ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26apr1643

A Comparative Analysis of Machine Learning Classifiers for Medicinal Plant Leaf Identification Using Image-Based Features

Sakthi Saranya S.[1]; Dr. W. Rose Varuna[2]

[1]Research Scholar, Department of Information Technology, Bharathiar University, Coimbatore, India.
[2]Assistant Professor, Department of Information Technology, Bharathiar University, Coimbatore, India.

Publication Date: 2026/04/29

Abstract: The automatic detection of medicinal plant species based on the leaf image is a challenging area of study owing to high inter-species similarity, morphological variations within a class, and noisy image acquisition processes. In this study, an algorithmic technique involving image processing and machine learning methods for recognising the type of medicinal plants based on their leaves. The proposed approach involves image acquisition, contrast improvement by CLAHE and median filtering, segmentation of the leaf image using k-means clustering, and feature extraction through GLCM textures and geometrical shapes. This study analyses five different types of ML classifiers: KNN, Naive Bayes, Decision Tree, Random Forest, and SVM. In this study, the experiments were conducted on the publicly available Medicinal Leaf dataset that contains 1800 images of 30 medicinal plant species. The experimental evaluation reveals that the Random Forest classifier model obtains a higher level of accuracy of 93.80% compared to other ML classifiers with a precision of 92.60%, recall of 93.10%, and F1-score of 92.85%.

Keywords: *CLAHE, K-Means Segmentation, SVM, Random Forest, Naive Bayes, Decision Tree.*

How to Cite: Sakthi Saranya S.; Dr. W. Rose Varuna (2026) A Comparative Analysis of Machine Learning Classifiers for Medicinal Plant Leaf Identification Using Image-Based Features. *International Journal of Innovative Science and Research Technology*, 11(4), 2446-2449. https://doi.org/10.38124/ijisrt/26apr1643

I. INTRODUCTION

Medicinal plants represent an indispensable asset both in conventional and modern medical practice. According to WHO, the current estimate shows that more than 80 percent of the world's population residing in developing countries depends on medicinal plants for their health care. Consequently, the accurate identification of plant species used in medicine is a paramount task; however, this aspect requires attention for biodiversity protection and pharmaceutical investigation as well [1]. The conventional methodology of plant classification involves examination of morphological traits of the plant (leaf shape, venation, texture, color), conducted by experienced botanists. It is a relatively long, biased, and impractical method, especially in poor areas. The emergence of computer vision and artificial intelligence presents an unprecedented opportunity for developing an advanced system for plant recognition [2].

The leaf, which can be obtained at any stage in plant development, is one of the main objects used in automated plant recognition. Nonetheless, the problematics of this classification task include intra-class variability, close resemblance between species, complicated background, and image degradation during acquisition. Thus, the choice of proper preprocessing steps and feature extraction strategy plays a crucial role in achieving good results [3].

II. RELATED WORK

Wang et al. [5] reviewed SVM-based plant identification using 30 leaf features comprising shape, color, and texture descriptors, demonstrating the effectiveness of feature-driven ML classification for botanical identification tasks.

Kan et al. [6] proposed an automated classification framework using ten shape features and five texture features extracted from preprocessed leaf images, subsequently classified by SVM. Their system achieved an average identification accuracy of 93.3% across twelve medicinal plant species. The study demonstrated that multi-feature extraction significantly enhances classifier discriminability for closely related species.

Kaur and Singh [7] conducted a comparative evaluation of SVM and KNN classifiers on herbal plant leaf datasets, reporting that SVM outperformed KNN in accuracy for multi-class leaf classification. This finding is consistent with

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26apr1643

SVM's well-documented strength in high-dimensional feature spaces with clear margin separation.

Kaur et al. [8] applied Random Forest to classify herbal plant species using leaf shape descriptors derived from binary leaf images. Their study highlighted RF's natural resistance to overfitting and its ability to handle high-dimensional feature sets with correlated attributes, both of which are relevant in plant leaf classification.

Bhatt et al. [9] demonstrated deep ensemble learning by combining MobileNetV2, InceptionV3, and ResNet50 for automatic medicinal leaf identification across 30 species, achieving high validation accuracy. While deep learning delivers strong results, the computational overhead and data dependency justify the continued importance of hybrid classical ML approaches.

Karnan and Ragupathy [10] surveyed plant classification using ML models, systematically cataloguing feature extraction techniques, publicly available datasets, and classifier performance across multiple studies. Their survey confirmed that no single classifier dominates universally, performance varies with feature sets, dataset characteristics, and species diversity. This observation motivates the multi-classifier comparative framework adopted in the present work.

III. WORKFLOW FOR MEDICINAL PLANT LEAF CLASSIFICATION

This work follows a structured six-stage pipeline as illustrated in Figure 1. Each stage progressively refines raw leaf image data into discriminative feature representations for ML classification.

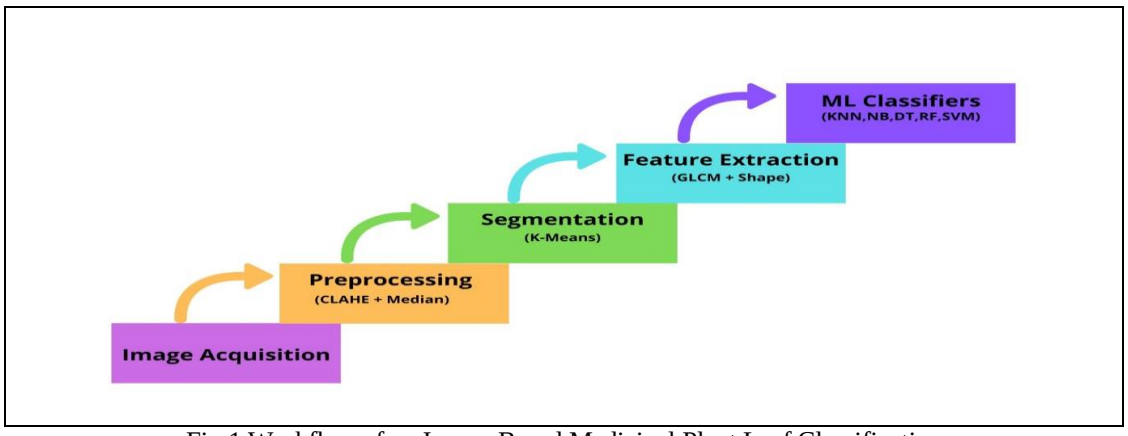

Fig 1 Workflow of an Image-Based Medicinal Plant Leaf Classification

➢ *Dataset*

The Medicinal Leaf Dataset published by Roopashree S. on Mendeley Data [10] is used in this study. The dataset comprises 1,800 high-resolution leaf images spanning 30 medicinal plant species, with 60 images per class. Images were captured under controlled laboratory conditions against a uniform white background. The dataset is split into 1,260 training images and 540 test images.

➢ *Image Processing*

Leaf images are scaled uniformly to 256×256 pixels through bilinear interpolation technique. RGB image is then segmented into HSV color space while converting to gray scale for extracting features related to texture and morphology. Median Filter with a kernel of 5×5 removes salt & pepper noise while retaining edge information. CLAHE with clip limit of 2.0 and a tile grid of 8×8 improves contrast by compensating for non-uniform illumination [11].

➢ *Leaf Region Segmentation*

K-means clustering is used with k=3 in HSV color space in order to classify between leaf foreground and background objects. The cluster containing leaf pixels is obtained through mean values of hue and saturation. The noise present in the binary image mask is removed using morphological opening and closing operations with a 3×3 elliptical structural element [12].

➢ *Feature Extraction*

- *Features for GLCM Textures:*

The Gray Level Co-occurrence Matrix is obtained in four orientations: 0, 45, 90, and 135 degrees, considering pixel distance of 1. Six features per each orientation are derived: Energy, Contrast, Correlation, Homogeneity, Entropy, and Dissimilarity. The average value for each of these features per orientation provides a 6D texture vector.

- *Geometric Shape Features:*

Eight parameters are calculated from the leaf mask: leaf area, perimeter, aspect ratio, eccentricity, solidity, extent, compactness, and minimum area of surrounding rectangle.

The overall feature vector is formed by concatenating the above-mentioned texture vector and shape parameters, resulting in 14 dimensions. Normalization of all features into [0,1] range is performed via min-max normalization.

➢ *Machine Learning Classifiers*

- *K-Nearest Neighbors (KNN):*

KNN was configured with k=7 neighbors and Euclidean distance. During testing, each of the 540 test samples is classified by identifying its seven closest training neighbors and assigning the majority class. KNN recorded 84.20% accuracy on this dataset. The performance is hampered by the

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26apr1643

fact that several shape descriptors, particularly extent and compactness, carry limited discriminative signal for certain species pairs, weakening the quality of nearest-neighbor assignments in the 14-dimensional feature space.

- *Naïve Bayes (NB):*

Gaussian Naive Bayes estimates the class-conditional distribution for each feature separately, assuming all features are mutually independent given the class. Training on 1,260 samples, it attained the lowest accuracy of 79.60% when tested on the 540-sample test set. This outcome is traceable to the correlations that exist among GLCM features — Energy and Homogeneity, for instance, tend to vary together — which cause the model to overcount shared information and generate unreliable posterior probability estimates for visually similar species.

- *Decision Tree (DT):*

A CART tree with Gini impurity criterion was grown to a maximum depth of 15 on the 1,260 training images and evaluated on the 540 test images, achieving 86.50% accuracy. The tree depth constraint partially controlled overfitting, but the single-tree structure still struggled to generalize for species pairs sharing very similar leaf geometry, such as Ocimum sanctum and Ocimum basilicum, where small training-time irregularities produced decision boundaries that did not hold on unseen test samples.

- *Random Forest (RF):*

An ensemble of 200 decision trees was trained, each on a bootstrap sample from the 1,260 training images, with sqrt(14) ≈ 4 features randomly sampled at every split. Test predictions on 540 images were determined by majority vote. Random Forest yielded the highest accuracy of 93.80%, alongside precision of 92.60%, recall of 93.10%, and F1-score of 92.85%. The two sources of randomness, bootstrap sampling and feature subsampling, decorrelated the individual trees and enabled the ensemble to achieve robust, well-balanced performance across all 30 species, including both large-leaf species such as Aloe vera and fine-leaf species such as Spearmint [8].

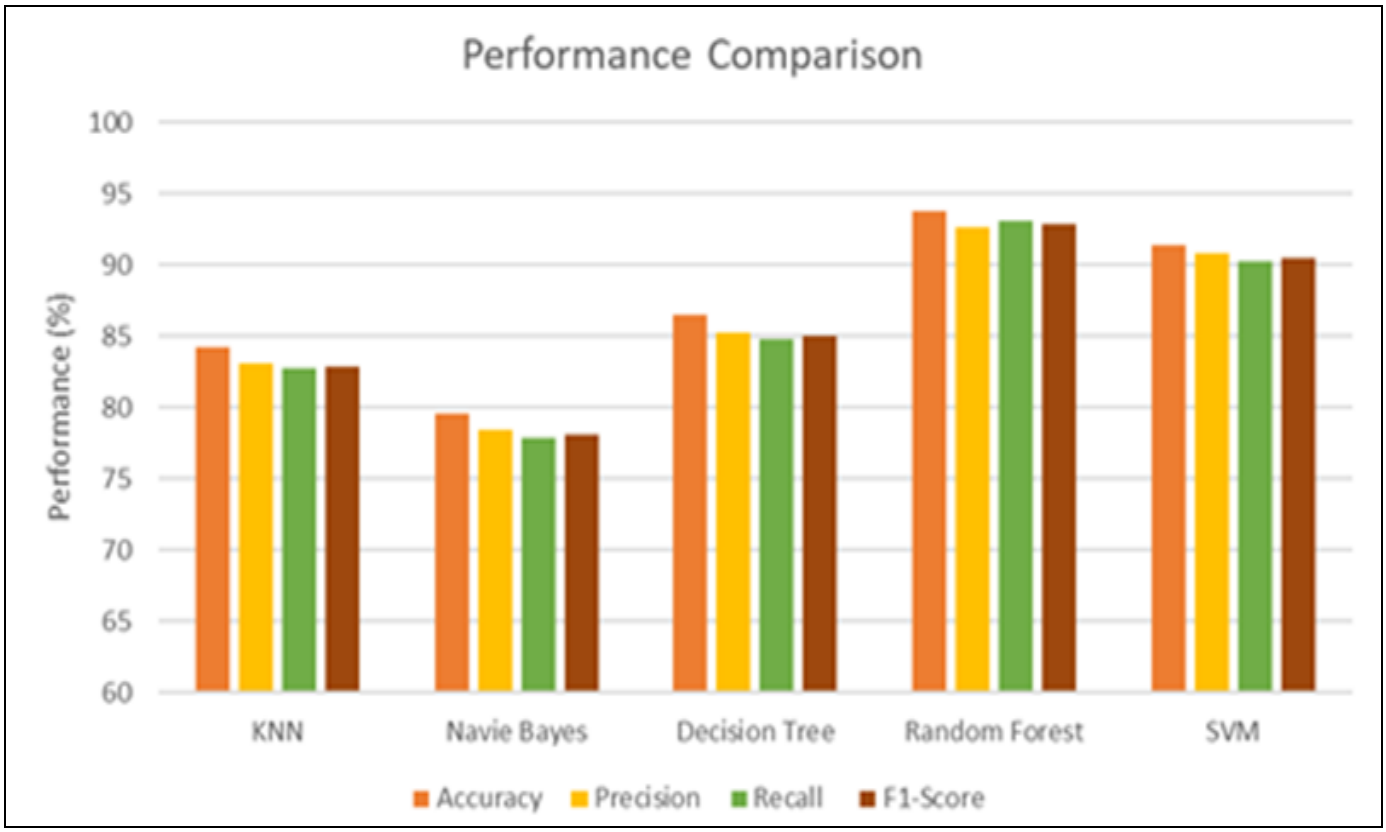

Fig 2 Classifier Performance Comparison

- *Support Vector Machines (SVM):*

SVM with an RBF kernel (C=10, gamma='scale') was trained using one-vs-one decomposition across 30 classes, generating 435 binary sub-classifiers. Test-time class assignment for each of the 540 test samples was determined by aggregating votes from all binary classifiers. SVM achieved 91.40% accuracy, ranking second overall. The RBF kernel effectively mapped the 14-dimensional feature space into a higher-dimensional representation where species clusters are more linearly separable, with the margin-maximization objective providing strong generalization from the 1,260 training samples. Figure 2 visualises the comparative performance across all three evaluation metrics.

IV. RESULT AND ANALYSIS

Random Forest achieved the highest classification accuracy of 93.80%. Its ensemble nature aggregates predictions from 200 independently trained decision trees, effectively reducing variance and correcting individual tree errors.

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26apr1643

Table 1 Table Styles

Classifier	Accuracy (%)	Precision (%)	Recall (%)	F1-Score(%)
KNN	84.20	83.10	82.70	82.90
Naïve Bayes	79.60	78.40	77.90	78.10
Decision Tree	86.50	85.20	84.80	85.00
Random Forest	93.80	92.60	93.10	92.85
SVM	91.40	90.80	90.20	90.50

The decorrelated tree structure, enforced by random feature subsampling at each split, makes RF particularly resilient to the inter-feature correlations present in fused GLCM-shape vectors. SVM ranked second at 91.40% accuracy. The RBF kernel projects the 14-dimensional feature space into a higher-dimensional space, enabling non-linear decision boundaries that effectively separate the 30 plant species. Table 1 presents the performance of all five classifiers.

V. CONCLUSION

This work presented a systematic comparative analysis of five machine learning classifiers, KNN, Naive Bayes, Decision Tree, Random Forest, and SVM, for medicinal plant leaf identification using an image-processing pipeline. The pipeline combined CLAHE contrast enhancement, Median filtering noise removal, K-means segmentation, and a fused 14-dimensional GLCM-shape feature vector, applied to the Medicinal Leaf Dataset of 30 species and 1,800 images. Experimental results established that Random Forest is the most effective classifier for this task, achieving high accuracy. SVM and Naive Bayes produced the lowest performance due to violated independence assumptions. The study contributes a reproducible comparative benchmark and provides practical guidance for classifier selection in resource-constrained medicinal plant recognition systems.

Future work will explore integrating additional spectral and vein-pattern features, incorporating deep learning-based feature extractors to replace handcrafted features, and expanding the dataset to include endemic South Indian and Ayurvedic medicinal species to improve regional applicability.

REFERENCES

[1]. Mulugeta, A. K., Sharma, D. P., & Mesfin, A. H. (2024). Deep learning for medicinal plant species classification and recognition: a systematic review. Frontiers in Plant Science, 14, 1286088.
[2]. Hussain, S., et al. (2021). The Classification of Medicinal Plant Leaves Based on Multispectral and Texture Feature Using Machine Learning Approach. Agronomy, 11(2), 263.
[3]. Prabha, B., & Kavitha, K. (2025). Deep Learning Based Medicinal Plants Identification Using CNN Architecture. In: Proceedings of ICMMCS 2025. Lecture Notes in Networks and Systems, vol 1400. Springer, Cham.
[4]. Wang, B., et al. (2024). AI-Driven Pattern Recognition in Medicinal Plants: A Comprehensive Review. Computers, Materials & Continua, 81(2).
[5]. Kan, H. X., Jin, L., & Zhou, F. L. (2017). Classification of Medicinal Plant Leaf Image Based on Multi-feature Extraction. Pattern Recognition and Image Analysis, 27(3), 581–587.
[6]. Kaur, P. P., & Singh, S. (2021). Classification of Herbal Plant and Comparative Analysis of SVM and KNN Classifier Models on the Leaf Features Using Machine Learning. In: Soft Computing for Intelligent Systems. Springer, Singapore.
[7]. Kaur, P. P., Singh, S., et al. (2022). Random Forest Classifier Used for Modelling and Classification of Herbal Plants. In: Mobile Radio Communications and 5G Networks. Springer, Singapore.
[8]. Bhatt, D., et al. (2022). Deep ensemble learning for automatic medicinal leaf identification. International Journal of Information Technology, 14(6), 3089–3097.
[9]. Karnan, A., & Ragupathy, R. (2024). A Comprehensive Study on Plant Classification Using Machine Learning Models. In: ICT: Smart Systems and Technologies. ICTCS 2023. Springer, Singapore.
[10]. Roopashree, S. (2020). Medicinal Leaf Dataset. Mendeley Data.
[11]. Jebadass, J. R., & Balasubramaniam, P. (2023). Preprocessing of leaf images using brightness preserving dynamic fuzzy histogram equalization technique. International Journal of Artificial Intelligence, 12(3), 1149–1157.
[12]. Sheth, V., et al. (2022). A Comparative Analysis of Machine Learning Algorithms. Procedia Computer Science, 215, 422–431.