



Implementing Medicinal Leaf Species Classification Using Hybrid Preprocessing, CNN Benchmarking, Squeeze-and-Excitation Attention

Amruta K. Jadhav^{1,*}, P. C. Shetiye²

¹Department of Computer Science and Engineering, School of Engineering and Technology (SOET), MGM University, Chhatrapati Sambhajanagar (M.S.) – 431003, India

²Department of MCA, Government College of Engineering, Chhatrapati Sambhajanagar (M.S.)–431005, India

Abstract

Accurate medicinal-plant identification is important for botanical research and herbal quality assurance, yet visual similarity among species makes manual identification difficult. This study evaluates an auditable medicinal-leaf classification workflow rather than proposing a new transfer-learning or attention architecture. The workflow combines file-integrity auditing, fixed stratified partitions, controlled augmentation, three model configurations, held-out evaluation, persistent artifacts, and qualitative Grad-CAM inspection. The controlled dataset contains 1,009 decodable images from nine species, partitioned into 706 training, 151 validation, and 152 held-out test images. SHA-256 comparison found no byte-identical duplicates; perceptual or specimen-level near-duplicate screening was not performed and is therefore a limitation. Under the same three-epoch CPU pilot protocol, ImageNet-initialized MobileNetV3-Small achieved 98.68% test accuracy and 98.68% weighted F1-score, compared with 71.71% for a randomly initialized baseline CNN and 62.50% for a segmentation-plus-CLAHE SE-CNN. Exact paired analysis showed that MobileNetV3-Small outperformed the baseline (McNemar $p < 0.001$); the advantage over the hybrid model also remained significant under every error-overlap pattern compatible with the reported confusion matrices. Class-wise metrics exposed severe hybrid-model failures for Coriander and Betel. Because the component changes were confounded, training was limited to three epochs, Grad-CAM was inspected only qualitatively, and no external dataset was used, the results support MobileNetV3-Small only as the strongest model for this controlled split and do not establish field generalization or superiority of the hybrid branch.

Keywords: medicinal plant identification; leaf classification; convolutional neural network; transfer learning; squeeze-and-excitation; Grad-CAM; Streamlit; reproducibility

1. Introduction

Medicinal plants contribute to traditional health systems, pharmaceutical discovery, and local biodiversity knowledge, but correct species-level identification remains a demanding visual task. Leaves may share global shape while differing only in venation, margin, texture, reflectance, or subtle color patterns. Inexperienced users can therefore confuse therapeutically useful plants with visually similar species, while conventional taxonomic inspection requires specialist knowledge and may not scale to field or educational settings. Contemporary computer vision provides a practical complement: convolutional neural networks (CNNs) learn hierarchical representations directly from images and can reduce dependence on manually engineered texture descriptors. Recent medicinal-leaf studies report high performance, including global-average-pooling CNN results near 99% on five species [1] and DenseNet-based performance up to 99.64% on a 30-species collection [2]. These outcomes establish feasibility, but they also make evaluation discipline increasingly important because accuracy can be inflated by controlled backgrounds, correlated photographs, augmentation before splitting, or reuse of the same acquisition session across partitions.

Transfer learning is attractive when the domain dataset is modest. A network initialized on ImageNet can reuse low-level edge, color, and texture filters and then adapt its higher-level representation to medicinal leaves. Multi-task and feature-fusion systems extend this idea by combining complementary local and global cues; MTJNet, for example, reported 99.71% accuracy on an Indian medicinal-plant dataset [3]. Similarly, ensembles of CNNs and vision transformers have produced near-perfect results on established leaf benchmarks [4]. Such headline scores are useful reference points, but they are not directly comparable across species counts, image sources, train/test protocols, or metrics. A scientifically useful implementation must therefore connect each result to an immutable split and expose the full prediction record rather than present accuracy alone.

This work also investigates channel attention. Squeeze-and-Excitation (SE) blocks recalibrate feature channels by compressing spatial information through global average pooling and learning channel-specific gates [5]. In principle, SE attention can emphasize useful venation or texture responses and suppress background-related channels. Explainability is addressed with gradient-weighted class activation mapping (Grad-CAM), which projects class gradients onto the final convolutional feature maps to show image

regions that influenced a prediction [6]. For laptop deployment, MobileNetV3 provides a resource-conscious alternative designed through hardware-aware search and efficient architectural components [7]. The study asks four bounded questions: (i) can a fixed, auditable split support an inspectable medicinal-leaf benchmark; (ii) how do a compact CNN, ImageNet-initialized MobileNetV3-Small, and a segmentation-plus-CLAHE SE-CNN behave under the same short CPU protocol; (iii) can checkpoints and prediction artifacts be reused after application restart; and (iv) do illustrative Grad-CAM maps focus on plausible leaf regions? The novelty is practical and experimental, not architectural: transfer learning, SE attention, CLAHE, and Grad-CAM are established methods. The contribution is their integration into a traceable workbench that links dataset audit, split membership, validation-based checkpointing, prediction-level exports, confidence intervals, persistent inference, and explicit negative results.

- a traceable nine-species workflow that records decoding results, SHA-256 file hashes, class profiles, fixed split membership, configurations, checkpoints, and predictions;
- a controlled pilot comparison showing that a standard pretrained MobileNetV3-Small substantially outperforms two randomly initialized CNN configurations under a three-epoch CPU budget;
- class-wise error analysis and paired statistical comparisons that complement aggregate accuracy;
- a persistent checkpoint registry and Streamlit interface for repeatable inference and artifact inspection; and
- an explicit account of what was not established: near-duplicate independence, component-level ablation, convergence of randomly initialized models, systematic explanation validity, and external generalization.

2. Related Work

2.1 Handcrafted descriptors and convolutional learning

Earlier plant-recognition pipelines typically segmented the leaf and derived shape, color, and texture descriptors before applying a conventional classifier. Gray-level co-occurrence matrix (GLCM) descriptors remain interpretable and inexpensive, but their fixed statistics are sensitive to illumination, scale, and background variation. Deep CNNs replace this manually selected representation with learned multi-scale filters. The global-average-pooling model evaluated by Azadnia and colleagues reported 99.66%, 99.32%, and 99.45% accuracy at 64, 128, and 256-pixel resolutions, respectively, for five medicinal species [1]. Dey and colleagues evaluated seven deep networks on 5,878 leaf images from 30 medicinal species; DenseNet201 provided the strongest result at 99.64% accuracy, while field photographs remained harder than controlled images [2].

2.2 Transfer learning, fusion, and attention

Transfer learning reduces the optimization burden for moderate datasets. In an Ethiopian dataset of 1,853 images from 35 expert-labeled species, fine-tuned VGG19 achieved 94% test accuracy, outperforming VGG16, InceptionV3, and Xception [8]. MTJNet fused local and global features after vein and Sobel-based preprocessing and reported 99.71% accuracy, 99.60% precision, 99.62% recall, and 99.58% F1-score [3]. Plant-CNN-ViT concatenated representations from ViT, ResNet50, DenseNet201, and Xception and obtained 100%, 100%, 100%, and 99.83% on Flavia, Folio, Swedish Leaf, and MalayaKew, respectively [4]. These results illustrate the strength of fusion but also reinforce the need to state the dataset and protocol beside every score.

Attention modules explicitly weight salient features. SE networks model inter-channel dependencies through global pooling and a bottleneck gating function [5]. Their small computational overhead makes them suitable for leaf classification, where texture and venation channels may carry different discriminative value. For mobile or laptop inference, MobileNetV3 combines depthwise separable convolution, squeeze-and-excitation components, and hardware-aware architecture search [7]. Grad-CAM complements these models by producing class-discriminative localization without changing network training [6].

Recent reviews indicate that deep learning has become a dominant approach for medicinal-plant species recognition, with leaf images representing the most frequently analyzed plant organ. Mulugeta et al. reported that 96.7% of the reviewed studies used leaf images and that transfer learning was adopted in 83.8% of the investigated studies, highlighting the importance of pretrained representations for medicinal-plant classification [9]. Similarly, Yao et al. reviewed deep-learning approaches for plant identification and leaf-image analysis and emphasized the increasing use of CNN-based models, multi-output learning, and multi-task learning for plant-related classification problems [10]. These findings support the use of transfer learning and standardized evaluation protocols when datasets are relatively small and computational resources are constrained.

Recent research has also explored hybrid architectures and preprocessing strategies to improve the representation of visually similar medicinal leaves. Kayaalp evaluated several pretrained deep-learning models together with image-processing and edge-detection strategies for medicinal plant leaf classification, reporting 98.69% test accuracy with DenseNet121 before edge-based processing, while the performance decreased substantially after edge detection [12]. This observation is relevant to the present study because it demonstrates that additional preprocessing does not necessarily improve classification performance and that transformations should be empirically validated rather than assumed to be

beneficial. Pushpa et al. similarly investigated hybrid combinations of deep CNN architectures for medicinal plant identification, demonstrating the potential of feature fusion while also highlighting the computational and generalization considerations associated with hybrid approaches [13].

Lightweight architectures have become particularly relevant for practical medicinal-plant recognition because many applications are intended for mobile, laptop, or resource-constrained environments. Widneh and Alemu investigated depth-wise separable convolution for medicinal plant-part identification using a large collection of leaf images and compared the approach with conventional pretrained architectures [14]. More recently, Gautam et al. proposed a feature-fusion framework combining handcrafted and deep features for medicinal-leaf classification and reported 98.90% test accuracy on their dataset [15]. These studies demonstrate the continuing interest in computationally efficient architectures and feature-fusion strategies, while also reinforcing the importance of evaluating models under clearly defined datasets and experimental protocols.

Alternative approaches have also investigated conventional feature engineering and optimization as computationally efficient alternatives to fully deep architectures. Sachar and Kumar proposed a medicinal-plant identification framework combining leaf segmentation, BRISK feature extraction, Levy-flight-based Dragonfly feature selection, and machine-learning classification. Their experiments achieved 92.0% accuracy on a custom medicinal-leaf dataset, illustrating that carefully engineered feature-selection pipelines remain relevant when computational efficiency is prioritized [16]. Such approaches provide useful complementary baselines against which lightweight deep-learning systems can be evaluated.

Recent plant-recognition research has increasingly moved beyond conventional CNN architectures toward transformer-based and self-supervised representations. Ong et al. evaluated the pretrained DINOv2 self-distillation model for cross-domain plant species identification and demonstrated its potential for transferring knowledge between herbarium and field imagery, addressing the scarcity of labeled field samples [17]. In another direction, Lapkovskis et al. investigated automatic multimodal fusion using images from leaves, flowers, fruits, and stems, achieving 82.61% accuracy across 979 plant classes and demonstrating the potential of combining complementary plant organs for species recognition [18]. These developments suggest that future medicinal-plant recognition systems may benefit from self-supervised learning, multimodal information, and stronger cross-domain evaluation rather than relying exclusively on controlled leaf-image datasets.

A recent comprehensive review by Zulfiqar et al. further describes the progression of automated plant identification from conventional supervised CNNs toward Vision Transformers, self-supervised learning, and few-shot learning [19]. This evolution is particularly relevant to medicinal-plant recognition because the availability of labeled images is often limited and performance obtained on controlled datasets may not represent real-world field conditions. The present study therefore focuses on a complementary problem: establishing a reproducible and leakage-aware benchmark using a compact pretrained CNN, while explicitly reporting model failures, confidence intervals, and explainability rather than optimizing solely for a high accuracy value [19].

2.3 Summary of representative work

Study	Year	Dataset	Technique	Reported outcome	Relevance / caveat
Azadnia et al. [1]	2022	5 species	Deep CNN + GAP	99.66% (64 px); 99.32% (128 px); 99.45% (256 px)	Controlled, small species set
Dey et al. [2]	2024	5,878 images; 30 species	Seven DCNNs; DenseNet201 best	99.64% accuracy	Field images were more difficult
Lee et al. [4]	2023	Four public leaf datasets	CNN-ViT ensemble	99.83–100%	Large ensemble; cross-study comparison limited
Sharma & Vardhan [3]	2024	Indian medicinal plant + leaf data	MTJNet local/global fusion	99.71% accuracy; 99.58% F1	Multi-task model; different task composition
Kiflie et al. [8]	2024	1,853 images; 35 species	Fine-tuned VGG family	VGG19: 94% test accuracy	Expert labels; modest dataset
Hu et al. [5]	2018	ImageNet and vision benchmarks	Squeeze-and-Excitation	Channel recalibration improved CNNs	Architecture method, not medicinal dataset
Selvaraju et al. [6]	2017	Classification/VQA	Grad-CAM	Class-discriminative localization	Explanation, not an accuracy model

Mulugeta et al. [9]	2024	Systematic review; 31 studies	Deep-learning medicinal-plant classification review	Leaf images used in 96.7%; transfer learning in 83.8%	Identifies dataset availability and trustworthiness as major research gaps
Yao et al. [10]	2024	Multiple plant/leaf datasets	Multi-prediction deep learning; CNNs	Comprehensive comparison of plant identification and disease-classification approaches	Review and empirical comparison; not restricted to medicinal plants
Kayaalp [12]	2024	10 plant-leaf categories	Pretrained CNNs + edge detection	DenseNet121: 98.69% before edge processing; 67.92% after edge processing	Demonstrates that additional preprocessing may reduce performance
Pushpa et al. [13]	2025	Medicinal/plant leaf images	Hybrid VGG16–MobileNet and MobileNet–ResNet50 models	85.85% and 88% reported for hybrid models	Hybrid architectures; performance depends on dataset and model configuration
Widneh & Alemu [14]	2025	15,100 medicinal-leaf images	Depth-wise separable CNN + pretrained CNN comparison	Evaluated lightweight DWS-CNN against VGG16, ResNet50 and InceptionV3	Relevant to computationally efficient medicinal-plant recognition
Gautam et al. [15]	2025	6,904 images; 80 medicinal-leaf types	Feature-fusion CNN with handcrafted + deep features	98.90% test accuracy	Lightweight feature-fusion approach; dataset/protocol differs from current study
Sachar & Kumar [16]	2025	PlantVillage + custom medicinal-leaf dataset	BRISK + Dragonfly optimization + MLNN	88.6% PlantVillage; 92.0% custom medicinal dataset	Computationally efficient alternative; relies on handcrafted features
Ong et al. [17]	2025	Plant species / herbarium and field imagery	DINOv2 self-supervised transformer	17.7% improvement over compared state-of-the-art model	Strong relevance to cross-domain generalization and limited labeled data
Lapkovskis et al. [18]	2025	Multimodal-PlantCLEF; 979 classes	Automatic multimodal deep-learning fusion	82.61% accuracy	Uses leaves, flowers, fruits and stems; broader plant-identification task
Zulfiqar et al. [19]	2025	Multiple plant-identification datasets	Comprehensive deep-learning review	Reviews CNN, ViT, self-supervised and few-shot approaches	Highlights emerging directions and need for stronger generalization
Arun [20]	2025	Plant-species image dataset	CNN with transfer learning	Plant-species classification using transfer learning	Supports transfer learning as a practical strategy for plant recognition

3. System Development and Proposed Methodology

3.1 System overview

The proposed system is an end-to-end research workbench rather than a single classifier. Each interaction is routed through a central pipeline that creates a timestamped experiment directory and stores the configuration, dataset audit, image index, stratified splits, checkpoints, learning histories, prediction records, summary metrics, and figures. The application can discover registered checkpoints after restart; classification and Grad-CAM therefore do not depend on retraining in the current session.

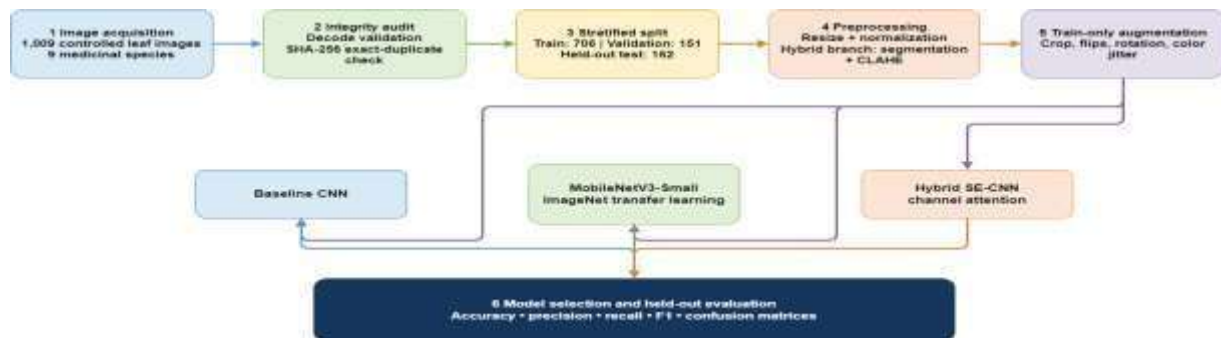


Figure 1. Complete experimental workflow from acquisition to classification, evaluation and inference.

3.2 Dataset acquisition and integrity auditing

The project dataset contains 1,009 leaf photographs organized in one folder per species. Each candidate file was opened and decoded before indexing; unreadable files would have been excluded. Exact duplicates were screened before splitting by computing a SHA-256 digest over the complete file bytes and grouping records with identical digests, irrespective of filename or folder. The audit found 1,009 unique digests, no invalid files, and no exact-duplicate groups. The split was then generated from the retained index, so a byte-identical file could not occur in more than one subset. No perceptual-hash, feature-embedding, image-registration, or specimen-identifier procedure was applied in the reported run. Consequently, visually similar recaptures, resized or recompressed copies, and multiple photographs of the same biological specimen may cross partitions. The terms leakage-aware and unique in this paper therefore refer only to exact-file duplication and split ordering; they do not establish specimen-level independence. A complete near-duplicate audit should compute perceptual hashes and pretrained-image embeddings before splitting, cluster candidates above prespecified similarity thresholds, manually adjudicate clusters, and assign each accepted specimen cluster to a single partition.

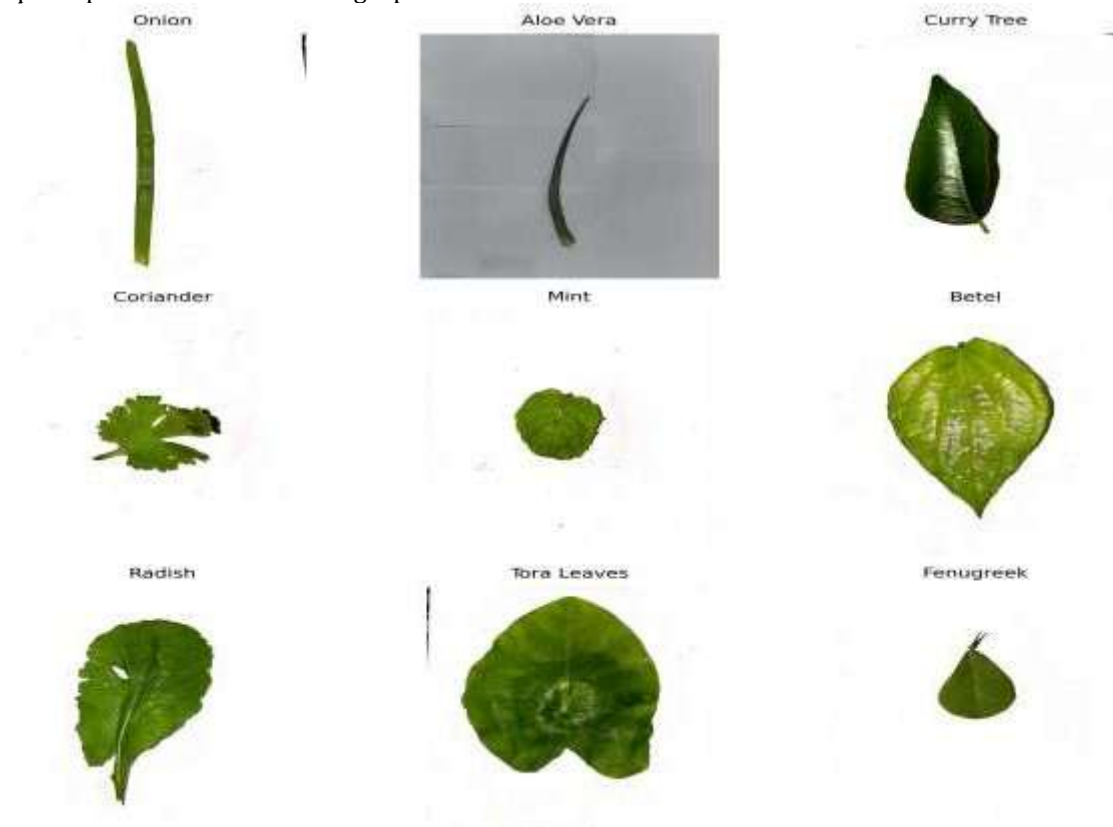


Figure 2. Representative images from the nine medicinal-leaf classes.

Species	Images
Allium cepa (Onion)	107
Aloe vera (Aloe Vera)	103
Bergera koenigii (Curry Tree)	112
Coriandrum sativum (Coriander)	107
Mentha spp. (Mint)	114
Piper betle (Betel)	114
Raphanus sativus (Radish)	136
Senna tora (Tora Leaves)	108
Trigonella foenum-graecum (Fenugreek)	108

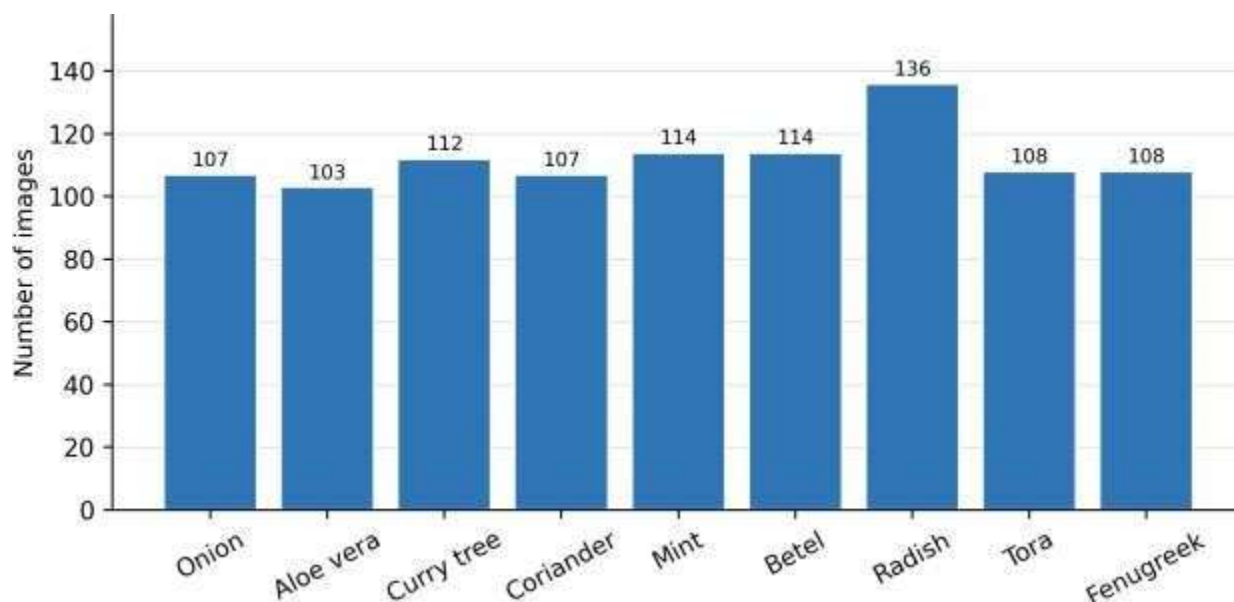


Figure 3. Audited class distribution (total: 1,009 unique images).

3.3 Leakage-aware stratified splitting

A fixed random seed of 42 generated a stratified 70/15/15 partition: 706 training images, 151 validation images, and 152 test images. Class proportions were preserved. Augmentation was applied only by the training data loader after the split; validation and test images were resized and normalized deterministically. The held-out test set was not used for checkpoint selection.

Partition	Images	Share	Role
Training	706	69.97%	Parameter optimization; stochastic augmentation
Validation	151	14.97%	Loss-based checkpoint selection and early stopping
Held-out test	152	15.06%	One-time comparative evaluation

3.4 Hybrid preprocessing and segmentation

The hybrid branch first converts RGB input to HSV and uses a saturation threshold to identify colored foreground. An inverse Otsu threshold on grayscale contributes a dark-object cue. The union mask is cleaned with morphological opening and closing, and only the largest connected component is retained. Foreground pixels are composited on a near-white background. The segmented image is then converted to CIELAB, and contrast-limited adaptive histogram equalization (CLAHE; clip limit 1.5, 8×8 tiles) is applied only to the lightness channel before conversion back to RGB. This branch is used by the hybrid SE-CNN; the baseline and MobileNetV3 consume RGB images directly.

3.5 Data augmentation

Training samples are resized, randomly cropped within a scale range of 0.82–1.00, horizontally flipped, vertically flipped with probability 0.15, rotated up to 18° , and perturbed by modest brightness, contrast, and saturation jitter. Images are then converted to tensors and normalized with ImageNet channel statistics. Evaluation samples receive only resizing, tensor conversion, and normalization.

Betel leaf example — MobileNetV3 prediction: Piper betle(Betel) (92.3%)

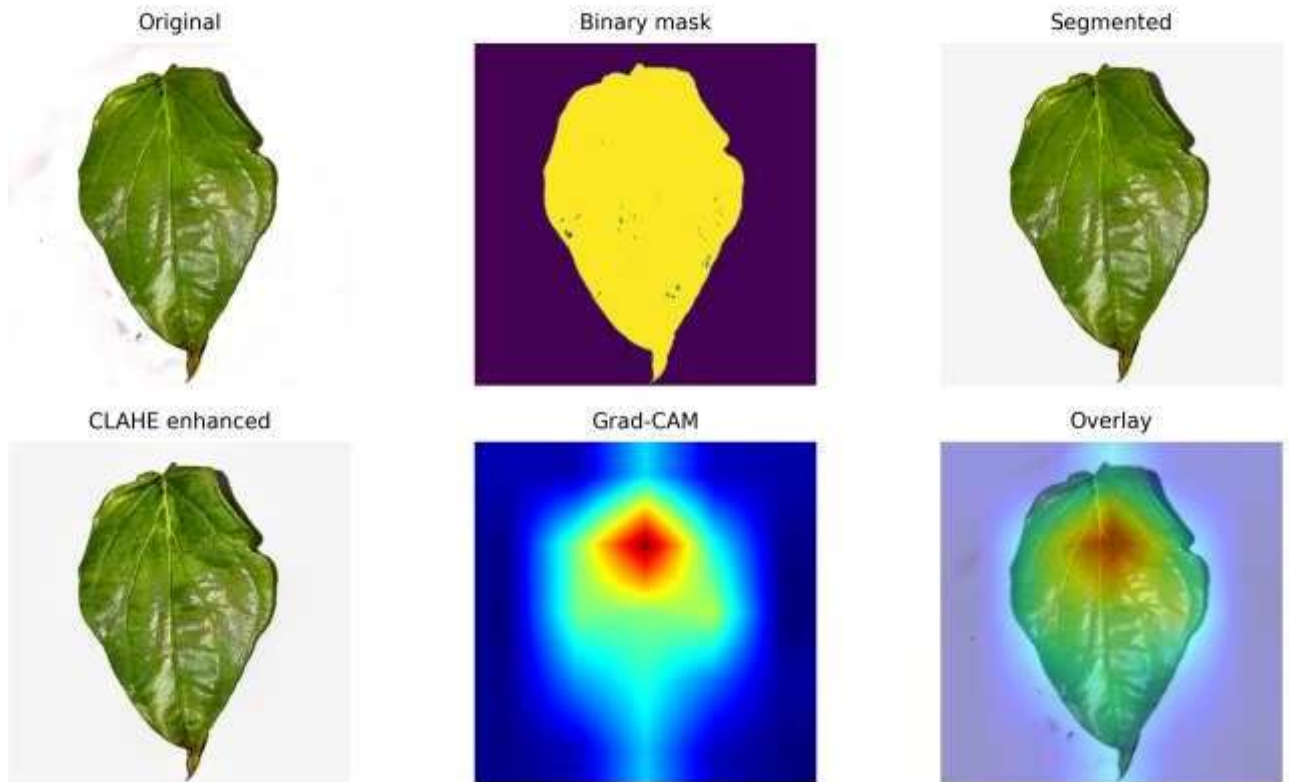


Figure 4. Hybrid preprocessing stages and MobileNetV3 Grad-CAM for a held-out Betel leaf; prediction confidence was 92.3%.

3.6 Compared deep-learning models

Model	Architecture	Parameters	Initialization	Input branch
Baseline CNN	Four Conv-BN-ReLU×2 blocks; max pooling; GAP; dropout; linear head	1,175,529	Random initialization	RGB
MobileNetV3-Small	Efficient inverted residual/depthwise network; replaced 9-class head	1,527,081	ImageNet	RGB
Hybrid SE-CNN	Baseline blocks plus SE attention after each block	1,197,829	Random initialization	Segmentation + CLAHE

3.7 Proposed SE attention model

Let U contain C feature channels. Spatial squeeze computes z_c as the average activation of channel c . Two 1×1 transformations form a bottleneck with reduction ratio $r = 8$; ReLU supplies nonlinearity and sigmoid produces a channel gate s . The recalibrated feature map is obtained by channel-wise multiplication:

$$z = \text{GAP}(U); \quad s = \text{sigmoid}(W_2 \text{ReLU}(W_1 z)); \quad \tilde{U} = s \odot U$$

Each of the four convolutional blocks uses 32, 64, 128, and 256 output channels, respectively. Global average pooling and 0.35 dropout precede the nine-class linear classifier.

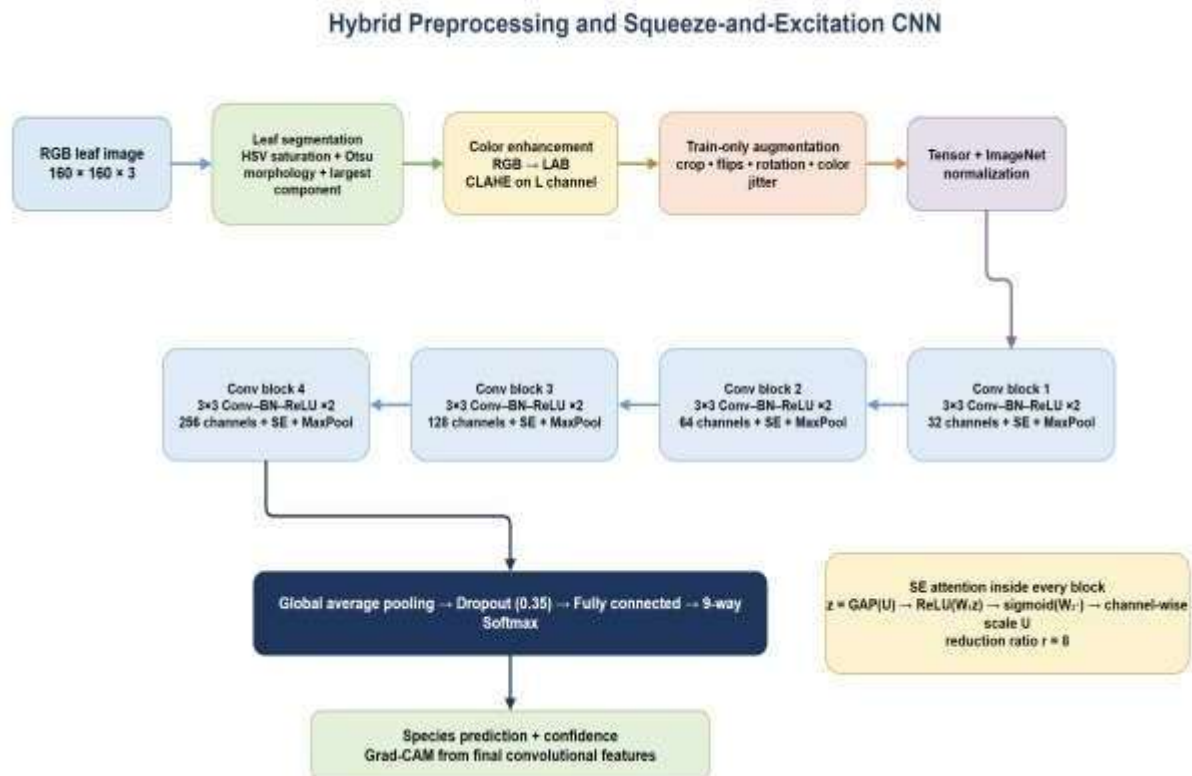


Figure 5 Hybrid preprocessing and SE-CNN architecture.

3.8 Training and checkpoint selection

All three configurations used a 160×160 input, batch size 4, AdamW optimization, initial learning rate 3×10^{-4} , weight decay 10^{-4} , label smoothing 0.08, ReduceLROnPlateau scheduling, early-stopping patience of two epochs, and random seed 42 for the documented split. Training was stopped after three epochs to demonstrate execution within a low-memory CPU pilot budget. The minimum-validation-loss checkpoint, never the test score, was retained. This budget was sufficient for the pretrained MobileNetV3-Small to reach a low validation loss by epoch 2, but it does not demonstrate convergence and is not a fair capacity comparison for the randomly initialized baseline and SE-CNN. The results must therefore be interpreted as a resource-constrained pilot, not as evidence that the baseline or SE mechanism is intrinsically inferior after adequate optimization. A confirmatory comparison requires longer training to convergence, identical stopping rules, and repeated seeds.

3.9 Evaluation

Evaluation reports accuracy, weighted precision, weighted recall, weighted F1-score, class-wise precision/recall/F1, confusion matrices, and a nonparametric 95% bootstrap confidence interval for accuracy. Because every model was evaluated on the same 152 images, model comparisons are paired. Exact McNemar tests were used on discordant correctness outcomes where those counts could be reconstructed from the reported confusion matrices. For MobileNetV3-Small versus the hybrid model, the identity-level overlap of errors was not retained in the supplied artifacts; therefore, the analysis reports the most conservative exact p-value over every overlap pattern compatible with the matrices. No multiplicity adjustment was applied because only two prespecified comparisons against MobileNetV3-Small were made.

3.10 Software implementation

The modular Python implementation separates configuration, auditing and splitting, preprocessing, PyTorch model definitions, training, evaluation, inference, Grad-CAM, utilities, orchestration, and Streamlit presentation. Each timestamped run stores its configuration, split membership, histories, checkpoint, predictions, metrics, and figures; a stable registry supports inference after restart. The retained manuscript artifacts document CPU execution, seed 42, the hyperparameters above, the exact split sizes, model parameter counts, and run identifier 20260910_000719. Exact Python and library versions, processor model, RAM capacity, operating-system build, and deterministic-backend flags were not captured in the retained run manifest. These omissions prevent bitwise reproduction and are reported as a limitation rather than reconstructed retrospectively. Future runs should export `python --version`, `pip freeze`, PyTorch and torchvision versions, OpenCV and scikit-learn versions, operating-system details, processor/RAM information, and all Python, NumPy, and PyTorch seeds with deterministic settings.

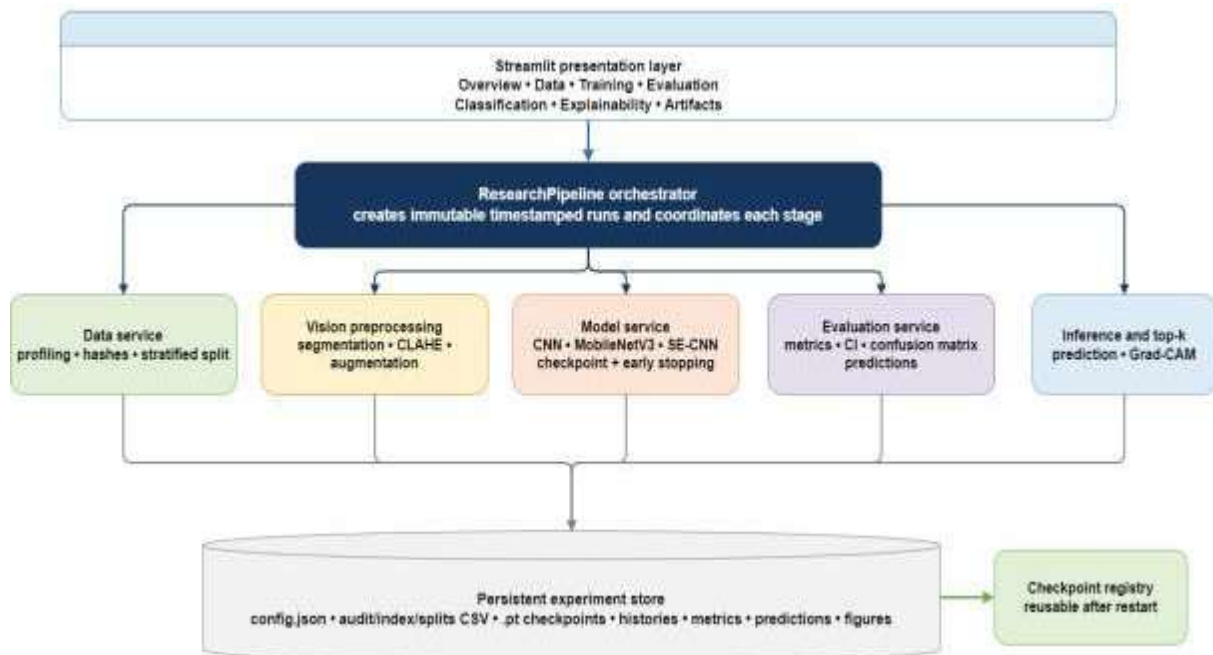


Figure 6. Software architecture and persistent artifact flow.

4. Results and Discussion

4.1 Dataset and execution evidence

The application detected nine class folders and three registered models while running on CPU. The dark-mode interface retained readable contrast across controls, metrics, tabs, and explanatory text. Figure 7 documents the implementation overview; Figure 8 records the live integrity audit; Figure 9 shows confirmation of the stratified, leakage-safe split.

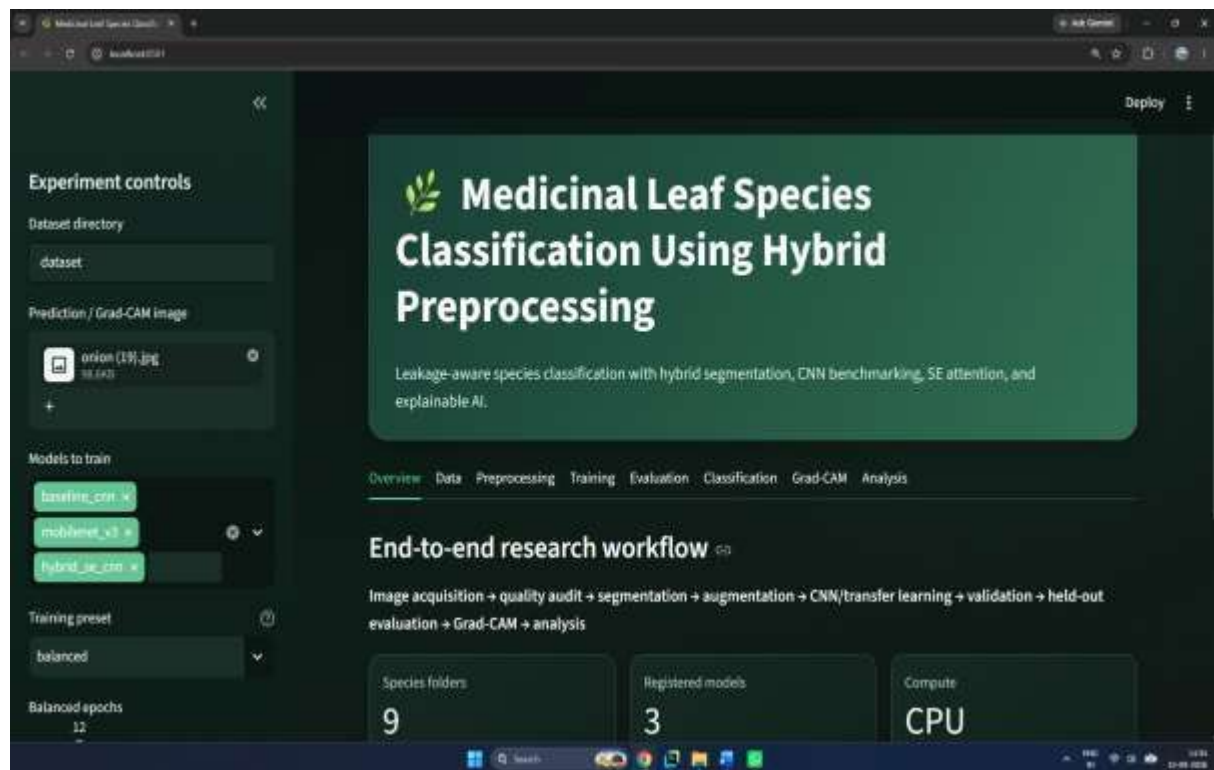


Figure 7. Overview showing nine species, three reusable models, CPU execution, and the scientific-integrity warning.

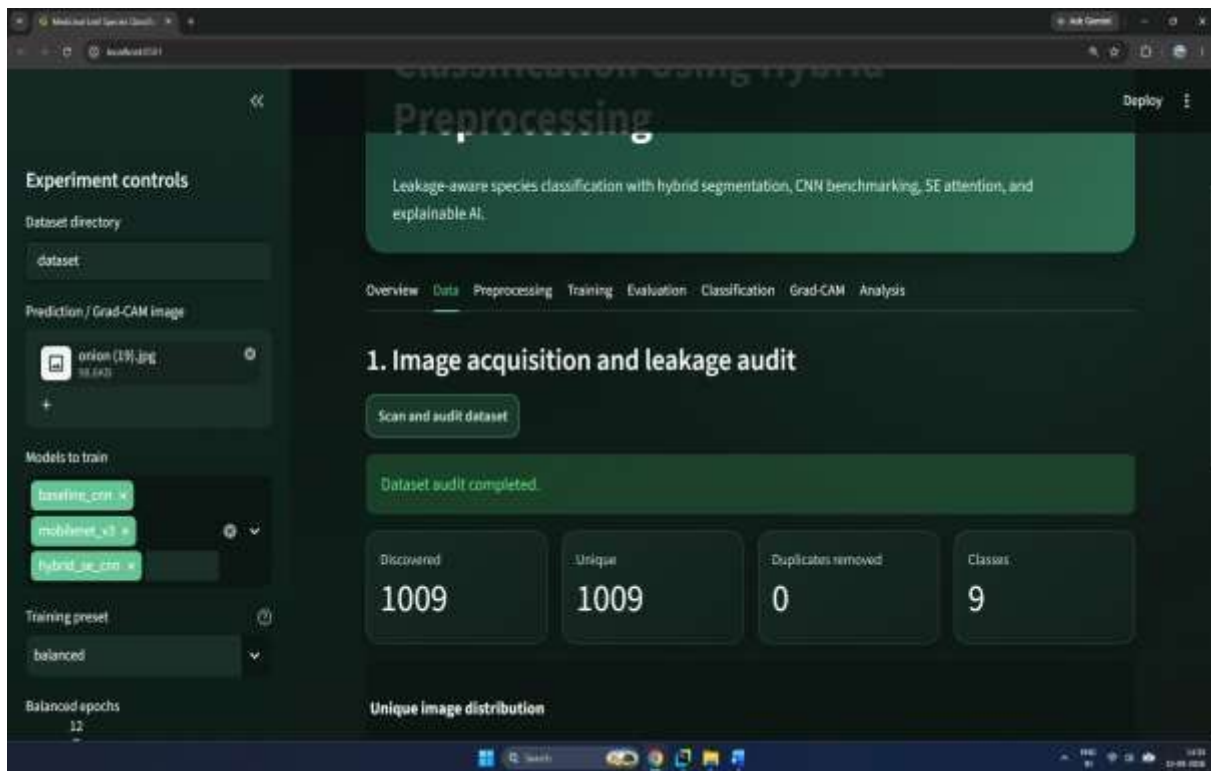


Figure 8. Execution screenshot of the dataset audit: 1,009 discovered and unique images, zero exact duplicates, and nine classes.

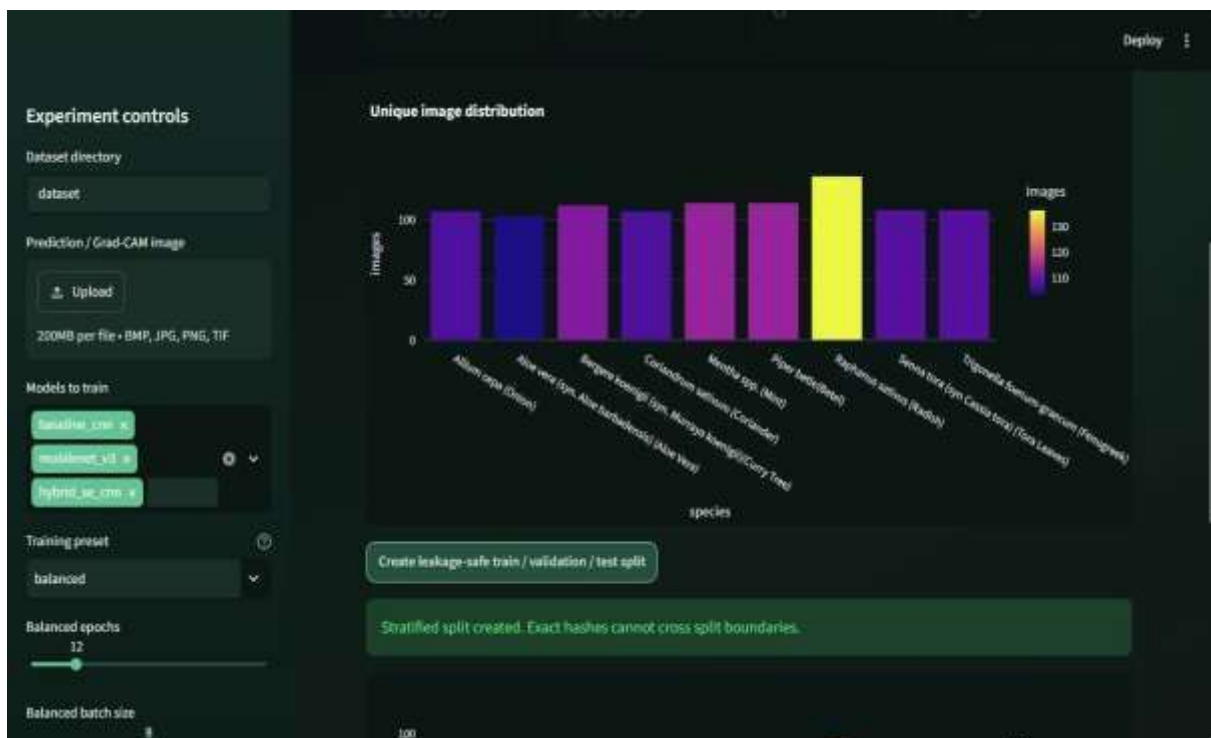


Figure 9. Execution screenshot confirming class distribution and the leakage-safe stratified split.

4.2 Training behavior

MobileNetV3 learned rapidly from its pretrained representation: validation accuracy increased from 94.70% in epoch 1 to 99.34% in epoch 2 and ended at 97.35% in epoch 3. Its minimum validation loss (0.5136) occurred in epoch 2, so that checkpoint was used. The baseline CNN improved from 40.79% to 62.32% training accuracy and reached 71.52% validation accuracy at epoch 3; its best validation accuracy was 74.17% at epoch 1, but minimum validation loss occurred at epoch 3. The hybrid SE-CNN improved steadily and reached 64.24% validation accuracy at epoch 3, yet remained undertrained. The large gap between MobileNetV3 and both randomly initialized networks demonstrates the practical value of transfer learning under a strict compute budget.

Model	Epoch	Train acc.	Val. acc.	Train loss	Val. loss
Baseline CNN	1	40.79%	74.17%	1.6972	1.1898
Baseline CNN	2	54.53%	62.25%	1.4211	1.1631
Baseline CNN	3	62.32%	71.52%	1.2615	1.0914
MobileNetV3	1	73.80%	94.70%	1.0444	0.6223
MobileNetV3	2	91.36%	99.34%	0.6859	0.5136
MobileNetV3	3	93.63%	97.35%	0.6053	0.5408
Hybrid SE-CNN	1	31.30%	40.40%	1.8808	1.5168
Hybrid SE-CNN	2	50.42%	39.74%	1.5302	1.4882
Hybrid SE-CNN	3	58.07%	64.24%	1.3054	1.1663

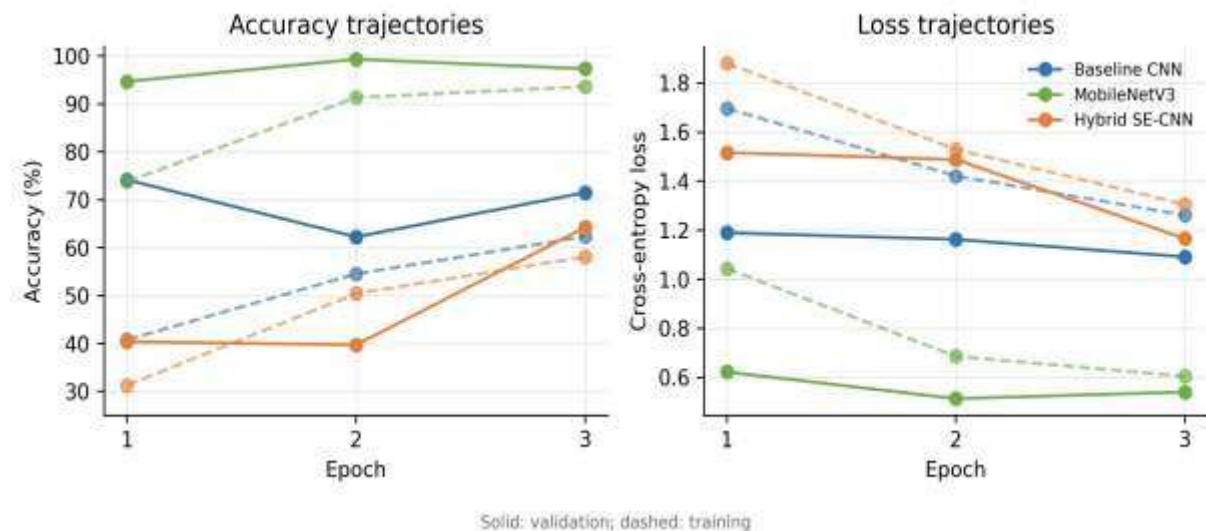


Figure 10. Training and validation trajectories for the three-epoch laptop protocol.

4.3 Held-out test performance

MobileNetV3-Small achieved 98.68% accuracy, weighted precision of 98.76%, weighted recall of 98.68%, and weighted F1-score of 98.68% on the 152-image test set. Its reported bootstrap interval was 96.71–100.00%; the upper boundary reflects the small test set and is not evidence of perfect population performance. The baseline CNN achieved 71.71%, while the hybrid SE-CNN achieved 62.50%. These results describe the selected checkpoints under the three-epoch pilot and do not establish convergence-equivalent model rankings.

4.3.1 Paired statistical comparisons

For MobileNetV3-Small versus the baseline CNN, the confusion matrices imply 43 images correct only for MobileNetV3-Small and two correct only for the baseline; the exact two-sided McNemar p-value is 5.89×10^{-11} . For MobileNetV3-Small versus the hybrid SE-CNN, MobileNetV3-Small made two errors and the hybrid made 57, but identity-level overlap was unavailable. Evaluating all feasible overlaps (zero, one, or two shared errors) gives exact two-sided McNemar p-values from 5.55×10^{-17} to 6.14×10^{-15} ; the conservative value is therefore $p < 6.2 \times 10^{-15}$. Statistical significance does not remove the design limitations caused by a single split, one seed, short training, and unequal initialization.

Model	n	Accuracy	Precision (w)	Recall (w)	F1 (w)	Accuracy 95% CI
Baseline CNN	152	71.71%	69.07%	71.71%	67.99%	65.13–78.29%
MobileNetV3	152	98.68%	98.76%	98.68%	98.68%	96.71–98.92%
Hybrid SE-CNN	152	62.50%	73.07%	62.50%	58.31%	54.61–69.74%

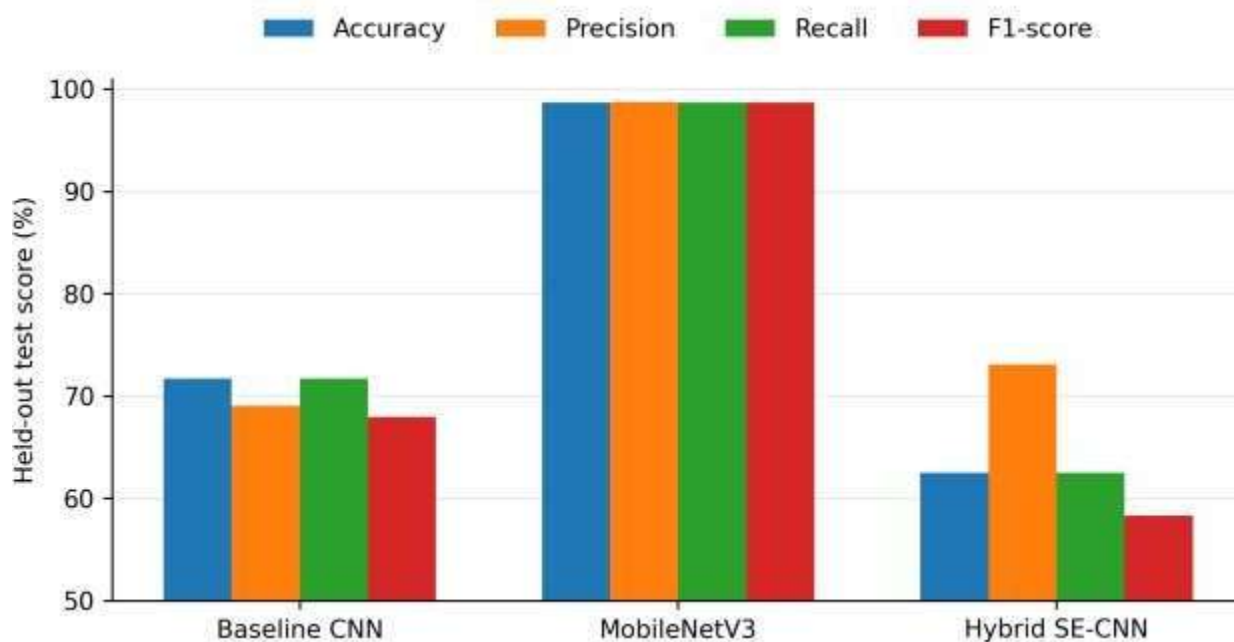


Figure 11. Held-out comparison of accuracy and weighted precision, recall, and F1-score.

4.4 Class-wise performance

Class-wise metrics reconstructed from the reported confusion matrices are shown in Table 8. MobileNetV3-Small maintained F1-scores of at least 96.55% for every class. Weighted averages concealed severe hybrid-model failures: Coriander recall was 6.25% (F1 11.76%) and Betel recall was 11.76% (F1 21.05%). The baseline also struggled with Mint (F1 40.00%) and Fenugreek precision (51.61%).

Table 8. Class-wise precision, recall, and F1-score reconstructed from the held-out confusion matrices.

Class	MobileNet P/R/F1 (%)	Baseline P/R/F1 (%)	Hybrid P/R/F1 (%)
Onion	94.12 / 100.00 / 96.97	94.12 / 100.00 / 96.97	100.00 / 81.25 / 89.66
Aloe Vera	100.00 / 93.33 / 96.55	100.00 / 100.00 / 100.00	81.25 / 86.67 / 83.87
Curry Tree	94.44 / 100.00 / 97.14	92.31 / 70.59 / 80.00	88.89 / 94.12 / 91.43
Coriander	100.00 / 100.00 / 100.00	100.00 / 93.75 / 96.77	100.00 / 6.25 / 11.76
Mint	100.00 / 100.00 / 100.00	46.15 / 35.29 / 40.00	36.84 / 41.18 / 38.89
Betel	100.00 / 100.00 / 100.00	100.00 / 94.12 / 96.97	100.00 / 11.76 / 21.05
Radish	100.00 / 100.00 / 100.00	92.86 / 61.90 / 74.29	46.43 / 61.90 / 53.06
Tora Leaves	100.00 / 100.00 / 100.00	94.44 / 100.00 / 97.14	70.83 / 100.00 / 82.93
Fenugreek	100.00 / 93.75 / 96.77	51.61 / 100.00 / 68.09	41.94 / 81.25 / 55.32

4.5 Confusion-matrix analysis

The MobileNetV3 confusion matrix contained only two errors: one Aloe vera image was predicted as Onion, and one Fenugreek image was predicted as Curry Tree. Seven classes attained perfect recall. In contrast, the baseline CNN failed to recover Tora leaves and showed confusion involving Mint, Radish, and Fenugreek. The hybrid SE-CNN strongly recognized Curry Tree, Tora, Onion, and Aloe but had very low recall for Coriander (6.25%) and Betel (11.76%). These class-specific failures explain why the hybrid model's weighted precision (73.07%) was substantially higher than its weighted recall (62.50%).

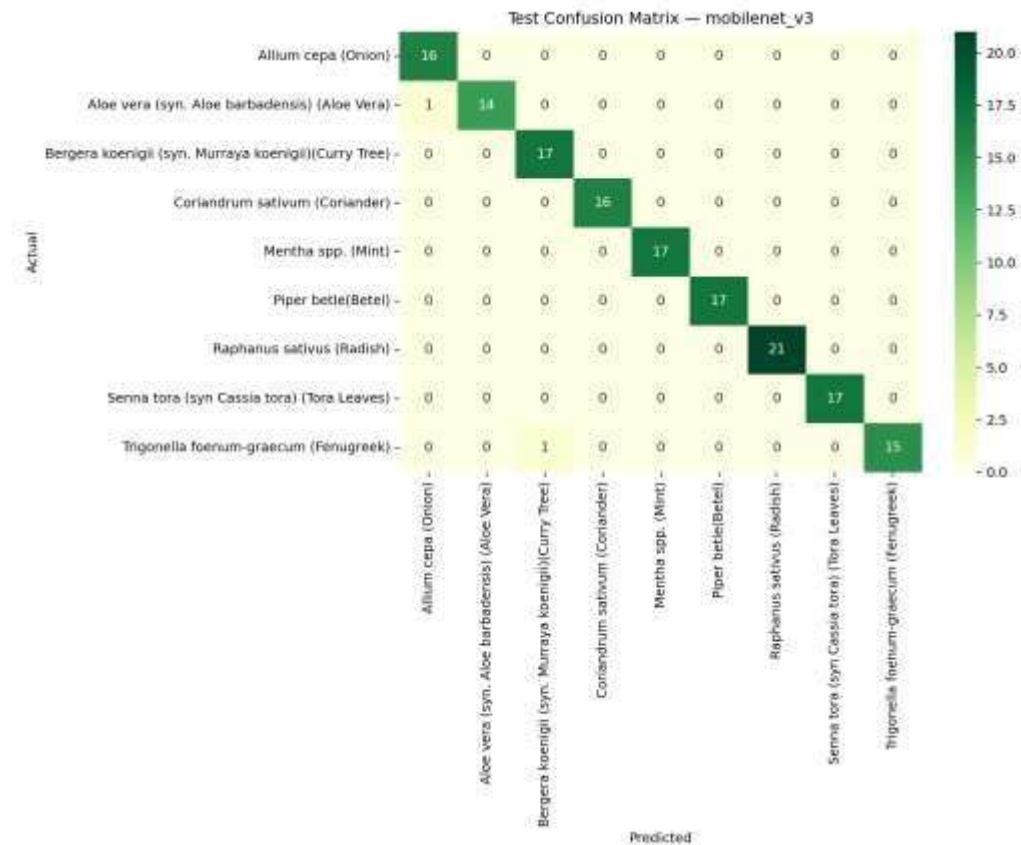


Figure 12. MobileNetV3 held-out confusion matrix (98.68% accuracy).

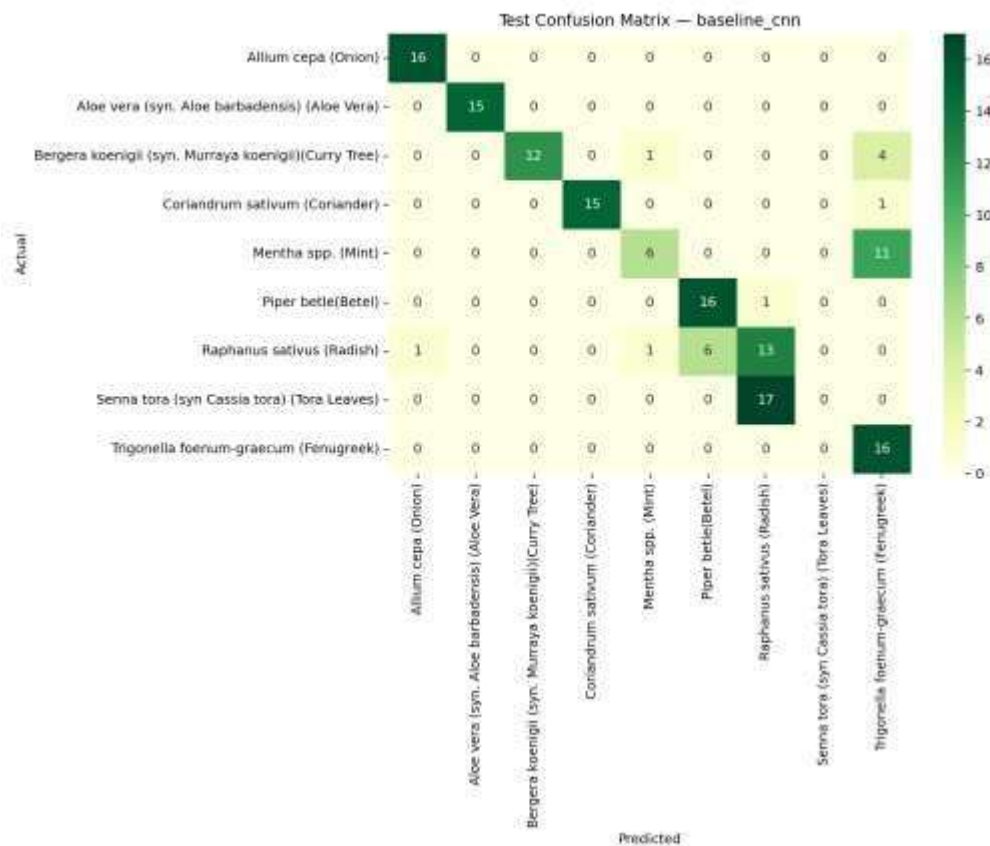


Figure 13. Baseline CNN held-out confusion matrix (71.71% accuracy).

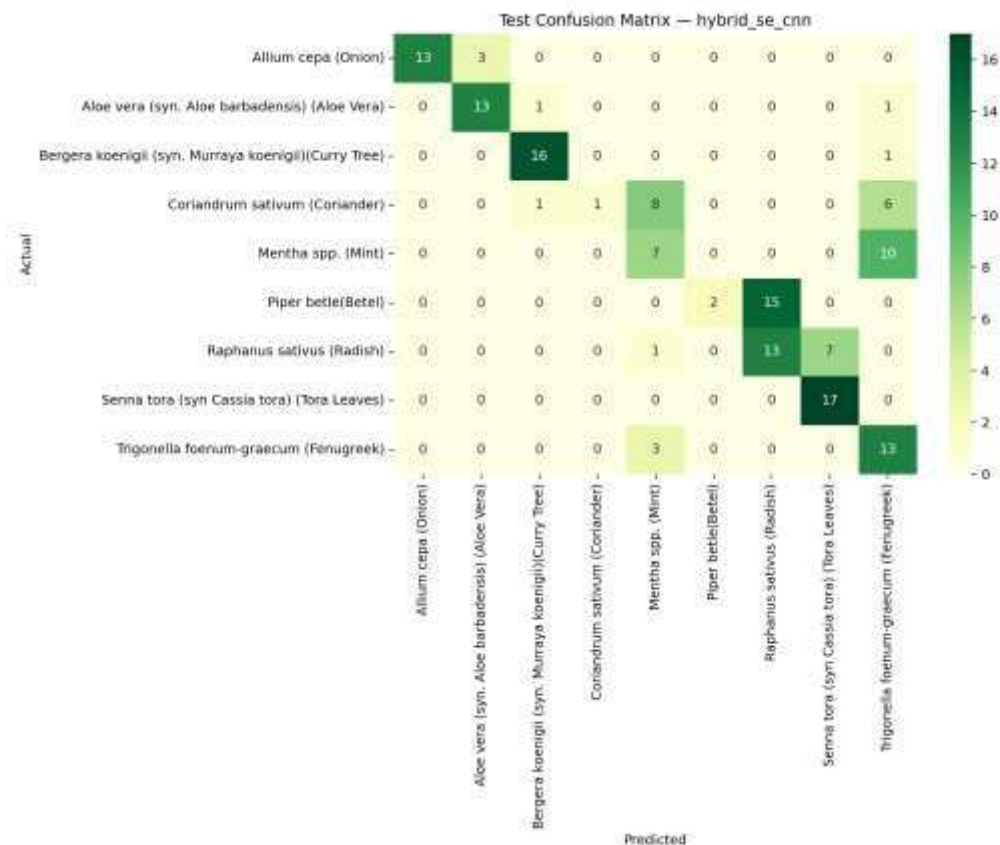


Figure 14. Hybrid SE-CNN held-out confusion matrix (62.50% accuracy).

4.6 Explainability and qualitative inference

A held-out Betel image was correctly classified as Piper betel with 92.3% confidence. Its Grad-CAM overlay concentrated on the upper central lamina and midrib-adjacent texture rather than exclusively on the white background. This was a single illustrative, qualitative inspection. The study did not define localization annotations, a pointing-game or intersection-over-union criterion, a blinded botanical rating protocol, or a representative sampling scheme across classes and errors. Accordingly, Grad-CAM is presented only as a diagnostic example and provides neither systematic explanation validation nor causal evidence that the model used botanically meaningful features.

4.7 Why the hybrid SE-CNN did not outperform transfer learning

MobileNetV3-Small achieved 98.68% accuracy, whereas the hybrid SE-CNN achieved 62.50%. The hybrid model is therefore not the superior method in this experiment. Three confounded factors may explain the gap: pretrained versus random initialization, a short optimization budget, and the combined use of segmentation, CLAHE, and SE blocks. Hard segmentation or CLAHE may remove useful color, margin, or reflectance cues, while SE gates add parameters that may require longer optimization. Because all three hybrid components changed together, the current run cannot attribute the loss to any individual component. The result is a negative system-level finding under the stated protocol, not evidence that segmentation, CLAHE, or SE attention is generally ineffective.

4.8 Component ablation status

A valid ablation must hold initialization, architecture, training duration, and data split constant while toggling one component at a time. The available configurations do not meet that requirement: the baseline differs from the hybrid in three preprocessing or attention components, while MobileNetV3-Small additionally changes the backbone and pretrained initialization. Table 7 is therefore an ablation-coverage audit, not a causal ablation result. It makes the missing cells explicit and prevents the 62.50% hybrid score from being misinterpreted as the isolated effect of SE attention, segmentation, or CLAHE. A confirmatory factorial experiment should train the five listed configurations to the same convergence rule over multiple seeds and compare paired predictions on the unchanged test set.

Table 7. Coverage audit for the requested component ablation; unrun cells are reported explicitly.

Configuration	Segmentation	CLAHE	SE	Initialization	Test accuracy	Attribution
Baseline CNN	No	No	No	Random	71.71%	Observed reference

Configuration	Segmentation	CLAHE	SE	Initialization	Test accuracy	Attribution
Baseline + segmentation	Yes	No	No	Random	Not run	Required to isolate segmentation
Baseline + CLAHE	No	Yes	No	Random	Not run	Required to isolate CLAHE
Baseline + SE	No	No	Yes	Random	Not run	Required to isolate SE
Segmentation + CLAHE + SE-CNN	Yes	Yes	Yes	Random	62.50%	Joint effect only; components confounded
MobileNetV3-Small	No	No	Built-in architecture	ImageNet	98.68%	Different backbone and initialization

4.9 Comparison with internal previous work and baseline techniques

Previous implementation and historical internal reports of approximately 59% for GLCM+SVM and 100% for MobileNetV2; another prior controlled implementation reported 100% for both ResNet50 and MobileNetV2. Those values were obtained under different code, splits, and reporting conditions and are therefore not interchangeable with the current held-out experiment. The current run provides stronger traceability: all 152 test predictions, exact split membership, configuration, histories, and checkpoints are saved. MobileNetV3 reached 98.68%, a more credible high score accompanied by two observed errors and a confidence interval.

Technique	Evidence source	Accuracy	Interpretation
GLCM + SVM	Internal previous report	≈59.0%	Handcrafted texture baseline; different protocol
MobileNetV2	Internal previous report	100.0%	Controlled-data result; generalization concern
ResNet50	Internal previous report	100.0%	Controlled-data result; generalization concern
Baseline CNN	Current held-out run	71.71%	Three epochs from random initialization
MobileNetV3	Current held-out run	98.68%	Best current model; two test errors
Hybrid SE-CNN	Current held-out run	62.50%	Undertrained; preprocessing/attention require ablation

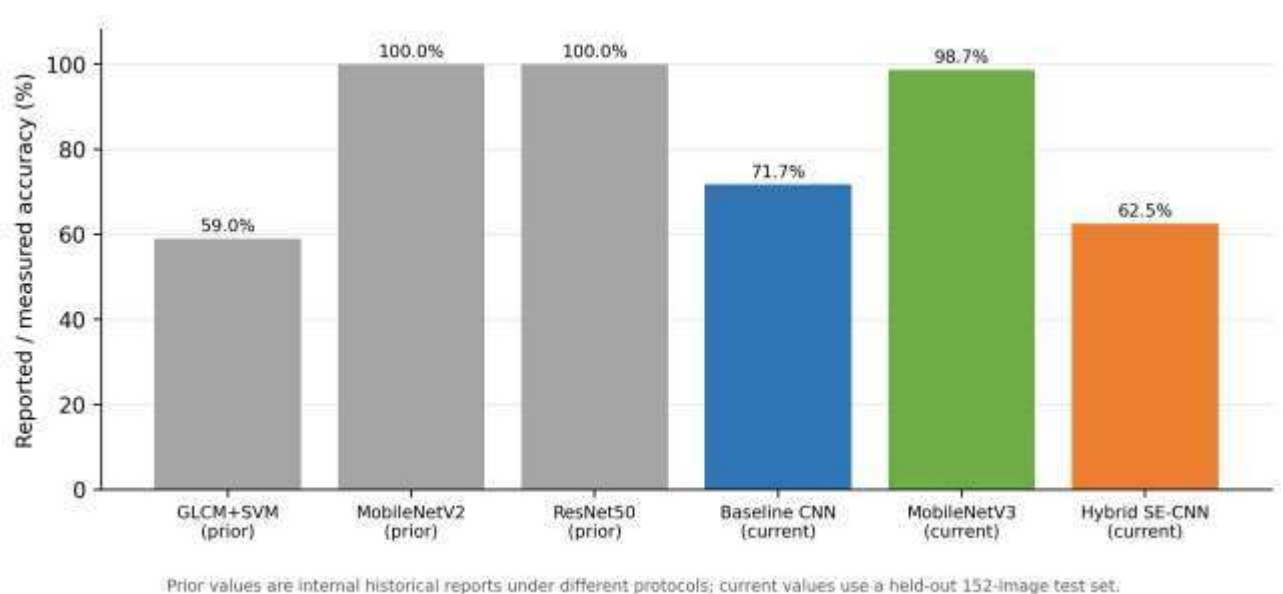


Figure 15. Internal historical values versus the current measured run; protocols differ and bars are not a direct ranking.

4.10 Comparison with published work

The 98.68% MobileNetV3-Small result is reported beside prior medicinal-leaf studies only to establish context, not to rank methods. The cited studies differ in species identity and count, sample size, image provenance, acquisition conditions, specimen dependence, augmentation, partitioning, pretraining, checkpoint selection, and metric definitions. These differences can be larger than the reported accuracy gaps and invalidate a direct claim of superiority. The current study can compare its three models internally because they share one test set and protocol; comparisons with published work remain descriptive and hypothesis-generating.

Study	Images classes /	Best technique	Reported accuracy	Protocol note
Current work	1,009 / 9	MobileNetV3-Small	98.68% test	Hash audit; 70/15/15; CPU; full predictions
Azadnia et al.	5 species	CNN + GAP	99.32–99.66%	Three input resolutions
Dey et al.	5,878 / 30	DenseNet201	99.64%	Public + field images
Sharma & Vardhan	Indian plant + leaf	MTJNet	99.71%	Local/global multi-task fusion
Kiflie et al.	1,853 / 35	VGG19	94% test	Expert-labeled Ethiopian species
Lee et al.	Four benchmarks	CNN-ViT ensemble	99.83–100%	High-cost ensemble; different domains

4.11 Reproducibility and application behavior

Training outputs are written to immutable timestamped run directories and copied to a stable checkpoint registry. Restarting Streamlit changes only the interactive session; it does not erase models. A user can therefore audit a dataset, create a split, train when desired, stop the process externally if necessary, and later classify or generate Grad-CAM from registered checkpoints. The saved run used for this paper is 20260910_000719. The paper-specific result JSON and benchmark CSV reproduce every aggregate number in this section.

5. Threats to Validity and Limitations

- SHA-256 screening removes byte-identical duplicates only; visually similar captures, transformations of the same image, and repeated photographs of one specimen were not detected or grouped.
- The dataset is controlled and locally assembled; performance under field clutter, occlusion, disease, seasonal change, camera variation, and unseen geography remains unknown.
- Only one deterministic split and one random seed were evaluated. Repeated stratified runs or specimen-grouped cross-validation are needed for variance estimates.
- The three-epoch CPU protocol does not establish convergence and disadvantages randomly initialized models; longer repeated-seed training is required for a fair architecture comparison.
- The available model configurations do not form a component-wise ablation, so the individual effects of segmentation, CLAHE, and SE attention cannot be estimated.
- Grad-CAM was assessed on one illustrative case without localization ground truth or blinded expert ratings.
- Exact software versions and complete hardware metadata were not retained in the run manifest, limiting exact replication.
- No external dataset was evaluated; all performance and statistical conclusions are restricted to the controlled local split.

6. Conclusion

This study evaluated an auditable medicinal-leaf classification workbench on one controlled nine-species split. Under a three-epoch CPU pilot, ImageNet-initialized MobileNetV3-Small achieved 98.68% held-out accuracy and 98.68% weighted F1-score, substantially exceeding the randomly initialized baseline CNN and segmentation-plus-CLAHE SE-CNN. Paired statistical analysis confirmed that these within-split differences were unlikely to arise from discordant predictions alone, and class-wise metrics showed that MobileNetV3-Small avoided the severe species-specific failures observed in the hybrid configuration. The result supports MobileNetV3-Small as the strongest tested checkpoint for this dataset and protocol; it does not establish novelty of the underlying architecture, isolate the effects of preprocessing or SE attention, demonstrate convergence-equivalent superiority, validate Grad-CAM systematically, or support generalization beyond the local controlled data.

7. Future Work

Collect specimen identifiers and create specimen-grouped splits to ensure that images of the same leaf or plant do not cross partitions. Add perceptual-hash and embedding-based near-duplicate detection, and then conduct a manual review of the suspected groups. Acquire an external field dataset that encompasses various factors such as phones, illumination, backgrounds, seasons, leaf age, damage, and disease states.

Run repeated-seed experiments and nested cross-validation while reporting the mean, standard deviation, confidence intervals, and paired significance tests. Perform controlled ablations for segmentation, CLAHE, augmentation, SE placement, reduction ratio, image resolution, and pretrained initialisation. To begin, initialise the SE-CNN using a pretrained backbone, or consider extending the training process by implementing gradual unfreezing, mixed precision, and conducting a learning-rate search. Evaluate probability calibration, abstention thresholds, out-of-distribution detection, and top-k botanical decision support. Validate Grad-CAM localisation with botanists and compare it with complementary explanation methods like occlusion sensitivity. Package the workbench for offline deployment, ensuring to include model cards, dataset documentation, and automated regression tests.

References

- [1] S. Azadnia, A. A. Al-Amidi, H. Mohammadi, A. C. D. Cavallo, and colleagues, "An AI Based Approach for Medicinal Plant Identification Using Deep CNN Based on Global Average Pooling," *Agronomy*, vol. 12, no. 11, art. 2723, 2022. doi: 10.3390/agronomy12112723.
- [2] B. Dey, J. Ferdous, R. Ahmed, and J. Hossain, "Assessing deep convolutional neural network models and their comparative performance for automated medicinal plant identification from leaf images," *Heliyon*, vol. 10, no. 1, e23655, 2024. doi: 10.1016/j.heliyon.2023.e23655.
- [3] S. Sharma and M. Vardhan, "MTJNet: Multi-task joint learning network for advancing medicinal plant and leaf classification," *Knowledge-Based Systems*, vol. 299, art. 112147, 2024. doi: 10.1016/j.knosys.2024.112147.
- [4] C. P. Lee, K. M. Lim, Y. X. Song, and A. Alqahtani, "Plant-CNN-ViT: Plant Classification with Ensemble of Convolutional Neural Networks and Vision Transformer," *Plants*, vol. 12, no. 14, art. 2642, 2023. doi: 10.3390/plants12142642.
- [5] J. Hu, L. Shen, and G. Sun, "Squeeze-and-Excitation Networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7132–7141.
- [6] R. R. Selvaraju et al., "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 618–626.
- [7] A. Howard et al., "Searching for MobileNetV3," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 1314–1324.
- [8] M. A. Kiflie, D. P. Sharma, and M. A. Haile, "Deep learning for Ethiopian indigenous medicinal plant species identification and classification," *Journal of Ayurveda and Integrative Medicine*, vol. 15, no. 6, art. 100987, 2024. doi: 10.1016/j.jaim.2024.100987.
- [9] A. K. Mulugeta, D. P. Sharma, and A. H. Mesfin, "Deep learning for medicinal plant species classification and recognition: A systematic review," *Frontiers in Plant Science*, vol. 14, art. 1286088, 2024. doi: 10.3389/fpls.2023.1286088.
- [10] J. Yao, S. N. Tran, S. Garg, and S. Sawyer, "Deep Learning for Plant Identification and Disease Classification from Leaf Images: Multi-prediction Approaches," *ACM Computing Surveys*, vol. 56, no. 6, art. 153, 2024. doi: 10.1145/3639816.
- [11] S. Hossain, R. Hasan, and J. Uddin, "Medicinal Plant Leaf Classification using Deep Learning and Vision Transformers," *Baghdad Science Journal*, vol. 22, no. 3, pp. 1065–1076, 2025. doi: 10.21123/bsj.2024.10844.
- [12] K. Kayaalp, "Classification of Medicinal Plant Leaves for Types and Diseases with Hybrid Deep Learning Methods," *Information Technology and Control*, vol. 53, no. 1, pp. 19–36, 2024. doi: 10.5755/j01.itc.53.1.34345.
- [13] B. R. Pushpa, S. Jyothsna, S. Lasya, et al., "HybNet: A hybrid deep models for medicinal plant species identification," *MethodsX*, vol. 14, art. 103126, 2025. doi: 10.1016/j.mex.2024.103126.
- [14] M. A. Widneh and A. A. Alemu, "Identification of medicinal plant parts using depth-wise separable convolutional neural network," *PLOS ONE*, vol. 20, no. 5, e0322936, 2025. doi: 10.1371/journal.pone.0322936.
- [15] V. Gautam, G. Kaur, G. S. P. Ghantasala, P. Vidyullatha, S. Allabun, M. Othman, and A. Zheleznyak, "A lightweight deep learning method for medicinal leaf image classification using feature fusion," *Scientific Reports*, vol. 15, art. 34417, 2025. doi: 10.1038/s41598-025-17436-w.
- [16] S. Sachar and A. Kumar, "Levy's flight-based dragonfly optimized machine learning framework to identify medicinal plants through leaf images," *Pattern Recognition*, vol. 168, art. 111879, 2025. doi: 10.1016/j.patcog.2025.111879.
- [17] C. A. O. Ong, F. S. Tay, Y. L. Then, et al., "An evaluation of a pre-trained transformer-based self-distillation model (DINOv2) for cross-domain plant species identification," *Neural Computing and Applications*, vol. 37, pp. 21969–21995, 2025. doi: 10.1007/s00521-025-11499-6.
- [18] A. Lapkovskis, N. Nefedova, and A. Beikmohammadi, "Automatic fused multimodal deep learning for plant identification," *Frontiers in Plant Science*, vol. 16, art. 1616020, 2025. doi: 10.3389/fpls.2025.1616020.

- [19] A. Zulfiqar, E. Izquierdo, and K. Chandramouli, "Harnessing deep learning for plant species classification: A comprehensive review," *Computers and Electronics in Agriculture*, vol. 237, art. 110663, 2025. doi: 10.1016/j.compag.2025.110663.
- [20] A. Arun, "Identification of Plant Species Using Convolutional Neural Network with Transfer Learning," *Journal of Phytopathology*, vol. 173, 2025. doi: 10.1111/jph.70032.