A Robust Deep Learning Framework for Automated Medicinal Plant Identification Through Leaf Image Analysis

Sophia Lobo

Assistant Professor , Department of Computer Science

St Joseph's First Grade College, Jayalakshmipuram, Mysuru, Karnataka.

Abstract- *The primary objective of this research is to develop a robust Medicinal Plant Classification System by leveraging the Deep Learning concept. Accurate identification and classification of these plants are critical for advancing pharmacological applications and ensuring effective treatment protocols.*

The system was built upon a proprietary dataset comprising 1,500 images gathered across 30 distinct Indian medicinal plant species, with each species represented by approximately 50 to 60 images. The feature set utilized for classification incorporates the plants' leaf texture, shape, color, and key physiological/morphological characteristics.

A Convolutional Neural Network (CNN) architecture was employed to train the data and achieve a highly accurate classification model. Through training several model architectures, the best-performing iteration yielded a 96.67% success rate in identifying the correct medicinal plant species. This significantly high accuracy validates the model's potential as a valuable advisory and early warning tool for quick and reliable identification.

.***Key Words****-Deep learning, Convolutional Neural Network (CNN) algorithm, DenseNet architecture*

1. Introduction

The accurate classification of medicinal plants is a critical, interdisciplinary field that merges botany, data science, and medical research. By applying deep learning algorithms to plant images, researchers can extract complex features and patterns, enabling highly precise species identification and detailed analysis. This approach is essential for exploring the vast biodiversity of plant life and pinpointing potential sources of therapeutic compounds.

Medicinal plants—whether wild-collected or cultivated—have served as vital sources for healing remedies for millennia, contributing significantly to the current pharmacopeia. Given this extensive history and the hundreds of known classes, reliable identification is crucial. In recent years, computational methods, particularly in the domain of image processing, have revolutionized plant classification. Among these novel techniques, Neural Networks stand out, with the Convolutional Neural Network (CNN) being the most widely adopted architecture for image-based tasks.

This project focuses on leveraging the capabilities of deep learning to create robust models for medicinal plant classification. Our goal is to accurately categorize plant species based purely on their visual characteristics. By harnessing this potential, we aim to deepen the scientific understanding of medicinal plants, accelerating the discovery of natural remedies and ultimately leading to improved healthcare outcomes.

2. Existing System

The existing system for automatic identification of medicinal plants was developed to streamline the crucial identification step in Ayurvedic medicine production, a process traditionally performed manually since Vedic times. This system employs a classic computer vision and machine learning pipeline: input plant images are initially processed by converting them to grayscale and applying edge detection for optimization. Subsequently, a wide range of features—including color, geometrical properties, and texture (extracted through a specialized Fuzzy C-means approach and selective scale and orientation filters to preserve frequency details)—are extracted. These features are then used to train an ensemble model, combining the Random Forest algorithm with a Convolutional Neural Network (CNN), which is extended into a multinomial logistic regression model to handle the multiple plant classes. However, the system's effectiveness is severely hampered by its resulting low accuracy, making the medicinal plant identification difficult and unreliable. Furthermore, its performance is not robust to environmental variations, specifically failing to predict correctly for low-light plant images. While the automation was intended to replace expensive and repetitive manual searching, the current limitations mean the system does not yet provide a reliable, scalable alternative for the essential task of correct plant identification.

3. Related Works

In this section, we present a highlight of the existing research relevant to the study stated in this paper.

1] Surahleen Kaur's et.al work investigates automatic plant species identification using Computer Vision and Machine Learning (ML) applied to leaf images, stressing the value of technology in botanical knowledge. The study's methodology focuses on image pre-processing and the extraction of texture and colour features, demonstrating their superiority over shape features for this task. Utilizing the Swedish dataset of 15 plant species, the research employs statistical methods to characterize these features, which are then used to train and test a multiclass Support Vector Machine (SVM) classifier. The ultimate goal of this research is to enhance the accuracy of plant identification systems, with future work proposed to include a wider variety of features and explore more advanced or hybrid classifier approaches.

2] N Duong-Trung's et.al research, developed in Vietnam, focuses on the classification of medicinal plants using Deep Learning and Transfer Learning techniques. The study emphasizes the critical role of medicinal plants in traditional medicine, noting Vietnam's extensive flora, which includes approximately 3,780 species with medicinal properties out of 10,500 identified species. Utilizing a dataset of approximately 2,300 samples collected from the Botanic Garden of Tay Do University (featuring species like *Cleistocalyx Operculatus*, *Polyscias Fruiticosa*, and *Cassia Alata L.*), the authors leveraged transfer learning with state-of-the-art deep networks to overcome the challenges of training machine learning models from scratch. This approach yielded a high classification accuracy of 98.7% during testing, demonstrating the effectiveness of the proposed combination, though the actual implementation accuracy was noted to be 93.2%.

3] Raisa Akter et.al work focuses on plant classification and medicinal plant identification from leaf images using deep learning techniques, specifically introducing an attention architecture for more effective feature extraction. The study utilized a proprietary dataset of Bangladeshi medicinal plants, which underwent pre-processing before being classified using Convolutional Neural Network (CNN) models. While achieving an identification rate of 71.3%, the authors emphasized the need to explore advanced CNN models to improve performance across different image types and plan to expand their dataset. Furthermore, the long-term goal of the research is the development of a mobile application based on their trained model to facilitate practical identification.

4] Manik et.al research details a methodology for plant classification that relies on the k-Nearest Neighbour (k-NN) algorithm and Gray Level Co-occurrence Matrix (GLCM) feature extraction. The process began with Data Collection, where the plant image data was partitioned into training and test sets. To extract features, the GLCM was used to represent the spatial relationship between pixels in different directions and distances. The model's training involved using the k-NN algorithm to classify new objects based on the closest training data points. For robust validation, k-fold cross-validation was employed for Data Sharing, ensuring comprehensive use of the data for both training and testing subsets. Finally, the Evaluation of the Model was conducted by calculating accuracy using a confusion matrix, which incorporates true positive, false positive, true negative, and false negative predictions.

5] Gokhale A et.al work details a leaf image classification study where Feature Extraction was performed on pre-processed images, focusing on both colour (mean and standard deviations of R, G, and B channels) and texture features (including contrast, correlation, inverse difference, and entropy). The study compared four machine learning algorithms—k-Nearest Neighbour (KNN), Logistic Regression, Naïve Bayes, and Support Vector Machine (SVM)—for classification. Among these Classifier Algorithms, KNN yielded the highest accuracy of 79.49%. The Performance Assessment of the top-performing KNN classifier was rigorously evaluated using metrics like precision, recall, and F1-score for each class, reporting overall average scores of 0.814, 0.813, and 0.805 for precision, recall, and F1-score, respectively.

6] R. Geerthana's et.al paper explores the classification of Indian medicinal plant species using Convolutional Neural Networks (CNN), focusing on features derived from leaf texture, shape, and colour. The research utilized a substantial dataset of 58,280 images distributed across five plant species, with roughly 10,000 images representing each species. The developed CNN model demonstrated high efficacy by achieving a significant success rate of 96.67%, thereby establishing the strong potential of deep learning techniques for the accurate identification of medicinal plants to aid in both treatment and conservation efforts.

7] Panni Lage Sanduni Tharaka Perera's et.al paper details the development of an automated plant identification system specifically for Sri Lankan medicinal plants using leaf images and image processing techniques. The methodology covered steps including image acquisition, pre-processing, and feature extraction. The system utilizes a dataset of 33 different plant species, with 40 to 70 images collected for each leaf. For classification and pattern recognition, the study implemented a Probabilistic Neural Network (PNN) technique, which successfully achieved an overall identification accuracy of 90%, underscoring the system's effectiveness for medicinal plant identification.

8] MA Widneh's et.al study addressed the classification of medicinal plant species by leveraging Convolutional Neural Network (CNN) algorithms, specifically employing the MobileNet and VGG16 architectures. The research utilized a substantial dataset comprising 52 medicinal plant species, with 300 images collected for each species, resulting in a total of 15,100 medicinal plant images. The application of the MobileNet and VGG16 models for classification proved effective, achieving a classification accuracy of 90%.

9] Owais A. Malik et.al project details the development of a real-time system for species identification of medicinal plants in the Borneo region using deep learning models. The study utilized a comprehensive dataset of over 25,000 images sourced from both open-access and local datasets to train and test these models, tackling common challenges such as small training samples and class imbalance. The system, which comprises a plant species classifier, a knowledge base, and a mobile front-end, demonstrated significant accuracy improvements over baseline models in laboratory testing, although a slight accuracy drop was observed during real-time implementation. A key innovation of this work is the integration of a unique crowdsourcing feedback mechanism and geo-mapping feature for plant species, which facilitates the continuous learning and improvement of the deep learning models within the region.[1]

10] Jafer Abdollahi et.al paper, "Identification of medicinal plants in Ardabil using deep learning," documents the successful application of Transfer Learning for plant classification. The study utilized the pre-trained MobileNetV2 Convolutional Neural Network (CNN) architecture. A custom medicinal plant dataset was compiled, consisting of 30 different plant classes with a total of 3,000 images. By leveraging the pre-trained weights of the model, the developed system was evaluated on a held-out test set and achieved a notably high classification accuracy of 98.05 percent.

11] Vina Ayumi et.al developed an Artificial Neural Network (ANN)-based classification system for medicinal plant species using leaf colour, texture, and shape features. The dataset used comprised 1,500 images of thirty species, with the data split for training and validation at an 80% to 20% ratio. Specifically, the study allocated 36 leaves for training, seven for validation, and twenty for testing, where the final testing set contained 43 images of the 30 species (1-3 images per class). The research found that the MobileNetV2 architecture, when subjected to finetuning, yielded the best performance, achieving a high validation .

12] Silky Sachar et.al research, focusing on the therapeutic value of medicinal plants, introduced the Ensemble Deep Learning-Automatic Medicinal Leaf Identification (EDLAMLI) classifier for plant identification, predominantly using leaf images despite the potential for using other parts like root, fruit, bark, and stem. The study utilized a dataset consisting of 30 classes of medicinal leaves. The EDLAMLI system leverages an ensemble of pre-trained Neural Networks—specifically MobileNetV2, InceptionV3, and ResNet50—by employing a weighted average of their outputs for final classification. This sophisticated approach achieved a high accuracy of 99.66% on the test set, with the robust validation using threefold and fivefold cross-validation methods yielding an even higher average accuracy of 99.9%.

4. Proposed System

The proposed paper addresses the challenge of medicinal plant classification by developing a system based on a Convolutional Neural Network (CNN) algorithm. The core of the system utilizes the DenseNet architecture, which was specifically selected for its efficient memory usage and enhanced feature reuse capabilities, making it highly effective for complex image classification tasks. The system was trained and validated using a dataset consisting of 1,500 images that encompass 30 distinct medicinal plant species..

The methodology employed involved splitting the dataset into training and testing subsets, allocated at an 80% and 20% ratio, respectively. To further improve the model's generalization and robustness, data augmentation techniques were strategically applied to the training data. This combination of a robust deep learning architecture (DenseNet), a substantial image dataset, and rigorous pre-processing steps aims to establish an accurate and efficient automated solution for identifying medicinal plant species.

4.1. Advantages of Proposed System

- Enables easy and rapid identification of plant species
- Delivers results with high accuracy, enhancing confidence in species knowledge for research, conservation, or medicinal applications.
- Features a very fast processing speed, leading to quicker results compared to existing, potentially manual or slower, automated systems.

4.2. Architecture Diagram

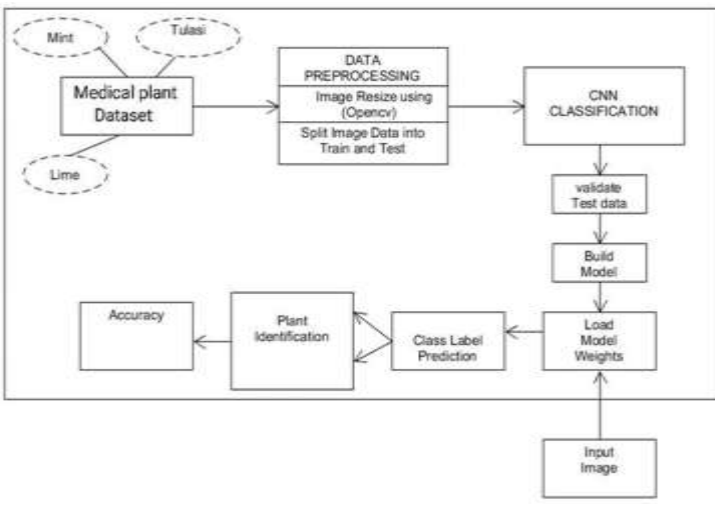

Fig.4.2 Architecture Diagram.

4.2.1. Dataset Acquisition and Structure

The data acquisition phase will involve sourcing relevant datasets from Kaggle.com. These datasets are characterized by having a large number of distinct classes for the predictive model.

4.2.2. Data Pre-processing

The initial phase of data pre-processing will involve three key steps applied to the selected image dataset:

- **Image Pre-processing Techniques:** Various image manipulation techniques will be performed on the raw data to standardize and enhance image quality.
- **Image Resizing:** All images will be resized to a uniform dimension to ensure consistency for the deep learning model input.
- **Data Splitting:** The pre-processed dataset will be divided into training and testing subsets for model development and evaluation, respectively.

4.2.3. Data Modeling and Evaluation

- **Model Training (CNN):** The pre-processed and split training data is utilized as input to the Convolutional Neural Network (CNN) algorithm to perform the model training.
- **Model Evaluation:** The performance of the trained model is assessed by passing the held-out test data to the algorithm.
- **Accuracy Calculation:** The final performance metric, **accuracy**, is calculated based on the model's predictions on the test set.

4.2.4. Model Building and Saving

Once the training phase is complete and the model demonstrates a high accuracy rate, the final step is to build and save the trained model file (serialization). This file is the deployable artifact that can be used for real-time predictions without needing to retrain the model.

5.Conclusion

This study utilized Convolutional Neural Network (CNN) and deep learning methods, specifically the DenseNet architecture , for the feature extraction and classification of medicinal plant species. The approach employed a dedicated Medicinal Plant dataset for both training and testing, and the initial image inputs were small and grayscale. The model achieved a high classification accuracy of 98.67%, with robust performance validated by nearly equal training and validation accuracies, suggesting minimal overfitting. Comprehensive evaluation included standard performance metrics such as precision, recall, F1-score, and support.

The current CNN model's utility is limited, as the conclusion highlights that it is not generic for all plant species. Despite the initial statement of four classes, the model was ultimately developed to classify a total of 30 distinct plant species. These include: *Alpinia Galanga* (Rasna), *Amaranthus Viridis* (Arive-Dantu), *Artocarpus Heterophyllus* (Jackfruit), *Azadirachta*

Indica (Neem), *Basella Alba* (Basale), *Brassica Juncea* (Indian Mustard), *Carissa Carandas* (Karanda), *Citrus Limon* (Lemon), *Ficus Auriculata* (Roxburgh fig), *Ficus Religiosa* (Peepal Tree), *Hibiscus Rosa-sinensis*, *Jasminum* (Jasmine), *Mangifera Indica* (Mango), *Mentha* (Mint), *Moringa Oleifera* (Drumstick), *Muntingia Calabura* (Jamaica Cherry-Gasagase), *Murraya Koenigii* (Curry), *Nerium Oleander* (Oleander), *Nyctanthes Arbor-tristis* (Parijata), *Ocimum Tenuiflorum* (Tulsi), *Piper Betle* (Betel), *Plectranthus Amboinicus* (Mexican Mint), *Pongamia Pinnata* (Indian Beech), *Psidium Guajava* (Guava), *Punica Granatum* (Pomegranate), *Santalum Album* (Sandalwood), *Syzygium Cumini* (Jamun), *Syzygium Jambos* (Rose Apple), *Tabernaemontana Divaricata* (Crape Jasmine), and *Trigonella Foenum-graecum* (Fenugreek).

6. Future Enhancement

The current model is limited to classifying 30 medicinal plant species. Our future work is focused on expanding the system to handle an increased number of classes while maintaining high accuracy across heterogeneous data. This necessitates modifying the current network architecture; however, a critical challenge is to improve performance without increasing the model depth, as excessive depth can lead to the well-known overfitting problem often observed in DenseNet architectures. Careful consideration will be given to architectural changes to ensure both generalization and accuracy. Additionally, a crucial future step involves upgrading the model's input from grayscale to color image classification to leverage richer spectral features for enhanced medicinal plant recognition.

7. References

1] Kaur, S., & Kaur, P. (2019). Plant species identification based on plant leaf using computer vision and machine learning techniques. Journal of Multimedia Information System, 6(2), 49-60.

2] Duong-Trung, N., Quach, L. D., Nguyen, M. H., & Nguyen, C. N. (2019, February). A combination of transfer learning and deep learning for medicinal plant classification. In Proceedings of the 2019 4th International Conference on Intelligent Information Technology (pp. 83-90).

3] Akter, R., & Hosen, M. I. (2020, December). CNN-based leaf image classification for Bangladeshi medicinal plant recognition. In 2020 Emerging Technology in Computing, Communication and Electronics (ETCCE) (pp. 1-6). IEEE.

4] Manik, F. Y., Saputra, S. K., & Ginting, D. S. B. (2020, June). Plant classification based on extraction feature gray level co-occurrence matrix using k-nearest neighbour. In Journal of Physics: Conference Series (Vol. 1566, No. 1, p. 012107). IOP Publishing.

5] Gokhale, A., Babar, S., Gawade, S., & Jadhav, S. (2020). Identification of medicinal plant using image processing and machine learning. In Applied Computer Vision and Image Processing: Proceedings of ICCET 2020, Volume 1 (pp. 272-282). Singapore: Springer Singapore.

6] Geerthana, R., Nandhini, P., & Suriyakala, R. (2021). Medicinal plant identification using deep learning. International Research Journal on Advanced Science Hub, 3, 48-53.

7] Perera, P. S. T., & Arudchelvam, T. (2021, April). Leaf Based Plant Identification System for Sri Lankan Medicinal Plant. In Future of Information and Communication Conference (pp. 831-836). Cham: Springer International Publishing.

8] : Widneh, M. A., Workneh, A. T., & Alemu, A. A. (2021). Medicinal plant parts identification and classification using deep learning based on multi label categories. Ethiopian Journal of Science and Sustainable Development, 8(2), 96-108.

9] Malik, O. A., Ismail, N., Hussein, B. R., & Yahya, U. (2022). Automated real-time identification of medicinal plants species in natural environment using deep learning models—a case study from Borneo Region. Plants, 11(15), 1952.

10] Abdollahi, J. (2022, February). Identification of medicinal plants in Ardabil using deep learning: Identification of medicinal plants using deep learning. In 2022 27th International computer conference, computer society of Iran (CSICC) (pp. 1-6). IEEE.

11] Ayumi, V., Ermatita, E., Abdiansah, A., Noprisson, H., Jumaryadi, Y., Purba, M., ... & Putra, E. D. (2022). Transfer learning for medicinal plant leaves recognition: a comparison with and without a fine-tuning strategy. International Journal of Advanced Computer Science and Applications, 13(9).

12] : Sachar, S., & Kumar, A. (2022). Deep ensemble learning for automatic medicinal leaf identification. International Journal of Information Technology, 14(6), 3089-3097.