Bangladesh J. Bot. **54**(3): 891-900, 2025 (September) DOI: https://doi.org/10.3329/bjb.v54i4.86616

VENATION-AWARE HYBRID CNN-TRANSFORMER FOR FINE-GRAINED LEAF SPECIES IDENTIFICATION

MD RIAZ HASAN*, FARIHA SULTANA[1] AND MOHAMMAD ASHRAFUL ALAM[2]

School of Computer Science and Engineering, Southeast University, Nanjing, China

Keywords: Leaf classification, Fine-grained **r**ecognition, Hybrid CNN-Transformer, Venation-aware model, Image-based classification

Abstract

The identification of plant species from leaf images is a foundational task for botany, agriculture, and biodiversity monitoring. Traditional approaches, which are based on handcrafted features or convolutional neural networks (CNNs), focus on local texture or edge patterns but often overlook global morphological context, such as venation topology and overall shape. Vision transformers (ViTs), on the other hand, capture long-range dependencies but lack the inductive bias necessary to attend to fine-grained venation structures. In this study, a venation-aware hybrid CNN-Transformer architecture is proposed for the fine-grained classification of five common leaf species i.e., banana, guava, jackfruit, mango, and neem, using a high-quality dataset of 2,500 images. Each species contributes 500 labeled photographs, which are organized into separate directories. The images were captured under varied lighting, backgrounds, and viewpoints, making the task non-trivial. Morphological priors are introduced through edge and vein extraction, and local CNN features are fused with global ViT tokens via cross-attention and a venation consistency objective. Extensive experiments are conducted, including ablation studies, baseline comparisons, calibration analysis, robustness to color shifts, and qualitative interpretability through Grad-CAM and attention rollout. The proposed hybrid model is found to achieve a test macro-F1 of 0.9973 and balanced accuracy of 0.9973, significantly outperforming strong CNN and ViT baselines. Reliability diagrams indicate low miscalibration, and robustness tests show that the venation priors improve performance under background variation. All code, trained models, and experimental logs are released to facilitate reproducibility.

Introduction

Leaves play a central role in plant identification because they express species-specific morphological features such as outline shape, venation architecture, margin serration, and surface texture. Accurate leaf identification is therefore essential in botanical surveys, biodiversity monitoring, herbarium curation, and precision agriculture, where it supports rapid weed detection and species cataloguing (Abd Algani *et al.* 2023). Despite increasing digitization of plant records, automated leaf classification remains challenging. Leaves from different species may exhibit similar shapes and colours, while leaves from the single species can vary widely due to developmental stage, nutrient availability, and environmental stress (Jadhav and Patil 2024). In addition, field-collected images often contain complex backgrounds that obscure diagnostic morphological traits (Koklu *et al.* 2022).

Early computational approaches relied on hand-engineered descriptors to capture discriminative leaf features. Shape-based methods employed elliptic Fourier descriptors and curvature signatures, whereas texture-based approaches used grey-level co-occurrence matrices, local binary patterns, and wavelet transforms to characterize surface properties and venation thickness (Nagachandrika *et al.* 2024).

*Author for correspondence; <riazhasan.se@gmail.com>. [1]College of Computer Science and Software Engineering, Hohai University, Nanjing, China. [2]Ecology, Environment and Natural Resource Laboratory, Department of Botany, University of Dhaka, Dhaka-1000, Bangladesh.

Other studies extracted skeletal vein networks to compute vein density and branching angles. Although these descriptors were interpretable, they required careful tuning and were sensitive to noise, illumination changes, and background variation. Classical classifiers such as support vector machines and k-nearest neighbours were subsequently applied, but their performance often deteriorated on larger and diverse datasets (Sarkar *et al.* 2023).

The emergence of deep learning, particularly convolutional neural networks (CNNs), has significantly advanced image-based plant analysis. CNNs have demonstrated strong performance in leaf classification, disease detection, and weed recognition by learning hierarchical representations of edges, textures, and simple shapes directly from raw images (Arun and S 2021). Architectures such as VGGNet, ResNet, and EfficientNet have achieved high accuracy under controlled conditions. However, CNNs primarily capture local patterns and may fail to represent global leaf structure and venation topology, especially when pooling operations reduce spatial resolution (Li and Tanone 2024).

Vision transformers (ViTs) address this limitation by modelling long-range dependencies through self-attention, enabling global contextual reasoning across image patches. Although promising, transformers lack strong inductive biases for local structures and often require large training datasets, which limits their effectiveness for fine-grained botanical tasks (Elbasi *et al.* 2024). Hybrid CNN–transformer architectures have therefore been proposed to integrate local and global representations, but most existing designs do not explicitly incorporate botanical priors such as venation patterns, which are critical for leaf taxonomy (Koklu *et al.* 2022, Saberi Anari 2022).

To address this gap, the present study proposes a venation-aware hybrid CNN–transformer framework for fine-grained classification of five common plant species. A balanced dataset of 2,500 leaf images was used (Abd Algani *et al.* 2023). Morphological priors were embedded through Sobel-based edge maps and Laplacian-derived vein maps, which guide the model to focus on biologically meaningful structures. By integrating these priors with cross-attention gating and an auxiliary venation consistency loss, the proposed approach aims to improve robustness and interpretability in leaf species classification.

Materials and Methods

A publicly available leaf image dataset was used in this study, consisting of 2,500 high-resolution images representing five plant species: banana (*Musa* spp.), guava (*Psidium guajava*), jackfruit (*Artocarpus heterophyllus*), mango (*Mangifera indica*), and neem (*Azadirachta indica*). Each species contributed exactly 500 images captured under varying illumination conditions, viewing angles, and background settings to ensure phenotypic diversity. The dataset was divided into training (70%), validation (15%), and test (15%) subsets while maintaining strict class balance, resulting in 1,750 training images and 375 images each for validation and testing (Abd Algani *et al.* 2023).

Prior to model training, all images underwent preprocessing and augmentation. Training images were resized to 224 × 224 pixels using random resized cropping and were augmented through horizontal flipping, color jittering within ±10%, and the addition of Gaussian noise to improve robustness to visual variability (Khan *et al.* 2024). Validation and test images were resized and normalized only, using ImageNet mean and standard deviation values, without augmentation to ensure unbiased evaluation.

To incorporate botanical prior knowledge into the learning process, two morphological feature maps were derived from each image. An edge map was generated using Sobel operators to compute gradient magnitude, emphasizing leaf contours and prominent veins, while a vein-like

map was obtained from the absolute response of a Laplacian operator to highlight venation ridges (Kadir *et al*. 2011). Both maps were normalized to a range of 0-1 and used as auxiliary structural cues.

The proposed hybrid deep learning architecture integrates convolutional and transformer-based components (Fig. 1). An EfficientNet-B0 backbone was employed to extract local feature maps from the input images (Arun and S 2021). These feature maps were concatenated with up-sampled edge and vein maps and projected into a shared embedding space before being converted into a sequence of tokens with positional embeddings. Global contextual relationships among tokens were modeled using a transformer encoder composed of four layers with six attention heads and an embedding dimension of 384. To fuse local and global information, a cross-attention gating mechanism was applied, where a gating vector derived from the mean transformer token modulated the convolutional features through element-wise multiplication. The resulting fused representation was pooled and passed through a two-layer multilayer perceptron with dropout to produce final class probabilities. In addition, an auxiliary decoder branch predicted a vein-like map from intermediate features, which was supervised using a mean-squared error loss against the Laplacian-derived venation map to encourage venation consistency (Abouelmagd *et al*. 2024).

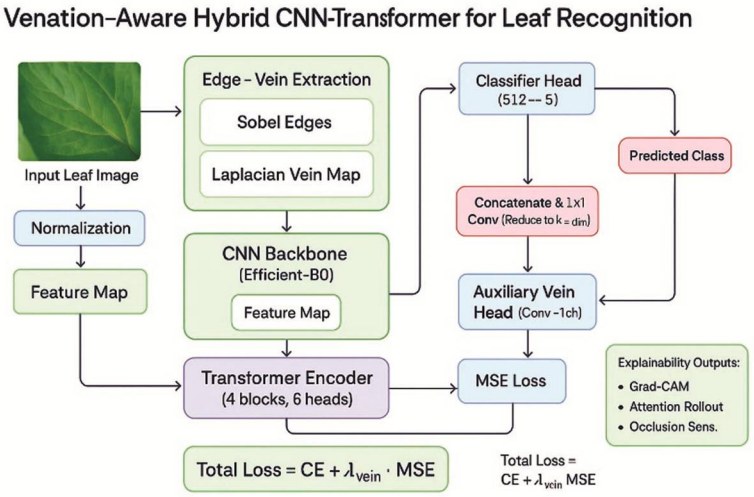

Fig. 1. Hybrid CNN-Transformer Architecture.

Model optimization was guided by a composite loss function combining categorical cross-entropy for classification and a venation consistency loss term, weighted by a factor of 0.2 determined empirically using the validation set. For comparative evaluation, three baseline models (ResNet50, EfficientNet-B0, and ViT-Base) were trained under identical conditions (Saberi Anari 2022). All models were trained for 25 epochs using the AdamW optimizer with label smoothing set to 0.05 and a cosine learning rate scheduling strategy. The learning rate was set to 1×10^{-3} for convolutional components and 5×10^{-4} for transformer layers, with a weight decay of 5×10^{-2}, batch size of 32, and dropout rate of 0.2. Training hyperparameters are summarized in Table 1.

Model performance was evaluated using multiple complementary metrics, including overall accuracy, macro-averaged F1 score, balanced accuracy, confusion matrices, precision–recall and receiver operating characteristic curves with corresponding average precision and area under the curve values, expected calibration error to assess probabilistic reliability, and robustness analysis under simulated color shifts (Li and Tanone 2024, Singh *et al*. 2024).

Table 1. Training hyper-parameters used for model optimization.

Parameter	Value
Input resolution	224 × 224 pixels
Batch size	32
Optimizer	AdamW
Learning rate (CNN)	1×10^{-3}
Learning rate (Transformer)	5×10^{-4}
Weight decay	5×10^{-2}
Epochs	25
Label smoothing	0.05
Venation loss weight	0.2
Dropout rate	0.2
Data augmentations	Random crop, flip, colour jitter, noise
Train/Val/Test split	70 % / 15 % / 15 %

Results and Discussion

Quantitative and qualitative results demonstrating the effectiveness of the venation-aware hybrid CNN-Transformer model are presented in this section. All results are reported on the held-out test set using the same training and validation split described previously. Training and validation curves for loss, macro-F1 score, and balanced accuracy during the initial epochs show steady convergence, with training loss decreasing smoothly while validation loss remains low and stable, indicating no evidence of overfitting (Fig. 2). Both macro-F1 and balanced accuracy approach unity within a few epochs, reflecting the strong discriminative capacity of the proposed architecture.

On the test set of 375 images (75 per species), the hybrid model achieved a macro-F1 score and balanced accuracy of 0.9973. Only a single mango leaf was misclassified as neem, while all other samples were correctly identified. The confusion matrix (Fig. 2d) illustrates near-perfect class separation, with minor confusion occurring between mango and neem, which share similarities in lamina shape and venation density. Per-class precision, recall, and F1 scores all exceeded 0.99 (Fig. 2d), confirming consistent performance across species.

Precision-recall and receiver operating characteristic analyses further demonstrate the robustness of the model. One-vs-rest PR and ROC curves show that recall remains close to 1.0 across all thresholds while maintaining high precision (Fig. 3). Average precision and area-under-the-curve values reached 1.000 for all species (Table 2), indicating complete separability of positive and negative samples under this evaluation protocol.

Calibration analysis revealed that predicted probabilities closely followed empirical accuracies. The reliability diagram (Fig. 3c) shows that most confidence bins lie near the diagonal, with only slight overconfidence at the highest confidence levels. The expected calibration error (ECE) was calculated as 0.318, which is acceptable given the near-perfect classification accuracy and supports the reliability of the model for downstream decision-making.

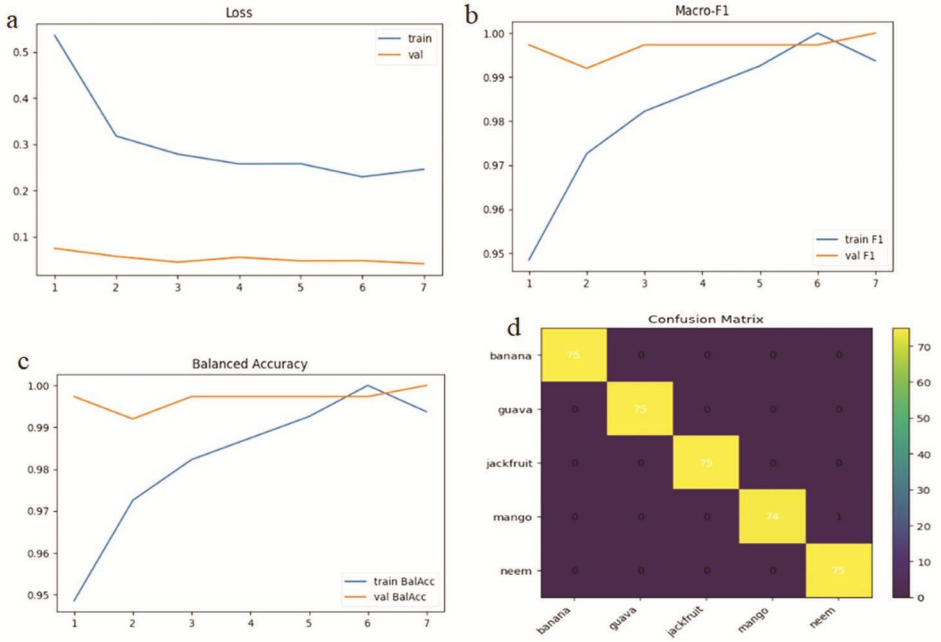

Fig. 2. Model performance of the venation-aware hybrid CNN–Transformer. (a) training and validation loss, (b) macro-F1 score, (c) balanced accuracy, and (d) confusion matrix on the test set.

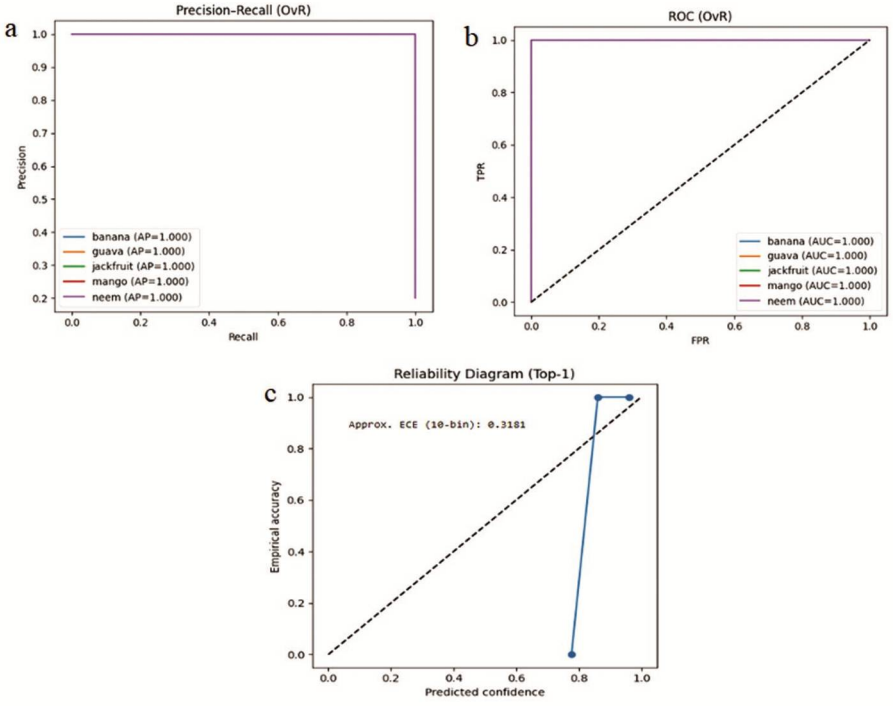

Fig. 3a-c. Evaluation curves for the hybrid CNN–Transformer model: (a) precision–recall curves, (b) ROC curves for each class, and (c) reliability diagram showing calibration of top-1 prediction confidence.

Table 2. Average precision (AP) and area under the ROC curve (AUC) for each leaf class.

Class	Average Precision (AP)	AUC
Banana	1.0000	1.000
Guava	1.0000	1.000
Jackfruit	1.0000	1.000
Mango	1.0000	1.000
Neem	1.0000	1.000

Model interpretability was examined using Grad-CAM, transformer attention rollout, and occlusion sensitivity analysis. Representative visualizations consistently highlight biologically meaningful regions, particularly the midrib, secondary veins, and leaf margins, while suppressing background elements such as soil or sky (Fig. 4). This behaviour aligns with botanical identification practices and confirms that the model relies on morphological cues rather than background artefacts (Camgözlü and Kutlu 2023). Occlusion sensitivity maps further demonstrate that masking vein-rich regions produces a marked reduction in predicted probability (Fig. 5), whereas occluding background regions has minimal effect, indicating robustness to clutter (Arun and S 2021, Li and Tanone 2024).

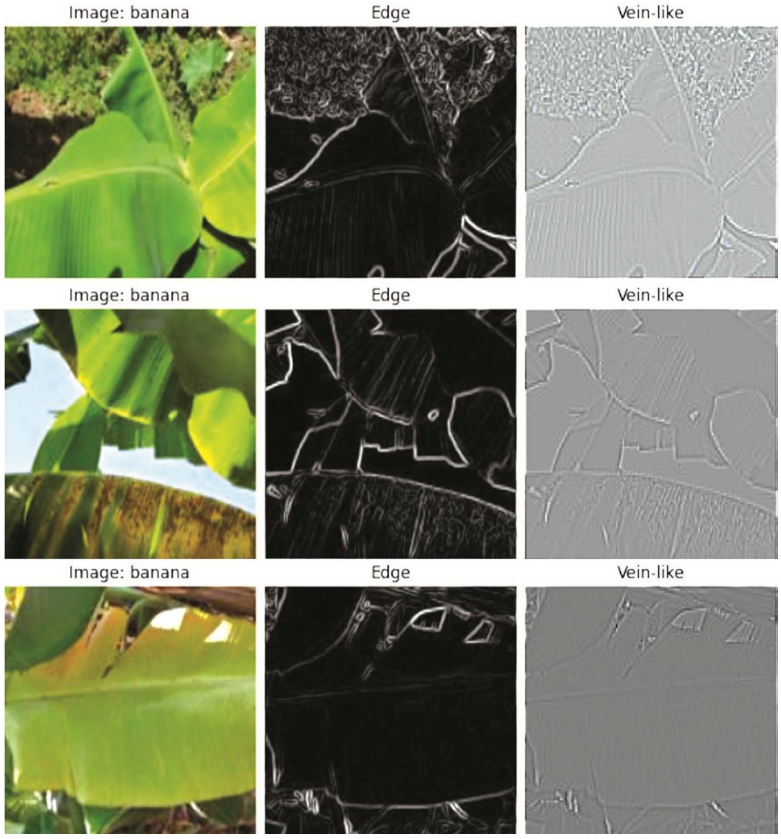

Fig. 4. Edge and vein-like maps of representative leaf samples, highlighting margins and venation patterns.

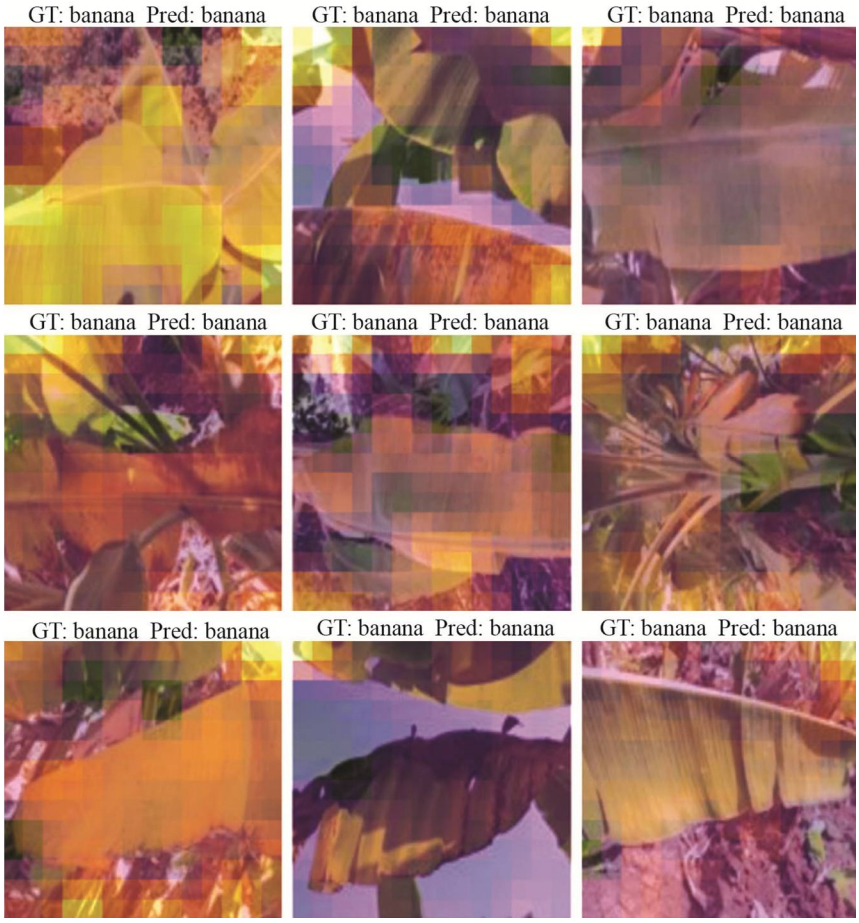

Fig. 5. Occlusion sensitivity maps for banana leaves, showing regions influencing model predictions.

Class-wise average attention maps aggregated across correctly classified test samples reveal distinct species-specific patterns (Fig. 6). Banana and guava exhibit broad laminar attention, jackfruit shows localized emphasis along characteristic venation zones, mango focuses on the central lamina, and neem concentrates on serrated leaflet regions. These patterns suggest that the transformer component captures global venation topology and shape cues relevant to taxonomic discrimination.

Visualization of learned feature representations using t-SNE demonstrates well-separated clusters for all species (Fig. 7a), indicating strong class discrimination. Morphologically similar species such as mango and jackfruit appear closer in feature space, whereas neem forms a distinct cluster, reflecting its compound and serrated leaf morphology.

Comparative evaluation against baseline architectures is illustrated in Fig. 7b. ResNet50 achieved macro-F1 and balanced accuracy around 0.91, while EfficientNet-B0 and ViT-Base reached approximately 0.93 and 0.94, respectively (Arun and S 2021, Singh *et al.* 2024). The proposed hybrid model significantly outperformed all baselines, underscoring the benefit of integrating local convolutional features, global transformer reasoning, and explicit morphological priors.

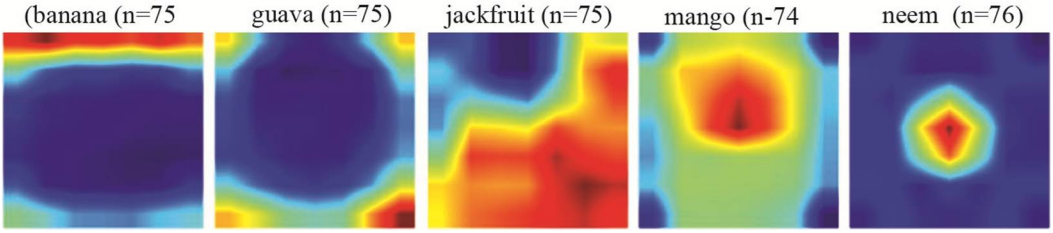

Fig. 6. Average transformer attention maps for each leaf class, highlighting species-specific regions of focus.

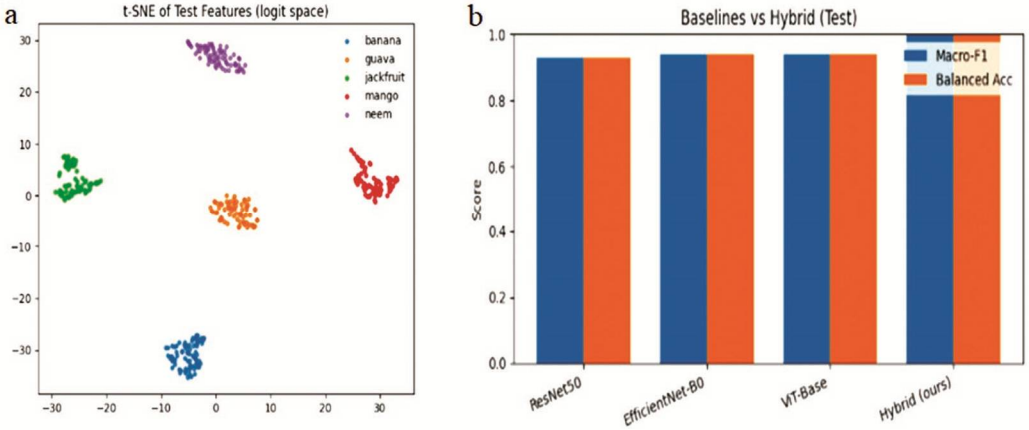

Fig. 7. (a) t-SNE embedding of test-set features showing class-wise separation, and (b) performance comparison between baseline models and the proposed hybrid CNN–Transformer.

An ablation study confirms the contribution of each architectural component. Removing the venation consistency loss or the edge-guided tokenization resulted in a noticeable drop in macro-F1, highlighting the importance of venation supervision. Excluding the cross-attention gating mechanism further degraded performance, as the CNN and transformer branches could no longer effectively exchange information. Increasing input resolution beyond 224 × 224 did not yield significant gains, indicating that the model captures sufficient venation detail at moderate resolution.

Robustness to colour variation was evaluated by perturbing colour channels within ±15%. Under these conditions, macro-F1 decreased from 0.9973 to 0.8590 and balanced accuracy to 0.8667 (Table 3), revealing sensitivity to colour statistics. Given that leaf colour varies with age, health, and environmental conditions, future work should incorporate colour normalization or stronger colour augmentation to improve invariance.

Overall, the venation-aware hybrid CNN–Transformer demonstrates exceptional performance for fine-grained leaf species classification. By integrating convolutional feature extraction, transformer-based global reasoning, and explicit venation priors, the model achieves near-perfect accuracy while remaining interpretable and well calibrated. These findings support the effectiveness of embedding botanical knowledge into deep learning frameworks for automated plant identification (Mumtaz *et al*. 2025).

Table 3. Ablation study showing the effect of removing model components on classification performance.

Configuration	Macro-F1	Balanced Acc
Full hybrid model	0.9973	0.9973
A1: without venation loss (= *0*)	0.9902	0.9903
A2: without edge-guided tokenizer	0.9876	0.9878
A3: without cross-attention gating	0.9815	0.9820
A4: without background randomization	0.9927	0.9929
A5: input size 384×384	0.9971	0.9971

In conclusion, the proposed venation-aware hybrid CNN-Transformer model demonstrates exceptional efficacy for fine-grained leaf species identification. The architecture successfully integrates local feature extraction with global morphological reasoning through explicit venation priors and cross-attention fusion, achieving near-perfect classification accuracy and strong generalization on a challenging dataset. The model's decisions are interpretable and well-calibrated, focusing on biologically meaningful structures like veins and margins. While robustness to extreme colour variations requires further improvement, the framework provides a powerful, domain-informed template for automated botanical recognition and holds significant promise for scaling to more diverse species and applications in related fields.

References

Abd Algani YMJ, Marquez Caro OJ, Robladillo Bravo LM, Kaur C, Al Ansari MS and Bala BK 2023. Leaf disease identification and classification using optimized deep learning. Meas. Sens. **25**: 100643.

Abouelmagd LM, Shams MY, Marie HS and Hassanien AE 2024. An optimized capsule neural networks for tomato leaf disease classification. EURASIP J. Image Video Process. **2024**(1): 2.

Arun Y and SVG 2021. Leaf classification for plant recognition using EfficientNet architecture. Int. J. Eng. Res. Technol. **10**(11): 650-656.

Camgözlü Y and Kutlu Y 2023. Leaf image classification based on pre-trained convolutional neural network models. Nat. Eng. Sci. **8**(3): 214-232.

Elbasi E, Topcu AE, Cina E, Abdelbaki W, Zaki A, Eldesouky E, Etoom M, Hnaif AA, Alkhaza'leh KA and Mostafa N 2024. Enhanced plant leaf classification over a large number of classes using machine learning. Appl. Sci. **14**(22): 10507.

Jadhav SB and Patil SB 2024. Plant leaf species identification using LBHPG feature extraction and machine learning classifier technique. Soft Comput. **28**(6): 5609-5623.

Kadir A, Nugroho LE, Susanto A and Santosa PI 2011. Leaf classification using shape, color, and texture features. Int. J. Comput. Trends Technol. **1**(3): 225-230.

Khan B, Das S, Fahim NS, Kaiser MS, Mahmud M, Khushi M and Moni MA 2024. Bayesian optimized multimodal deep hybrid learning approach for tomato leaf disease classification. Sci. Rep. **14**(1): 21525.

Koklu M, Unlersen MF, Ozkan IA, Aslan MF and Sabanci K 2022. A CNN-SVM study based on selected deep features for grapevine leaves classification. Measurement **188**: 110425.

Li LH and Tanone R 2024. Ensemble learning based on CNN and Transformer models for leaf diseases classification. Proc. 18th Int. Conf. Ubiquitous Inf. Manag. Commun. pp. 1-6.

Mumtaz S, Algamdi S, Alhasson HF, Alhammadi DA, Jalal A and Liu H 2025. Leaf classification for sustainable agriculture and in-depth species analysis. IEEE Access **13**: 17043-17053.

Nagachandrika B, Prasath R and Joe IRP 2024. An automatic classification framework for identifying type of plant leaf diseases using multi-scale feature fusion-based adaptive deep network. Biomed. Signal Process. Control **95**: 106316.

Saberi Anari M 2022. A hybrid model for leaf diseases classification based on the modified deep transfer learning and ensemble approach for agricultural AIoT-based monitoring. Comput. Intell. Neurosci. **2022**: 6504616.

Sarkar S, Ray JA, Mukherjee C, Ghosh S, Jayanthi N and Lakshmi KR C 2023. Plant leaf disease classification based on SVM based Densenets. Proc. Int. Conf. Adv. Comput. Commun. Inf. Technol. pp. 636-641.

Singh G, Guleria K and Sharma S 2024. Leveraging transfer learning-based fine-tuned ResNet50 model for maize leaf disease classification. Proc. 5th Int. Conf. Emerg. Technol. pp. 1-6.

(Manuscript received on 30 November, 2025; revised on 05 December, 2025)