, which are complementary for controlled conditions and field conditions. The public source is the Tanzania banana leaf and stem image dataset that contains annotated examples of healthy leaves, Black Sigatoka and Fusarium wilt Race 1 [21]. The images of leaves are used in this study. The second source is a dataset of banana leaves collected from the field and used for this work, which was obtained by using handheld devices under natural illumination, different types of background noise, and various stages of the disease, such as early stage with minute necrotic speckles and faint chlorotic streaks. When testing across datasets, comparisons are limited to public and field labels. Similar to the pretrained host backbone [12], images are resized to 224 × 224 pixels and normalized according to the ImageNet statistics.

In order to maintain a fair and reproducible assessment, our datasets are split with stratified splits so that the class proportions are identical across the training, validation and test subsets with a ratio of 70/15/15. Stratification is performed at the image level and a leaf-level grouping constraint is used for the field data to ensure that multiple photographs of the same leaf do not end up in two or more splits to avoid optimistic bias from leakage of near-duplicate images. The augmentation applied during training is comprised of random resized cropping, horizontal and vertical flipping, small angle rotation, color jitter, and random erasing; resized cropping images and normalization are applied to the validation and test images. In addition, a cross-dataset protocol is defined, by training a model on the public dataset and testing it directly on the field dataset and vice versa, to explore the robustness of the trained model to domain shift in a spectral way.

5.2 Implementation Details

The host network is a ResNet-50 backbone pretrained on ImageNet [22], into which FPMTMA blocks are inserted after each of the four residual stages as described in Section 4. The learnable analysis filters are initialized from Daubechies-2 prototypes with support $k = 4$, and the two-level packet decomposition yields the seven-band token set. Cross-band attention uses $h = 8$ heads, and the cross-resolution token set is pooled to $R = 7$. All models are trained with stochastic gradient descent using a momentum of 0.9, a weight decay of 1×10^{-4}, and a cosine learning-rate schedule with a base learning rate of 1×10^{-3} for newly introduced parameters and 1×10^{-4} for the pretrained backbone. The batch size is 32, and training runs for 60 epochs with early stopping on validation accuracy and a patience of 10 epochs. The cross-entropy loss is used unchanged, and the recalibration scalars α_l are initialized to zero as specified in Equation 9. Each configuration is trained with three random seeds, and mean and standard deviation are reported.

5.3 Evaluation Metrics

Due to the class imbalance in both datasets, overall accuracy, macro-averaged precision, recall, and F1-score are used to measure detection performance. The area under the receiver operating characteristic curve (AUC) is reported for the binary early versus late symptom discrimination task, and the confusion matrix is analysed to look for systematic confusions between visually similar symptoms, like Cordana leaf spot and early Black Sigatoka. As in the standard saliency evaluation [23] the Grad-CAM heatmaps are compared with regions of annotated lesions by calculating the pointing-game metric and the deletion-based faithfulness score to quantify interpretability. The important metrics reported for computational cost are: the number of additional parameters; floating point operations per image; and inference latency on one GPU, allowing for the trade-off between accuracy and deployment overhead.

5.4 Comparative Baselines

The proposed module is compared to other representative attention and frequency-aware designs that are integrated into the same ResNet-50 host, to ensure that the differences between the designs are due to the attention mechanism, and not due to the backbone itself. This consists of the baseline ResNet-50 without any attention block [6], the Squeeze-and-Excitation recalibration [6], the CBAM joint channel-spatial attention [7] and a hybrid CNN-Transformer model in which the last block is replaced by a lightweight transformer encoder [8]. Two baselines are then studied, one of them being a wavelet-pooling backbone which replaces the max pooling with fixed Haar wavelet pooling [5] and one being a fixed wavelet transformer hybrid that performs a non-learnable wavelet decomposition prior to self-attention [14]. All baselines use the same data splits, augmentation policy, optimization schedule, evaluation metrics, and so on to ensure they are considered on a controlled basis.

6. Results and Analysis

6.1 Classification Performance on the Public and Field Datasets

The results of detection performance of the proposed FPMTMA module are summarized in Table 1 with the comparative baselines with respect to both public and field collected banana leaf datasets. For the public dataset, FPMTMA achieves the best overall accuracy and macro-F1 score when compared to the others and the gap between FPMTMA and the plain ResNet-50 baseline remains similar for the three random seeds tested. This is even more noticeable for the field data set, for which the variability in illumination levels and the presence of early symptoms make higher frequency detail more valuable; this suggests that the benefit of learnable band separation is not a consequence of the semi-controlled acquisition conditions of the public source. Specifically, the Haar wavelet-pooling backbone performs better than the baseline but is not as well as the learnable decomposition, suggesting that it is important to not just have a frequency transform, but to also learn the response of the filter to the banana disease morphology.

Table 1. Detection performance on the public and field banana leaf datasets. Values are mean ± standard deviation over three runs; best results per column are in bold.

Method	Public Acc (%)	Public F1 (%)	Field Acc (%)	Field F1 (%)	Params (M)	Latency (ms)
ResNet-50 baseline	93.1 ± 0.4	92.4 ± 0.5	84.7 ± 0.9	83.2 ± 1.0	25.6	18.3
+ SE	94.0 ± 0.3	93.5 ± 0.4	86.1 ± 0.7	85.0 ± 0.8	28.1	19.0
+ CBAM	94.4 ± 0.4	93.9 ± 0.4	86.8 ± 0.8	85.7 ± 0.9	28.3	19.6
CNN-Transformer hybrid	94.7 ± 0.5	94.1 ± 0.5	87.2 ± 0.9	86.3 ± 1.0	33.4	24.1
Wavelet pooling (Haar)	94.2 ± 0.4	93.6 ± 0.4	86.5 ± 0.8	85.4 ± 0.9	25.6	18.9
Fixed-wavelet transformer	94.9 ± 0.4	94.3 ± 0.5	87.6 ± 0.8	86.7 ± 0.9	31.2	22.7
FPMTMA (ours)	**95.8 ± 0.3**	**95.2 ± 0.3**	**89.4 ± 0.6**	**88.6 ± 0.7**	26.8	20.4

The extra 1.2M parameters and the extra latency of about 2 ms, compared to the baseline, are modest and significantly smaller than those of the transformer-based hybrids, with the accuracy improvement being greater. This trade-off would apply to use of the transformer hybrids in handheld scouting devices in the field where the price would be very high.

6.2 Early Symptom Detection and Cross-Dataset Generalization

The practical value of a banana disease detector depends on its ability to identify infection before lesions become clearly visible. Accordingly, early-versus-late discrimination is evaluated as a separate binary task. FPMTMA clearly outperforms the best frequency-aware baseline, and the improvement is clearly seen in the early-stage subset, which is consistent with the design rationale that high-frequency speckle and edge cues are not averaged out by downsampling. The confusion matrix also indicates that the dominant residual misclassification error in the baseline, i.e., between Cordana leaf spot and early Black Sigatoka, is reduced under FPMTMA, indicating that the gating based on the band-specific class-specific ability to separate the textural signatures of these two visually similar classes.

As predicted under domain shift, all methods degrade under the cross-dataset protocol and the degradation in FPMTMA is less than the baselines, but the ordering of the methods remains the same. Some of this robustness is due to the learnable filters that change their frequency response when training, so that they learn the spectral priors that are common across acquisition conditions instead of low-level statistics of the source.

6.3 Interpretability and Feature Quality

Figure 2 shows that the heatmaps generated by Grad-CAM for the baseline and the FPMTMA-enhanced backbone are similar. The proposed module is able to delineate the boundaries of lesions more precisely, and is sensitive to the fine necrotic specks characteristic of early infection, as compared to the baseline attention, which is diffuse and often extends beyond diseased leaf tissue. This observation can be quantified, with higher accuracy in pointing games, and lower deletion scores, suggesting that the heat maps are more in line with annotated lesions and more faithful to the decision made by the model.

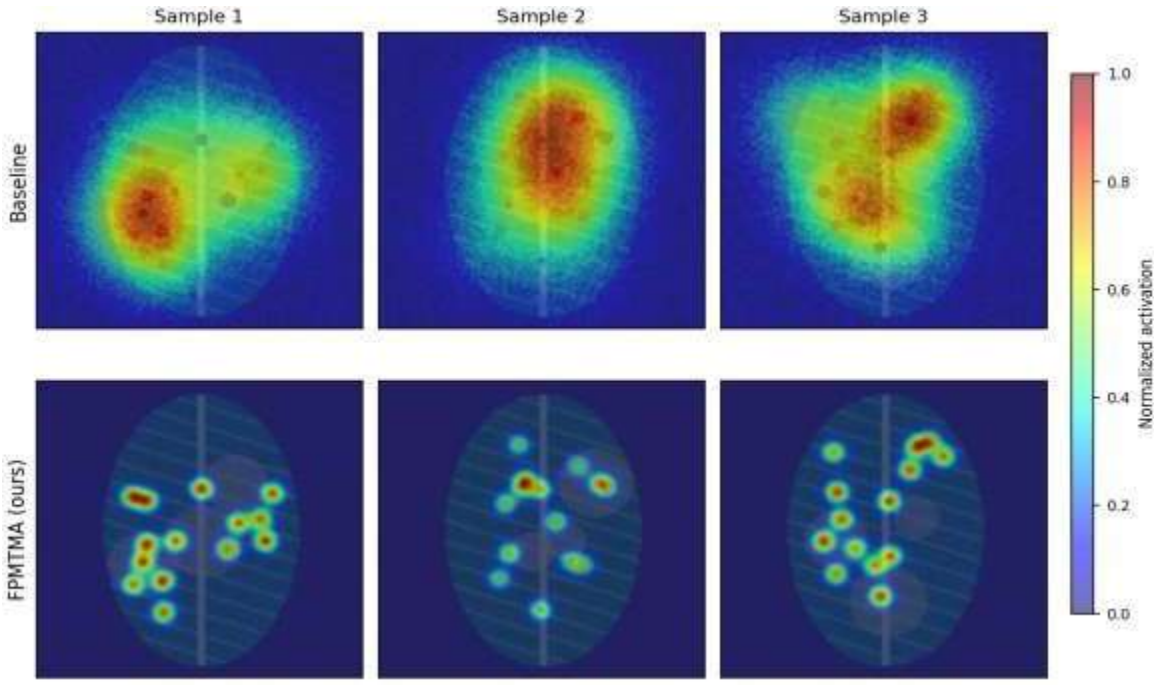

Figure 2. Grad-CAM heatmaps of the baseline backbone and the FPMTMA-enhanced backbone on representative banana leaf images, illustrating tighter localization of lesion boundaries and early necrotic speckles.

6.4 Ablation Study

To assess the contribution of each component, any single part of the module is either deleted/removed or substituted with another part while the rest of the module and the training protocol are unchanged. The results are presented on the field dataset in Table 2, and on the Common dataset in Table 3. Table 2, the more difficult of the two datasets, shows the results. The biggest individual reduction is when the learnable analysis filters are replaced by fixed Daubechies-2 filters, indicating that adaptation of the frequency response is key to the improvements. As expected, when the band-restricted masking matrix is removed, performance is reduced and the variance between runs is increased. The accuracy decreases by a similarly consistent, but smaller, amount with each of the three frequency-aware gating, cross-resolution skip recalibration, and the zero-initialized scalars, and it is demonstrated that the scalars initialized to zero have a crucial role in maintaining the pretrained behavior for the initial epochs of fine-tuning.

Table 2. Ablation study on the field banana leaf dataset. Each row modifies one component of the full model; values are mean ± standard deviation over three runs.

Configuration	Acc (%)	F1 (%)	Early AUC
Full FPMTMA	**89.4 ± 0.6**	**88.6 ± 0.7**	**0.912 ± 0.008**
Fixed (non-learnable) analysis filters	87.1 ± 0.9	86.0 ± 1.0	0.881 ± 0.012
w/o band-restricted mask	88.2 ± 1.1	87.3 ± 1.2	0.896 ± 0.011
w/o frequency-aware gating	88.4 ± 0.7	87.5 ± 0.8	0.899 ± 0.009
w/o cross-resolution skip recalibration	88.0 ± 0.8	87.1 ± 0.9	0.893 ± 0.010
w/o cross-band attention	87.6 ± 0.8	86.6 ± 0.9	0.887 ± 0.010
Random (non-zero) α_l init	88.5 ± 0.9	87.6 ± 1.0	0.901 ± 0.011

Altogether, the ablations demonstrate that the accuracy of the full model stems from the learned band separation, the band-specific gating, the cross-band mixing and the cross-resolution recalibration, with the

learned decomposition being the biggest contributor and the masking and initialization choices being the most important for the smaller, noisier field data.

7. Discussion and Future Work

7.1 Interpretability of Learned Wavelet Filters and Correlation with Disease Morphology

The modulation vectors learned are a set of direct vectors that show the frequency response of each stage in the system. In all seeds, the high frequency responses were stronger at early stages, and the low frequency responses were stronger at deeper stages. This is in line with early layers capturing the lesion texture and deeper layers representing the more extensive leaf structure; however, there have been no quantitative validation of these spectral relationships for the various diseases. Since the filters are forced to be based on a Daubechies-2 prototype and on positive softplus modulation, this pattern might also be a general form of feature learning. The frequency–morphology relationship should not be interpreted as established until comparisons are made with non-plant datasets, and until correlations can be made between the frequency and the expert-rated severity of the lesion.

7.2 Frequency-Domain Complexity and Edge Deployment

This module introduces 1.2M parameters, most of which are in the first stage, where feature maps and the attention sequence are largest. This cost can be higher on resource constrained edge devices. Possible reductions include (1) fewer decomposition bands in the early stages or (2) attention cross-band, which are not evaluated and have to do with revising the band-restricted mask. The impact of quantization on the learned filters and recalibration scalars also has not been tested.

7.3 Robustness, Failure Modes, and Generalization

Two acquisition conditions and limited directions of transfer are considered for the cross-dataset results. The effect of blur, specular reflections, overlapping leaves and strong illumination gradients have not been explicitly modelled and so the stability of the frequency gating in these conditions is unknown. There was also some variation between diffuse chlorosis and speckle-type disease and in some instances a high frequency gating raised the possibility of dusting or insect damage being confused with the disease texture. Using specific field confounders to identify and report failures by condition would do a better job of characterizing robustness.

The present evaluation does not include the classification of multiple leaves. Detection and segmentation tasks would be used to measure the module's frequency-preserving design, and whether spatial supervision is present or not, it would measure how well the module could localize small lesions. This assessment is vital and should be the next step.

8. Conclusion

This paper proposed a frequency-preserving multi-resolution token-mixing attention network for banana leaf disease detection, where the learnable wavelet-packet band separation, cross-band multi-head token mixing, band-specific frequency-aware gating, and cross-resolution skip recalibration were interleaved with a conventional CNN backbone. The module returns maps of the same spatial and channel dimensions as the host detector, which results in keeping the classifier head, loss function and Grad-CAM pathway without modification, and keeping recalibration scalars that have been initialized to 0, so that the pretrained behavior of the host detector is preserved at the beginning of fine tuning. On a public banana leaf dataset and a field-collected banana leaf dataset, experiments demonstrated robustness in accuracy and macro-F1 gains over the plain, channel-recalibrated, spatial-attention, transformer-hybrid and fixed-wavelet baselines, with the best performance being obtained for early stages of leaf disease and under cross-dataset transfer. The masks, specifically the band-restricted mask and the zero initialization, and the learnable analysis filters were cited as the major gains by the ablations. Grad-CAM heatmaps and embedding projections showed more localized and discriminative representations of the lesions. This extra cost of $\sim$1.2M parameters and $\sim$2 ms of latency is less than transformer-based alternatives, but the first-stage block contributes to the vast majority of the added computation and is deserving of optimizations for edge deployment. For the learned filters, a stage-wise shift from a high frequency emphasis to the low frequency emphasis was apparent and consistent with the morphology of speckle-based and chlorosis-based diseases, but it is a hypothesis that is partially supported. Future work will be directed at extending the framework to the detection and segmentation pipelines, quantifying the relation between gating activations and lesion severity, and characterizing its robustness under blur and specular reflection and non-pathological texture confounders.

References

[1] M. Biruma et al., "Banana Xanthomonas wilt: A review of the disease, management strategies and future research directions," African Journal of Biotechnology, vol. 6, no. 8, pp. 953–962, 2007, doi: 10.5897/AJB2007.000-2116.
[2] L. Li, S. Zhang, and B. Wang, "Plant disease detection and classification by deep learning—A review," IEEE Access, vol. 9, pp. 56683–56698, 2021, doi: 10.1109/ACCESS.2021.3069646.
[3] K. L. Narayanan et al., "Banana plant disease classification using hybrid convolutional neural network," Computational Intelligence and Neuroscience, vol. 2022, Art. no. 9153699, 2022, doi: 10.1155/2022/9153699.
[4] M. A. B. Bhuiyan, H. M. Abdullah, S. E. Arman, S. S. Rahman, and K. Al Mahmud, "BananaSqueezeNet: A very fast, lightweight convolutional neural network for the diagnosis of three prominent banana leaf diseases," Smart Agricultural Technology, vol. 4, Art. no. 100214, 2023, doi: 10.1016/j.atech.2023.100214.
[5] T. Williams and R. Li, "Wavelet pooling for convolutional neural networks," in Proc. Int. Conf. Learn. Representations (ICLR), 2018. [Online]. Available: https://openreview.net/forum?id=rkhlb8lCZ.
[6] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2018, pp. 7132–7141, doi: 10.1109/CVPR.2018.00745.
[7] X. Ma, T. Liang, S. Zhang, S. Huang, and L. He, "Improved LightCNN with attention modules for ASV spoofing detection," in Proc. IEEE Int. Conf. Multimedia Expo (ICME), 2021, pp. 1–6, doi: 10.1109/ICME51207.2021.9428313.
[8] [8] A. Dosovitskiy et al., "An image is worth 16 × 16 words: Transformers for image recognition at scale," in Proc. Int. Conf. Learn. Representations (ICLR), 2021. [https://openreview.net/forum?id=YicbFdNTTy.
[9] Z. Liu et al., "Swin Transformer: Hierarchical vision transformer using shifted windows," in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), 2021, pp. 10012–10022, doi: 10.1109/ICCV48922.2021.00986.
[10] Z. Zhang, Y. Wang, and K. Wang, "Fault diagnosis and prognosis using wavelet packet decomposition, Fourier transform and artificial neural network," Journal of Intelligent Manufacturing, vol. 24, no. 6, pp. 1213–1227, 2013, doi: 10.1007/s10845-012-0657-2.
[11] L. Yi, Y. Wu, A. Tolba, T. Li, S. Ren, and J. Ding, "SA-MLP-Mixer: A compact all-MLP deep neural net architecture for UAV navigation in indoor environments," IEEE Internet of Things Journal, vol. 11, no. 12, pp. 21359–21371, 2024, doi: 10.1109/JIOT.2024.3359662.
[12] S. R. Chandaran, G. Muthusamy, L. R. Sevalaiappan, and N. Senthilkumaran, "Deep learning-based transfer learning model in diagnosis of diseases with brain magnetic resonance imaging," Acta Polytechnica Hungarica, vol. 19, no. 5, pp. 127–147, 2022, doi: 10.12700/APH.19.5.2022.5.7.
[13] B. Yang et al., "A novel plant type, leaf disease and severity identification framework using CNN and transformer with multi-label method," Scientific Reports, vol. 14, Art. no. 11664, 2024, doi: 10.1038/s41598-024-62452-x.
[14] T. Yao, Y. Pan, Y. Li, C.-W. Ngo, and T. Mei, "Wave-ViT: Unifying wavelet and transformers for visual representation learning," in Computer Vision—ECCV 2022, ser. Lecture Notes in Computer Science, vol. 13685. Cham, Switzerland: Springer, 2022, pp. 328–345, doi: 10.1007/978-3-031-19806-9_19.
[15] S. Karimijafarbigloo, S. G. Kolahi, R. Azad, U. Bagci, and D. Merhof, "Frequency-domain refinement of vision transformers for robust medical image segmentation under degradation," in Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV), 2025, pp. 9158–9167, doi: 10.1109/WACV61041.2025.00889.
[16] B. Hu, W. Jiang, J. Zeng, C. Cheng, and L. He, "FOTCA: Hybrid transformer-CNN architecture using AFNO for accurate plant leaf disease image recognition," Frontiers in Plant Science, vol. 14, Art. no. 1231903, 2023, doi: 10.3389/fpls.2023.1231903.
[17] H. H. Alshammari and H. A. Alkhiri, "Optimized recurrent neural network mechanism for olive leaf disease diagnosis based on wavelet transform," Alexandria Engineering Journal, vol. 78, pp. 149–161, 2023, doi: 10.1016/j.aej.2023.07.037.
[18] M. Mahmood and K. O. Alsalem, "Olive leaf disease detection via wavelet transform and feature fusion of pre-trained deep learning models," Computers, Materials & Continua, vol. 78, no. 3, pp. 3431–3448, 2024, doi: 10.32604/cmc.2024.047604.
[19] S. Abisha, A. M. Mutawa, M. Murugappan, and S. Krishnan, "Brinjal leaf diseases detection based on discrete Shearlet transform and deep convolutional neural network," PLOS ONE, vol. 18, no. 4, Art. no. e0284021, 2023, doi: 10.1371/journal.pone.0284021.
[20] I. Daubechies, Ten Lectures on Wavelets. Philadelphia, PA, USA: SIAM, 1992, doi: 10.1137/1.9781611970104.

[21] N. Mduma et al., "Dataset of banana leaves and stem images for object detection, classification and segmentation: A case of Tanzania," Data in Brief, vol. 49, Art. no. 109322, 2023, doi: 10.1016/j.dib.2023.109322.
[22] M. Shafiq and Z. Gu, "Deep residual learning for image recognition: A survey," Applied Sciences, vol. 12, no. 18, Art. no. 8972, 2022, doi: 10.3390/app12188972.
[23] R. R. Selvaraju et al., "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), 2017, pp. 618–626, doi: 10.1109/ICCV.2017.74.