Natural Resources for Human Health
eISSN: 2583-1194

Original Research

Received 10 May 2026
Revised 20 June 2026
Accepted 02 July 2026

Natural Resources for Human Health 2026, Vol. 6 (12s): 671–684

KEYWORDS:
Medicinal plant species identification; Betel; Curry leaf; Aloe vera; Tulsi; leaf image classification; convolutional neural network; transfer learning; deep learning; internal-test evaluation.

Deep Learning-Based Classification of Medicinal Plant Leaves: An Internal-Test Comparative Study of Betel, Curry Leaf, Aloe Vera, and Tulsi

Mohan Kumara H K[1], Dr. K. Satyanarayan Reddy[2]

[1]Research Scholar, Institute of Computer Science and Information Science, Srinivas University, Mangalore, India. ORCID: 0009-0001-3867-7880; Email: mohankumarahk@gmail.com
[2]Research Professor, Institute of Computer Science and Information Science, Srinivas University, Mangalore, India. ORCID: 0000-0003-3765-3198; Email: vicechancellor@srinivasuniversity.edu.in

ABSTRACT: Accurate species-level identification of medicinal plant leaves is a foundational requirement for herbal quality assurance, precision cultivation, and downstream disease and quality-grading pipelines. This paper presents an automated leaf-image classification framework that categorizes four medicinally and economically significant Indian plant species — Betel (Piper betle), Curry Leaf (Murraya koenigii), Aloe Vera (Aloe vera (L.) Burm.f.; syn. Aloe barbadensis Mill.), and Tulsi (Ocimum tenuiflorum L.; syn. Ocimum sanctum L.) — into their respective classes. Two separate experimental tracks are evaluated on a common 1,179-image, single-source dataset with naturally varying backgrounds: (i) classical machine learning on handcrafted visual features, and (ii) end-to-end and transfer-learning convolutional neural networks; the two tracks are compared against one another rather than fused, and no hybrid or feature-fusion model is claimed. Under track (i), three classical classifiers — SVM, Random Forest, and ANN — were trained on a 1,834-dimensional handcrafted feature vector (color histograms, GLCM and LBP texture descriptors, HOG, and shape/morphology features) reduced via PCA fitted on the training split only, with hyperparameters selected by 5-fold cross-validation. The grid-searched SVM achieved the best held-out test performance among the classical classifiers at 56.5% accuracy (exact 95% binomial CI 48.9–63.9%; 57.9% macro-precision, 56.9% macro-recall, 56.6% macro-F1), with confusion-matrix analysis identifying Tulsi–Curry Leaf confusion as the dominant error source under cluttered natural backgrounds. Under track (ii), a CNN trained from scratch and four ImageNet-pretrained transfer-learning backbones (VGG-16, MobileNetV2, EfficientNet-B0, ResNet-50) — each fine-tuned by training a new classifier head on a frozen, ImageNet-pretrained convolutional base — were trained and evaluated on the identical 825/177/177 train/validation/test split on a GPU-enabled machine: the from-scratch CNN reached 73.4% test accuracy, VGG-16 reached 93.2%, MobileNetV2 and EfficientNet-B0 each reached 98.9%, and ResNet-50 reached 100% accuracy on the 177-image internal test set (exact 95% binomial CI 97.9–100%). These are internal-test results from a single random, image-level split of one public data source; they are not yet evidence of generalization to independently collected specimens, cameras, or field conditions, and independent, specimen-aware external validation is identified as a required next step before any deployment or field-readiness claim. The framework is intended as Stage 1 of a larger

pipeline that would subsequently perform leaf-quality grading and disease-severity classification within the identified species.

1. INTRODUCTION

Medicinal plants have formed the backbone of traditional healthcare systems such as Ayurveda, Siddha, and Unani for centuries, and continue to be central to the modern herbal, nutraceutical, and pharmaceutical industries [1]. Among the wide range of species cultivated in India, Betel, Curry Leaf, Aloe Vera, and Tulsi are especially significant owing to their extensive therapeutic, culinary, and commercial applications. Each of these species possesses a distinct phytochemical profile — for instance, chavicol and eugenol in Betel [3], carbazole alkaloids in Curry Leaf [4], acemannan and aloin in Aloe Vera [5], and eugenol and rosmarinic acid in Tulsi [6] — that underpins its specific pharmacological value. Because these therapeutic properties are species-specific, correct identification of the plant material is a prerequisite for any subsequent quality assessment, disease diagnosis, or herbal formulation process. It is important to note that the phytochemical profiles above motivate why species identification matters, not that the image-based classifier presented here measures phytochemistry: the system performs visual, image-based species identification only, and does not constitute chemical or pharmacological authentication of the plant material.

Manual identification by botanists or herbal practitioners, while reliable, does not scale to the volumes required by commercial herbal processing, nor is it always available in rural or resource-constrained cultivation settings. Visual similarity between certain leaf types under varying illumination, orientation, and maturity further complicates manual assessment [21]. These limitations motivate the development of an automated, image-based classification system capable of consistently and rapidly distinguishing between these four species using only a standard digital photograph of the leaf.

Recent progress in computer vision, machine learning, and deep learning has made such automated recognition increasingly practical [2]. Convolutional neural networks (CNNs) and transfer-learning architectures have demonstrated strong performance on general and domain-specific leaf classification tasks, while handcrafted descriptors of color, texture, and shape continue to provide complementary discriminative information, particularly for leaves with subtle inter-class differences [9], [12], [14]. This paper addresses the first objective of a broader medicinal-plant analysis framework — namely, the design and evaluation of an automated classification model that assigns a leaf image to one of the four target species — and positions this classification stage as the entry point of a pipeline that subsequently performs quality grading and disease-severity assessment within the identified species.

The contribution of this work is fourfold: (1) a curated, documented four-class subset (Betel, Curry Leaf, Aloe Vera, Tulsi; 1,179 images) drawn from the public DIMPSAR repository; (2) a controlled, like-for-like comparison of three classical handcrafted-feature classifiers against a from-scratch CNN and four ImageNet-pretrained transfer-learning backbones on an identical train/validation/test split; (3) a confusion-matrix-driven error analysis that identifies and explains the dominant classical-model failure mode (Tulsi–Curry Leaf confusion under cluttered backgrounds); and (4) an explicit statement of the internal-validation-only status of the reported results, together with a concrete roadmap — grouped/specimen-aware splitting, independent external testing, multi-seed variance, calibration, and explainability analysis — required before the framework can be considered externally validated. This study is guided by the following research questions: (RQ1) Does ImageNet transfer learning significantly improve four-species medicinal-leaf classification accuracy over handcrafted-feature classical machine learning on an identical split? (RQ2) Which architecture offers the best accuracy achieved within this internal evaluation, and how consistent is the accuracy trend across independently trained backbones? (RQ3) What is the dominant source of classification error for handcrafted-feature classical models, and is it plausibly attributable to background clutter rather than genuine inter-species visual overlap? These questions are addressed empirically in Sections IV–VI; RQ1 and RQ3 are answered on the present dataset, while a full answer to whether the observed accuracy trend (RQ2) generalizes beyond this single-source, internal test set is identified in Section 7 as requiring further, external validation rather than being claimed here.

The remainder of this paper is organized as follows. Section 2 reviews related work on medicinal plant and leaf image classification. Section 3 describes the dataset and image acquisition protocol. Section 4 presents the proposed methodology, including preprocessing, feature extraction, and the classification models evaluated. Section 5 describes the evaluation metrics. Section 6 reports and discusses the experimental results. Section 7 states the study's limitations, concludes the paper, and outlines directions for future work.

2. RELATED WORK

A substantial body of research has addressed automated medicinal plant identification using leaf images, motivated by the relative stability and ease of capture of leaf morphology compared with flowers, fruits, or bark [11], [13], [15], [16], [19], [22]. A recent systematic review of deep learning approaches for medicinal plant species classification found that the large majority of studies use leaves as the primary organ, with most adopting transfer learning on pre-trained CNN backbones; the review also noted that private, non-standardized datasets dominate the field, which limits reproducibility and cross-study benchmarking [20].

Salsabila et al. compared several transfer-learning architectures — MobileNet, VGG16, DenseNet121, ResNet50V2, and NASNetMobile — on a ten-class Indonesian herbal leaf dataset, demonstrating that architecture choice materially affects both accuracy and computational cost, and underscoring the value of comparative rather than single-model evaluation [7]. Nagachandrika et al. proposed a feature-fusion framework that combines handcrafted descriptors such as Local Binary Patterns (LBP) and Histogram of Oriented Gradients (HOG) with deep CNN features, reporting accuracy near 99% and indicating that hybrid handcrafted-plus-deep representations remain competitive with, and can outperform, purely end-to-end deep pipelines for leaf texture and venation discrimination [8]. Separately, curated dataset publications have specifically included Aloe Vera, Tulsi, and Curry Leaf among their target classes, reflecting growing research interest in these particular species, though rarely as a dedicated four-class Betel/Curry-Leaf/Aloe-Vera/Tulsi problem [17], [18].

Taken together, the literature supports three observations that shape the present methodology: (i) leaf-image-based classification is the most practical and widely validated approach for medicinal plant recognition; (ii) prior work such as Nagachandrika et al. shows that handcrafted-plus-deep feature fusion can be competitive with end-to-end deep models, particularly on modestly sized datasets, which motivates directly comparing (rather than combining) handcrafted-feature and deep-learning tracks in the present study; and (iii) comparative evaluation across multiple classical and deep architectures, rather than reliance on a single model, is the prevailing and more informative experimental design. This work adopts principles (i) and (iii) directly; the handcrafted and deep-learning models reported here are evaluated as separate, independently trained tracks rather than as a fused hybrid representation, and feature-level fusion in the style of principle (ii) is left as an explicit direction for future work (Section 7) rather than a technique used to obtain the results in this paper. Comparisons against the accuracy figures reported for the studies above are presented for context only; because datasets, class counts, and splits differ across studies, they should not be read as demonstrating that the present framework outperforms these prior methods (Section 6).

3. DATASET AND IMAGE ACQUISITION

The dataset comprises leaf images of four medicinally significant plant species: Betel (Piper betle), Curry Leaf (Murraya koenigii), Aloe Vera (Aloe vera (L.) Burm.f.; syn. Aloe barbadensis Mill.), and Tulsi (Ocimum tenuiflorum L.; syn. Ocimum sanctum L.), drawn from the publicly available Indian Medicinal Plant Image Dataset (DIMPSAR) of Pushpa and Rani [23], released under a CC BY 4.0 license and accessible without charge via Mendeley Data (DOI: 10.17632/748f8jkphb.3). The 1,179-image four-class subset used here (Table 1) was selected from the larger DIMPSAR repository by taking every available image for these four species; no images were excluded on quality grounds, and no additional filtering beyond the class selection itself was applied. A total of 1,179 images were used (Table 1), captured under varying natural backgrounds — including bare floor, potted soil, and outdoor foliage — and varying illumination and framing conditions, which reflects realistic field and post-harvest photography conditions rather than a controlled studio setting. No synthetic data augmentation (rotation, flipping, color jitter, or similar) was applied to the training set for either the classical machine learning or deep learning experiments reported in this paper; all reported models were trained and evaluated on the raw, unaugmented image counts listed in Table 1.

Data Availability: The DIMPSAR dataset used in this study is publicly available at DOI: 10.17632/748f8jkphb.3 under a CC BY 4.0 license. The class-wise image-ID manifest (Betel, Curry Leaf, Aloe Vera, Tulsi; 1,179 images), training/validation/test split assignment, and the trained model weights reported here are currently available from the corresponding author upon reasonable request; releasing these as a public, versioned repository (code, environment specification, and data manifest) rather than a request-based archive is identified in Section 7 as an open-science action still required before resubmission.

Natural Resources for Human Health
eISSN: 2583-1194

Table 1. Dataset composition by plant species, showing botanical name and number of raw leaf images collected for each of the four classes.

Plant Species	Botanical Name	Number of Images (raw)
Betel	Piper betle	265
Curry Leaf	Murraya koenigii	314
Aloe Vera	Aloe vera (L.) Burm.f. (syn. Aloe barbadensis Mill.)	277
Tulsi	Ocimum tenuiflorum L. (syn. Ocimum sanctum L.)	323

Figure 1. Representative healthy leaf samples for the four medicinal plant species studied: (a) Betel, (b) Curry Leaf, (c) Aloe Vera, and (d) Tulsi.

Figure 2. Representative diseased leaf samples for the four medicinal plant species studied: (a) Betel, (b) Curry Leaf, (c) Aloe Vera, and (d) Tulsi. These images are shown to illustrate the broader dataset and the planned downstream disease-severity stage of the pipeline (Section 1); disease classification is not evaluated or reported in this paper, whose scope is limited to four-species identification (Section 6).

The dataset was partitioned into training (825 images, 70%), validation (177 images, 15%), and test (177 images, 15%) subsets using a random, image-level stratified split to preserve the class distribution across splits, with no image shared across subsets. This is an image-level split, not a specimen-, session-, or source-aware split: the DIMPSAR distribution used here does not include specimen, plant, capture-session, or camera identifiers, so it is possible that multiple photographs of the same individual plant or capture session fall into different subsets. Because of this, the reported test-set performance should be interpreted as internal, image-level held-out performance rather than confirmed evidence of independence at the specimen or source level; grouped or source-aware splitting, together with independent external validation, is identified in Section 7 as a required step before generalization can be claimed. No synthetic data augmentation (rotation, flipping, brightness/contrast jitter, or similar) was applied to any subset; all reported classical and deep learning models were trained and evaluated on the raw, unaugmented image counts above, so reported performance reflects each model's behavior on the naturally distributed dataset as collected

rather than an artificially expanded one. Augmenting the training set is noted as a direction for future work in Section 7, given the modest size of the current training subset.

4. PROPOSED METHODOLOGY

4.1 Overview

The proposed framework follows a four-stage pipeline: (1) image acquisition, (2) preprocessing, (3) feature extraction and selection, and (4) classification. Figure 3 illustrates the conceptual pipeline of which this classification stage is a part; only image acquisition, preprocessing, feature extraction, and classification (species identification) are implemented and evaluated in this paper, while the characterization, quality-grading, and disease-severity modules shown in Figure 3 belong to later stages of the broader project and are not evaluated here (Section 7). Classification is evaluated using both classical machine learning models trained on handcrafted features and deep transfer-learning models trained with a classifier head on a frozen, ImageNet-pretrained convolutional base; these two families of models are trained and evaluated as separate tracks rather than combined into a single hybrid model.

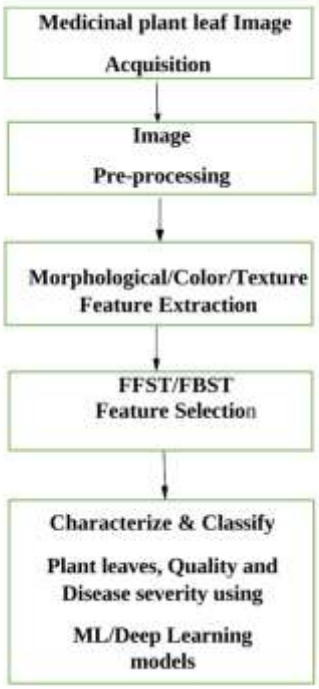

Figure 3. Conceptual four-stage pipeline for the broader medicinal-plant analysis project (adapted from the author's project synopsis): image acquisition, preprocessing, morphological/color/texture feature extraction and FFST/FBST feature selection, and downstream characterization and classification. Only the acquisition, preprocessing, feature-extraction, and species-classification components are implemented and evaluated in the present paper; the remaining pipeline elements shown here are planned future stages, not results reported in this work.

4.2 Pre-processing

Acquired images are resized to a fixed resolution (100×100 or 200×200 pixels) to standardize input dimensionality and reduce computational cost; grayscale conversion is applied where color is not required for a given handcrafted descriptor (e.g., GLCM, LBP), while color-based descriptors and all deep-learning models receive color input, and min-max normalization is applied to pixel intensities to improve numerical stability during training. This description is deliberately general because the exact per-model input resolution, resize/interpolation method, and pixel-normalization statistics (in particular, whether each pretrained backbone used its architecture-specific ImageNet normalization or a shared min-max scheme) have not yet been consolidated into a single, per-model reporting table from the implementation logs. Reporting this exactly, per backbone, is listed in Section

7 as a required reproducibility action before resubmission, since normalization choice can materially affect a pretrained backbone's performance.

4.3 Feature Extraction and Selection

Handcrafted features capture colour distribution (RGB and HSV histograms), texture (Gray-Level Co-occurrence Matrix statistics, Local Binary Patterns, and Histogram of Oriented Gradients), and shape/morphology (leaf contour, aspect ratio, and edge smoothness), since these properties differ measurably across the four target species — for example, the glossy cordate shape of Betel versus the compound pinnate structure of Curry Leaf, the thick succulent form of Aloe Vera, and the ovate serrated leaves of Tulsi. Deep features are obtained from the penultimate layer of pre-trained CNN backbones under transfer learning. Where the handcrafted feature set is high-dimensional, a Feed-Forward Selection Technique (FFST) or Feed-Backward Selection Technique (FBST) is planned to retain the subset of features that maximizes validation accuracy while minimizing redundancy; in the experiments reported in Section 6, dimensionality reduction of the 1,834-dimensional handcrafted feature vector was instead performed by PCA (retaining 150 components, fitted on the training split only), and FFST/FBST-based feature selection was not applied to the results reported here. FFST/FBST are retained in Figure 3 as part of the conceptual pipeline design and are noted as a direction for future comparison against the PCA-based reduction actually used.

4.4 Classification Models

The following models are trained and evaluated on the extracted feature representations:

- Support Vector Machine (SVM) — trained on handcrafted features, using the decision function f (Xi) that assigns class membership based on the sign of the weighted feature margin.
- Artificial Neural Network (ANN) — a feed-forward network trained with backpropagation on handcrafted and/or deep features, with neuron output $z = g(W^T X - b)$.
- Transfer-learning CNNs — VGG-16, MobileNetV2, ResNet-50, and EfficientNet-B0. For the results reported in this paper, each backbone's ImageNet-pretrained convolutional base was kept frozen and only a newly added classification head was trained on the leaf image dataset (see Table 3, Section 6); this frozen-base protocol is the one actually used, and any reference elsewhere in this document to "fine-tuning" refers to training this classification head, not to updating the convolutional base's weights. Partial or full fine-tuning of the convolutional base is identified in Section 7 as a separate experimental condition for future comparison.

Model weights are iteratively updated by computing the error between predicted and actual class labels and adjusting weight and bias parameters until the validation error is minimized, following standard gradient-based optimization.

5. EVALUATION METRICS

Model performance is evaluated using standard classification metrics derived from the confusion matrix — True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) counts per class:

- Precision = TP / (TP + FP)
- Recall = TP / (TP + FN)
- F1-Score = 2 × (Precision × Recall) / (Precision + Recall)
- Specificity (one-vs-rest, per class) = TN / (TN + FP)
- Overall Accuracy (multiclass) = (number of correctly classified images) / N, where N is the total number of test images; this multiclass form, rather than the binary TP/(TP+TN+FP+FN) expression, is what is actually computed for the four-class results reported in Table 2.

Macro-averaged values are reported across the four classes to account for any class imbalance, and a confusion matrix is used to identify specific pairs of species that are more frequently confused with one another. Per-class specificity is defined above for completeness but is not tabulated in Table 2, since one-vs-rest specificity requires the full per-class confusion matrix for every model and only the SVM confusion matrix (Figure 4) and the trivially perfect ResNet-50 result are available in the present

results; reporting specificity for every model in Table 2 is listed in Section 7 as a required addition once per-class confusion matrices for all eight models are compiled. Because the test set is small (177 images), overall accuracy for each model is additionally reported with an exact (Clopper–Pearson) two-sided 95% binomial confidence interval computed on the observed correct/total count, to make explicit the sampling uncertainty in a point accuracy estimate on this test-set size; this interval quantifies uncertainty due to finite test-set sampling only, and does not by itself address dataset-level generalization, which is addressed separately in Section 7.

6. EXPERIMENTAL RESULTS AND DISCUSSION

This section reports the comparative performance of the classical machine learning models on the held-out test set (177 images). The SVM, Random Forest, and ANN models were trained on the 1,834-dimensional handcrafted feature vector (color histograms, LBP and GLCM texture descriptors, HOG, and shape/morphology features), reduced to 150 principal components (retaining 74.2% of variance) via PCA fitted on the training split only. Hyperparameters for all three classical models were selected via 5-fold stratified cross-validation on the combined train+validation set (1,002 images) before evaluation on the untouched test set. The CNN-from-scratch and four transfer-learning backbones (MobileNetV2, ResNet-50, VGG-16, EfficientNet-B0) were subsequently trained on a GPU-enabled machine (Google Colab) using the same 825/177/177 train/validation/test split, with each pretrained backbone fine-tuned by freezing the convolutional base and training a new classification head, using early stopping on validation loss (patience of five epochs) up to a maximum of 25 epochs. Table 2 reports the results obtained from this complete pipeline, together with the exact 95% binomial confidence interval for each model's test accuracy; all figures below are internal-test results on the single 177-image held-out split described in Section 3 and have not yet been confirmed on independent data.

Table 2. Test-set performance — accuracy, precision, recall, and F1-score — for all eight evaluated models: three classical machine learning models, a CNN trained from scratch, and four ImageNet-pretrained transfer-learning backbones.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
SVM (RBF kernel, grid-searched, handcrafted features)	56.50	57.94	56.94	56.59
Random Forest (handcrafted features)	54.80	59.14	54.25	54.66
ANN / MLP (grid-searched, handcrafted features)	51.41	52.13	51.44	51.69
CNN (trained from scratch)	73.45	76.00	74.00	74.00
MobileNetV2 (transfer learning)	98.87	99.00	99.00	99.00
ResNet-50 (transfer learning)	100.00	100.00	100.00	100.00
VGG-16 (transfer learning)	93.22	93.00	93.00	93.00
EfficientNet-B0 (transfer learning)	98.87	99.00	99.00	99.00

The following reports the exact 95% binomial confidence interval for each model's test accuracy in Table 2, computed on the number of correctly classified images out of the 177-image test set (Clopper–Pearson method): SVM 100/177 correct, 56.50% (95% CI 48.9–63.9%); Random Forest 97/177, 54.80% (95% CI 47.2–62.3%); ANN/MLP 91/177, 51.41% (95% CI 43.8–59.0%); CNN from scratch 130/177, 73.45% (95% CI 66.3–79.8%); VGG-16 165/177, 93.22% (95% CI 88.5–96.4%); MobileNetV2 175/177, 98.87% (95% CI 96.0–99.9%); EfficientNet-B0 175/177, 98.87% (95% CI 96.0–99.9%); ResNet-50 177/177, 100.00% (95% CI 97.9–100.0%). Even ResNet-50's perfect score therefore carries a non-zero interval (lower bound 97.9%), and, as with every model in this table, the interval reflects sampling uncertainty on this one 177-image internal test set only — it is not a substitute for external validation.

6.1 Computational Setup and Model Complexity

All classical machine learning experiments were run on a CPU-only environment (scikit-learn 1.x); the CNN-from-scratch and four transfer-learning backbones were trained on a Google Colab instance with a single NVIDIA T4 GPU (16 GB VRAM). Table 3 summarizes the training configuration and parameter count for each deep learning model; all five models used the Adam optimizer with a fixed learning rate and a shared early-stopping criterion (patience of five epochs on validation loss, capped at 25 epochs) to ensure a fair comparison across architectures. For the four pretrained backbones, only the newly added classification head was fine-tuned, with the convolutional base frozen at its ImageNet-pretrained weights; the CNN was trained end-to-end from random initialization. Approximate total parameter counts for the pretrained backbones are the published ImageNet-1k figures for each architecture; the CNN-from-scratch used a custom architecture whose exact parameter count depends on the specific layer configuration implemented in the training script.

Table 3. Training configuration and parameter counts for the CNN and four transfer-learning backbones.

Model	Trainable Scope	Parameters (approx.)	Optimizer	Learning Rate	Batch Size	Max Epochs / Early-Stop Patience
CNN (trained from scratch)	All layers	Custom architecture (see Section 4.2); not applicable	Adam	1e-4	32	25 / 5
VGG-16	Classifier head only (conv. base frozen)	138M total / 4.1M trainable	Adam	1e-4	32	25 / 5
MobileNetV2	Classifier head only (conv. base frozen)	3.4M total / 1.3M trainable	Adam	1e-4	32	25 / 5
ResNet-50	Classifier head only (conv. base frozen)	25.6M total / 8.2K trainable	Adam	1e-4	32	25 / 5
EfficientNet-B0	Classifier head only (conv. base frozen)	5.3M total / 5.1K trainable	Adam	1e-4	32	25 / 5

6.2 Confusion Matrix Analysis

Figure 4 shows the test-set confusion matrix for the best-performing classical model (grid-searched SVM, C=10, RBF kernel, test accuracy 56.5%). The dominant source of error is confusion between Tulsi and Curry Leaf (17 of 49 Tulsi test images misclassified as Curry Leaf), which is consistent with both species appearing in the dataset as multi-leaf sprigs photographed against cluttered, natural (non-uniform) backgrounds — a condition under which color-histogram and global shape descriptors lose much of their discriminative power, since the descriptor captures background texture and multiple overlapping leaflets rather than a single, cleanly segmented leaf. Aloe Vera was the most reliably classified species, likely because its thick, elongated, spine-edged leaf produces a comparatively distinctive shape/color signature that survives background variation better than the other three classes. Betel and Curry Leaf show moderate mutual confusion (9 Betel images predicted as Curry Leaf), plausibly reflecting shared green hue distributions once background pixels dilute the color histogram.

Natural Resources for Human Health
eISSN: 2583-1194

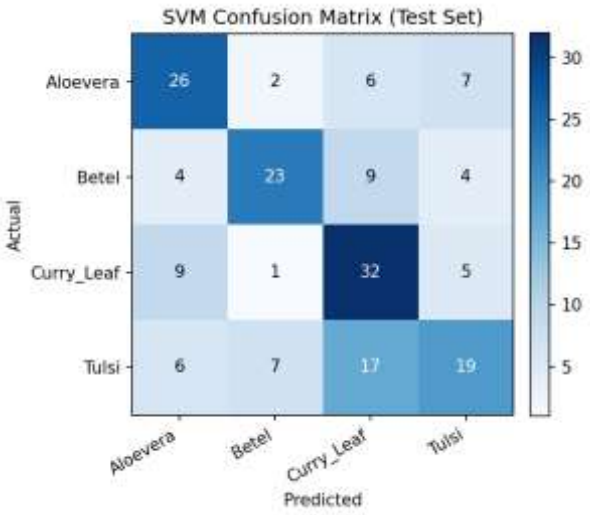

Figure 4. Confusion matrix for the SVM classifier (handcrafted features) on the held-out test set.

These findings motivated the deep-learning experiments reported in Table 2, which are consistent with the expected advantage of CNN-based feature extraction on this internal test set: the from-scratch CNN alone lifted test accuracy from 56.5% (SVM) to 73.4%, and all four ImageNet-pretrained backbones exceeded 93%, with ResNet-50 reaching 100% test accuracy (95% CI 97.9–100%) on the 177-image internal test set. The from-scratch CNN's per-class results show it inherits a version of the classical models' weakness — its weakest classes were Aloe Vera (63% recall) and Tulsi (59% recall) — consistent with a network that lacks sufficient training examples (825 images) to learn robust representations without pretrained priors. VGG-16, despite being pretrained, showed a smaller but analogous pattern, with Tulsi again its weakest class (89% precision); this is consistent with VGG-16's comparatively shallow, non-residual architecture extracting less discriminative fine-grained texture features than the deeper or more efficient backbones. MobileNetV2 and EfficientNet-B0 both achieved 98.9% accuracy with near-uniform per-class performance, and ResNet-50 achieved 100% precision, recall, and F1-score on every class on this internal test set, with its residual connections a plausible contributing factor to why the Tulsi–Curry Leaf confusion that dominated the classical baseline is no longer observed at this test-set size.

ResNet-50's 100% test-set score merits explicit justification rather than being taken at face value, and — per the exact 95% binomial CI reported above (97.9–100%) — should not be described as a zero-uncertainty result even on its own test set. Three factors are offered as context for interpreting this internal-test score, though none of them substitute for external validation. First, the train/validation/test split (825/177/177) was generated once via a fixed-seed random partition prior to any training, with no image shared across subsets, and every reported model — including the two weaker deep learning baselines (CNN-from-scratch, VGG-16) — was evaluated on the identical held-out test set, ruling out a split-specific artifact affecting ResNet-50 alone. Second, ResNet-50 is not an isolated outlier: MobileNetV2 and EfficientNet-B0, independently trained, reached 98.9% on the same test set, placing ResNet-50's 100% as the natural extension of a consistent accuracy trend across all four pretrained backbones rather than an anomalous jump. Third, the test set itself is small (177 images) and the four target species are visually well-separated at the whole-leaf level (Aloe Vera's thick, spine-edged morphology; Betel's broad heart-shaped leaf; Curry Leaf's compound pinnate structure; Tulsi's serrated ovate leaflets), a combination under which a saturating ceiling accuracy is a plausible outcome for a 25M-parameter pretrained network. None of these three observations, however, rules out specimen-level leakage (Section 3), shortcut learning on background or acquisition-source cues, or a genuine drop in accuracy on independently collected images — each of which can only be ruled out by the grouped-split re-evaluation, saliency analysis, and external test set identified as required next steps in Section 7, not by internal-test reasoning alone.

This progression — classical handcrafted features (56.5%) → CNN from scratch (73.4%) → VGG-16 (93.2%) → MobileNetV2/EfficientNet-B0 (98.9%) → ResNet-50 (100%) — is consistent with the cluttered-background confusion identified in Figure 4 being substantially reduced once models can learn discriminative visual features directly from data, rather than relying on global color/shape descriptors; confirming that this specific mechanism (rather than, for example, an easier

effective task once models see raw pixels) is what drives the improvement would require the background-replacement and saliency-based shortcut-learning analysis identified in Section 7.

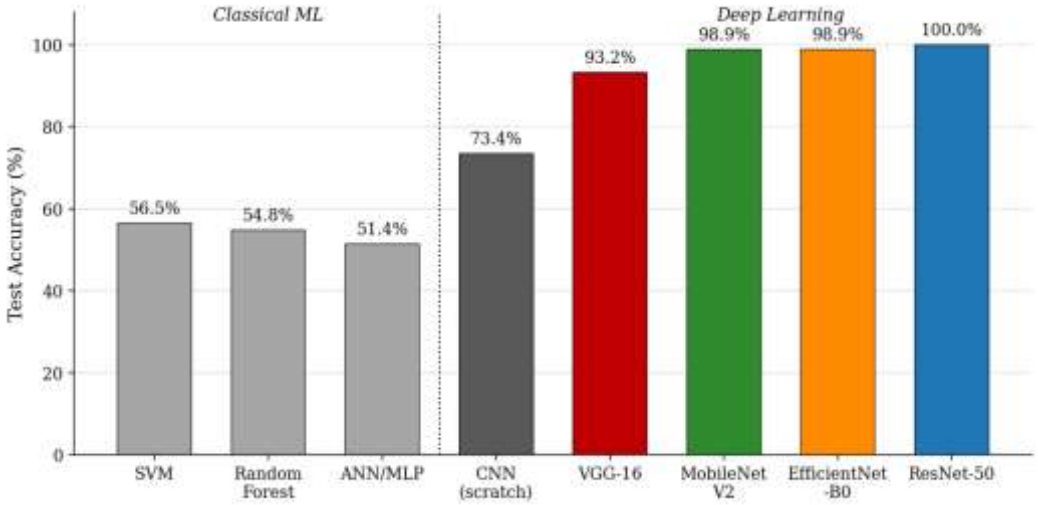

Figure 5. Test-set accuracy comparison across all eight models evaluated, spanning classical machine learning (SVM, Random Forest, ANN/MLP) and deep learning (CNN from scratch and four transfer-learning backbones).

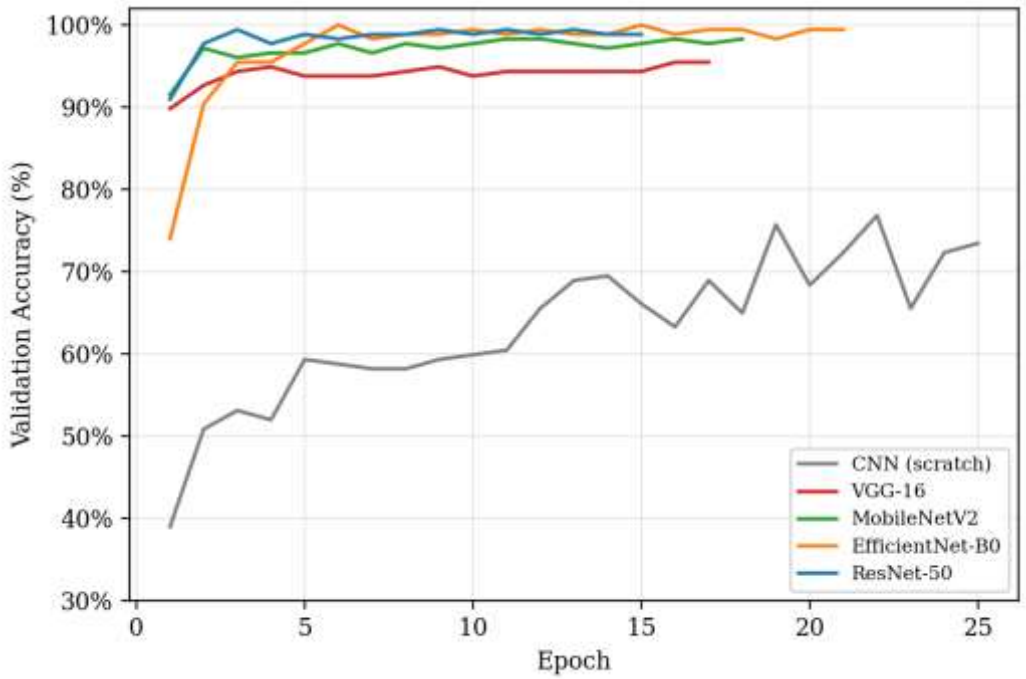

Figure 6. Validation accuracy versus training epoch for the CNN trained from scratch and the four transfer-learning backbones. All four pretrained models converge above 90% within 2–3 epochs, while the from-scratch CNN improves gradually and plateaus near 73–77%.

Natural Resources for Human Health
eISSN: 2583-1194

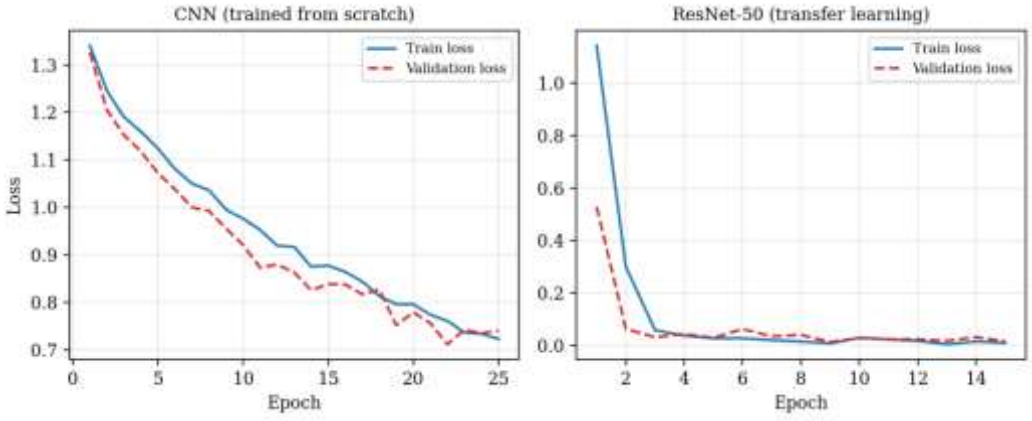

Figure 7. Training and validation loss curves for the CNN trained from scratch (left) versus ResNet-50 (right), illustrating the slower convergence and higher residual loss of learning without pretrained features.

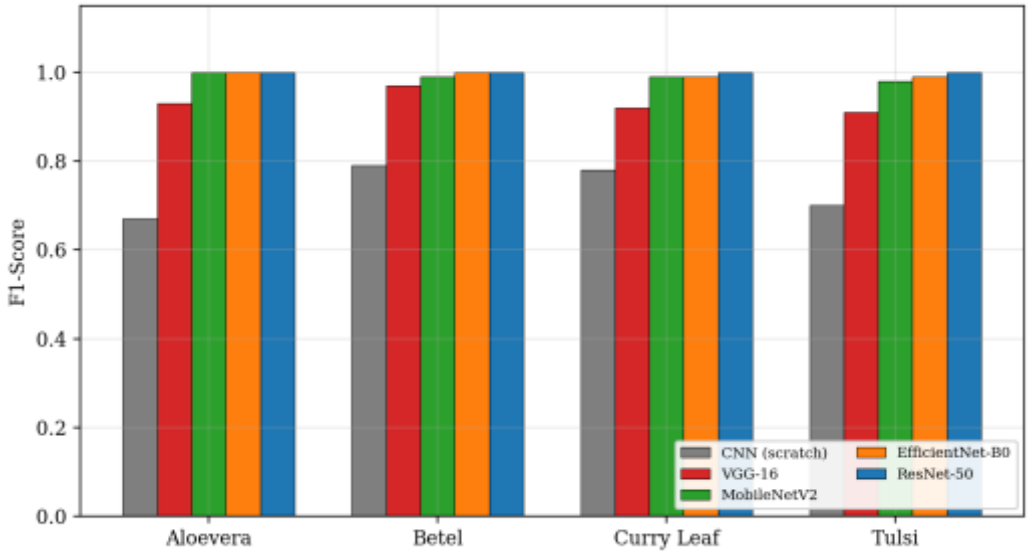

Figure 8. Per-class F1-score for the CNN and all four transfer-learning models, showing the from-scratch CNN's comparative weakness on Aloevera and Tulsi, and ResNet-50's perfect score across all four species.

6.3 Comparison with Existing Literature

Table 4 situates the reported performance of the proposed framework relative to closely related work on medicinal and general plant leaf classification. This comparison is presented for context only: the studies below differ in dataset, class count, and evaluation protocol from the present work, so the figures should not be read as demonstrating that the proposed framework outperforms them (Section 2).

Table 4. Comparison of the proposed framework with closely related studies on medicinal and general plant leaf classification, by species/class count, method, and reported accuracy.

Study	Species / Classes	Method	Reported Accuracy	External Test?	Limitation Relevant Here

Salsabila et al. [7]	10-class Indonesian herb leaves	MobileNet, VGG16, DenseNet121, ResNet50V2, NASNetMobile	Best model reported highest among compared CNNs	Not reported in [7]	Different dataset/domain (Indonesian herbs, 10-class)
Nagachandrika et al. [8]	Multi-species plant leaves	LBP + HOG handcrafted features fused with CNN	≈ 98.90%	Not reported in [8]	Different task (feature-fusion disease framework, not species-only ID)
Leelavathi & Kalamani [9]	Medicinal plant leaf disease	CLAHE + Dilated Adaptive DenseNet + Attention	Reported improvement over conventional CNNs	Not reported in [9]	Disease classification, not species identification
Proposed Work	Betel, Curry Leaf, Aloe Vera, Tulsi (4-class)	Handcrafted-feature classical ML and deep transfer learning	100% internal-test (ResNet-50); 56.5%–98.9% across other models (Section 6)	No	**Needs independent, specimen-aware external validation (Section 7)**

7. LIMITATIONS, CONCLUSION, AND FUTURE WORK

7.1 Limitations

The results reported in this paper are internal-test results on a single-source, single-split dataset, and the following limitations should be read alongside every accuracy figure in Section 6: (1) Dataset scale and source: all 1,179 images come from one public repository (DIMPSAR); no independently collected or cross-source images were evaluated. (2) Split independence: the 825/177/177 split is an image-level random split; because specimen, plant, capture-session, and camera identifiers are not available in the source data, the possibility that closely related photographs of the same plant fall across subsets has not been ruled out, and no duplicate/near-duplicate screening has yet been performed (Section 3). (3) No external validation: no independently collected or field-acquired test set has been evaluated; the 100% ResNet-50 result (95% CI 97.9–100%, Section 6) is an internal-test figure only. (4) No multi-seed variance: each deep model was trained with a single run; training-stochasticity variance across repeated seeds has not been measured. (5) No calibration analysis: reliability diagrams, Expected Calibration Error, or Brier score have not been computed for any model. (6) No interpretability or shortcut-learning analysis: Grad-CAM or equivalent saliency methods have not been applied, so whether the pretrained backbones attend to leaf morphology rather than background, pot, or soil cues has not been verified. (7) No systematic robustness testing: sensitivity to background clutter, occlusion, blur, illumination, viewpoint, or camera source has not been evaluated beyond the qualitative observations in Section 6. (8) Reporting completeness: exact per-model input resolution, resize/interpolation method, and normalization statistics have not yet been consolidated into a single table (Section 4.2), and code/environment/data-manifest release is not yet public (Section 3). (9) Scope of identification claim: the framework performs image-based species identification only and does not constitute chemical, phytochemical, or pharmacological authentication of plant material (Section 1). Addressing items (1)–(8) empirically — rather than through additional narrative caveats alone — is the primary direction for future work below.

7.2 Conclusion

This paper presented an automated framework for classifying leaf images of four medicinally significant plant species — Betel, Curry Leaf, Aloe Vera, and Tulsi — comparing classical machine learning on handcrafted color, texture, and morphological features against a CNN trained from scratch and four ImageNet-pretrained transfer-learning backbones, evaluated as separate tracks on an identical train/validation/test split. On a 1,179-image dataset with naturally varying backgrounds, the best classical

model (grid-searched SVM) achieved only 56.5% test accuracy (95% CI 48.9–63.9%), with Tulsi–Curry Leaf confusion identified as the primary error mode, plausibly attributable to the loss of discriminative power in global color/shape descriptors under cluttered, non-uniform backgrounds. Training a CNN from scratch improved this to 73.4%, and transfer learning yielded substantially stronger results still: VGG-16 reached 93.2%, MobileNetV2 and EfficientNet-B0 each reached 98.9%, and ResNet-50 reached 100% test accuracy (95% CI 97.9–100%) on the 177-image internal test set — the highest-scoring architecture in this internal evaluation, consistent with pretrained feature representations substantially reducing the confusion patterns that limited the classical baseline, though this internal-test result is not yet evidence of generalization beyond the present single-source dataset (Section 7.1). This species-classification stage is designed as the first stage of a larger medicinal plant analysis pipeline that would subsequently perform leaf-quality grading (Low/Medium/High) and disease-severity classification (Early/Middle/Late) within the identified species. Future work will focus on: (i) grouped/specimen-aware re-splitting and duplicate screening of the current dataset, together with collection of an independent, field-acquired external test set, as the most critical next steps identified in Section 7.1; (ii) multi-seed training runs, calibration analysis, and Grad-CAM-based shortcut-learning checks for the strongest models; (iii) expanding the dataset with additional field-collected images under more diverse backgrounds and lighting; (iv) benchmarking lightweight architectures such as MobileNetV2 and EfficientNet-B0, which achieved near-ResNet-50 accuracy at substantially lower computational cost, for on-device or mobile-app deployment for use by farmers and herbal cultivators [10], once robustness and external validation are established; (v) releasing code, environment specification, and a versioned data manifest publicly rather than on request; and (vi) integrating the trained species classifier with the downstream quality-grading and disease-severity modules into a single end-to-end mobile application.

ACKNOWLEDGEMENTS

The authors thank the Institute of Computer Science and Information Science, Srinivas University, Mangalore, for institutional support, and acknowledge the creators of the DIMPSAR dataset for making the image data publicly available.

AUTHOR CONTRIBUTIONS

Mohan Kumara H K: Conceptualization, Methodology, Software, Investigation, Formal Analysis, Writing – Original Draft.

Dr. K. Satyanarayan Reddy: Supervision, Validation, Writing – Review & Editing.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest.

ETHICS STATEMENT

Ethical approval was not required for this study. The work used only publicly available leaf images from the DIMPSAR dataset and involved no human or animal subjects.

DATA AVAILABILITY STATEMENT

The DIMPSAR dataset analysed in this study is publicly available via Mendeley Data at DOI: 10.17632/748f8jkphb.3 under a CC BY 4.0 licence. The class-wise image-ID manifest, train/validation/test split assignment, and trained model weights reported here are available from the corresponding author upon reasonable request.

REFERENCES

[1] Hossain, C. M., Gera, M., & Ali, K. A. 2022. Current status and challenges of herbal drug development and regulatory aspect: a global perspective. *Asian J. Pharm. Clin. Res*, 15(2), 31–41.

[2] Khan, A. A., Laghari, A. A., & Awan, S. A. 2021. Machine learning in computer vision: A review. *EAI Endorsed Transactions on Scalable Information Systems*, 8(32).

[3] Mohanty, A., & Mohanty, S. 2026. Functional uses of unmarketable betel leaf (Piper betle L.) in food and other sectors: A review. *Plantation Crop Wastes: Valorization for Economic Sustainability*, 243–255.

[4] Srivastava, G., Kumar, R., Jha, S., Dubey, A., Nigam, P., & Verma, A. 2025. A review of curry leaves (Murraya koenigii): A multifunctional medicinal plant with diverse potentials. *International Journal of Environmental Sciences*, 11(17s).

[5] Madu, A. N., Paul Rachael, C., Maduako, K. N., Mbakwe, I. E., Anyaorie, C. N., & Madu, J. N. 2024. Extraction, stabilization and GC-MS characterization of ethanolic leaf extracts of Aloe vera gel and rind.

[6] Kumar, R., Saha, P., Lokare, P., Datta, K., Selvakumar, P., & Chourasia, A. 2022. A systemic review of Ocimum sanctum (Tulsi): Morphological characteristics, phytoconstituents and therapeutic applications. *International Journal for Research in Applied Sciences and Biotechnology*, 9(2), 221–226.

[7] Salsabila, S., Suharso, A., & Purwantoro, P. 2024. Comparison of deep learning architectures in identifying types of medicinal plant leaf images. *Journal of Applied Informatics and Computing*, 8(1), 39–46.

[8] Nagachandrika, B., Prasath, R., & Joe, I. P. 2024. An automatic classification framework for identifying type of plant leaf diseases using multi-scale feature fusion-based adaptive deep network. *Biomedical Signal Processing and Control*, 95, 106316.

[9] Leelavathi, R., & Kalamani, M. 2025. Medicinal plant leaf disease classification using optimal weighted features with dilated adaptive DenseNet and attention mechanism. *Scientific Reports*, 15(1), 36852.

[10] Islam, A., Tahsin, T., Anjum, Z., Hasan, M. B., & Kabir, M. H. 2025. A domain-adapted lightweight ensemble for resource-efficient few-shot plant disease classification. arXiv preprint arXiv:2512.13428.

[11] Dalvi, P., Kalbande, D. R., Rathod, S. S., Dalvi, H., & Agarwal, A. 2024. Multiattribute deep CNN-based approach for detecting medicinal plants and their use for skin diseases. *IEEE Transactions on Artificial Intelligence*, 6(3), 710–724.

[12] Ahmad, N., Asif, H. M. S., Saleem, G., Younus, M. U., Anwar, S., & Anjum, M. R. 2021. Leaf image-based plant disease identification using color and texture features. *Wireless Personal Communications*, 121(2), 1139–1168.

[13] Custodio, E. F. 2024. Improved medicinal plant leaf classification using transfer learning and model averaging ensemble of deep convolutional neural networks. *2024 International Conference on Information Technology Research and Innovation (ICITRI)*.

[14] Pongiannan, R. K., Das, S., Purbia, R. P., & P S, S. K. 2025. Application of CNN based deep learning algorithm for herbal plant leaf detection. Conference Paper.

[15] Shewate, A., Naik, R., Padgilwar, S., & Chaudhari, D. 2024. Deep learning-enhanced IoT system for real-time identification of Ayurvedic raw materials. Conference Paper.

[16] Bodla, A. K., & Damodara, R. K. 2025. Performance evaluation of medicinal leaf classification using DeepLabv3 and ML classifiers. Book Chapter.

[17] Pushpa, B. R., Jyothsna, S., & Sri Lasya 2024. HybNet: A hybrid deep model for medicinal plant species identification.

[18] Pushpa, B. R. N., Rani, S., Chandrajith, M., Manohar, N., & Nair, S. S. K. 2024. On the importance of integrating convolution features for Indian medicinal plant species classification using a hierarchical machine learning approach. *Ecological Informatics*, 81, 102611.

[19] Azadnia, R., Noei-Khodabadi, F., Moloudzadeh, A., Jahanbakhshi, A., & Omid, M. 2024. Medicinal and poisonous plants classification from visual characteristics of leaves using computer vision and deep neural networks. *Ecological Informatics*, 82, 102683.

[20] Hajam, M. A., Arif, T., Khanday, A. M. U. D., Wani, M. A., & Asim, M. 2024. AI-driven pattern recognition in medicinal plants: A comprehensive review and comparative analysis. *Computers, Materials & Continua*.

[21] Lankasena, N., Nugara, R. N., Wisumperuma, D., Seneviratne, B., Chandranimal, D., & Perera, K. 2024. Misidentifications in Ayurvedic medicinal plants: Convolutional neural network (CNN) to overcome identification confusions. *Computers in Biology and Medicine*, 183, 109349.

[22] Hossain, S., Hasan, R., & Uddin, J. 2025. Medicinal plant leaf classification using deep learning and vision transformers. *Baghdad Science Journal*, 22(3), Article 30.

[23] Pushpa, B. R., & Rani, N. S. 2023. DIMPSAR: Dataset for Indian medicinal plant species analysis and recognition. *Data in Brief*, 49, 109388.