Identification of Different Medicinal Plants/Raw materials through Image Processing Using Machine Learning Algorithms

B MEENU[1], V DEEKSHITHA[2], SAI GADDAM LIKHITHA3, AISHWARYA VILAS PATIL[4], PROF. AFROZ PASHA[5]
[1, 2, 3, 4, 5]*Department of Computer Science and Engineering, Presidency University*

Abstract- In an era where nature's pharmacy brims with untapped potential, swiftly and accurately identifying medicinal plants is a transformative leap for healthcare, biodiversity, and sustainable innovation. This project unveils a groundbreaking hybrid approach, melding image processing with cutting-edge machine learning—fusing Conditional Generative Adversarial Networks (CGANs), Wasserstein GANs (WGANs), Deep Convolutional GANs (DCGANs), and the formidable VGG16 model, topped with logistic regression—to revolutionize plant recognition. Picture an AI botanist with a creative twist, not just spotting plants from snapshots but conjuring synthetic images to sharpen its skills, built on a rich, curated dataset of medicinal wonders amplified by sophisticated augmentation to echo real-world diversity. Our method weaves through data loading, preprocessing, and feature extraction, birthing two stellar models: Model 1 (CGAN with Logistic Regression), scoring a robust AUC-ROC of 0.94, and Model 2 (WGAN and DCGAN with Logistic Regression), rocketing to an awe-inspiring 0.99, flaunting near-perfect classification across 40 species across 40 species. Encased in a sleek web application, this tech marvel bridges complex AI with everyday use, its success vividly etched in confusion matrices and precision-recall metrics, with Model 2 shining in stability and accuracy—a bold stride toward preserving herbal wisdom, fueling research, and championing Sustainable Development Goals, where pixels and plants collide to spark a blooming future of botanical brilliance.

I. INTRODUCTION

The identification and classification of medicinal plants and raw materials play a crucial role in herbal medicine, pharmaceuticals, and the broader healthcare industry. Traditionally, plant identification has relied on botanical expertise, which can be time-consuming and prone to human error. However, with advancements in artificial intelligence, machine learning, and particularly image processing techniques, it is now possible to automate and enhance the accuracy of plant identification. By leveraging deep learning models, Generative Adversarial Networks (GANs), and the broader capabilities of Generative AI (Gen AI), a robust framework can be developed to recognize medicinal plants through image analysis with greater precision and efficiency.

In the context of medicinal plant identification, Gen AI enables the creation of high-quality synthetic images that simulate real plant appearances, which is especially useful when data availability is limited or class imbalance affects model training. This not only enhances model performance but also supports the development of more generalized and scalable identification systems. With the evolution of deep learning and Convolutional Neural Networks (CNNs), automated feature extraction has become more reliable than traditional image recognition techniques. Models like VGG16 offer a pre-trained framework that effectively learns intricate visual patterns from plant images, while GAN variants such as Conditional GANs (CGANs) and Wasserstein GANs (WGANs) address data limitations and improve image quality. CGANs enhance the training process by generating category-specific images through label conditioning, and WGANs stabilize the training process by using the Wasserstein distance metric, resulting in more realistic synthetic outputs. When combined with the transformative capabilities of Gen AI, including advanced GAN architectures and diffusion models, these techniques collectively elevate the quality and diversity of training datasets, overcoming environmental variability, class imbalance, and generalization issues that are common in real-world applications.

II. LITERATURE REVIEW

Praveen Kumar et al. [1] proposed PSR-LeafNet, a deep learning framework for classifying medicinal plant leaves using three subnetworks (P-Net, S-Net, R-Net) to extract shape, color, venation, and texture features, followed by classification with an SVM. The model employs the MRMR method for effective feature selection and achieved high accuracy on the MalayaKew (97.12%), IMP (98.10%), and Flavia (95.88%) datasets. Its strength lies in handling large-scale data and improving accuracy through multi-network integration. However, reliance on leaf images and limited model interpretability pose challenges. Future work may include multimodal data and explainable AI for better transparency.

In a focused exploration of Ethiopian indigenous medicinal plants, M.A. Kiflie et al. [2] leveraged transfer learning with pre-trained convolutional neural networks, including VGG16, VGG19, Inception-V3, and Xception, to enhance classification performance. On a dataset of 1,853 images from 35 species, VGG19 achieved the highest accuracy (94%),followed by VGG16 (92%), Inception-V3 (91%), and Xception (87%). The approach automates plant identification, reducing reliance on manual methods and addressing limited labeled data. However, relying solely on leaf images may not capture all species variations, and deep models often lack interpretability. Future work could integrate features like flowers or use explainable AI for greater transparency.

Utilizing image processing and machine learning techniques, Vikaho Swu et al. [3] investigated plant classification through the analysis of leaf characteristics such as shape, texture, and color. They tested classifiers like Naïve Bayes, SVM, and Random Forest, with SVM showing the highest accuracy. Naïve Bayes, though efficient for small datasets, assumes feature independence, which may not suit complex plant data. Challenges included lighting variations and background noise affecting image quality. The study highlights machine learning” role in reducing manual effort and errors. Future work could incorporate deep learning and multimodal data to improve classification of medicinal plants.

A novel approach to plant leaf classification was proposed by P.S. Kanda et al. [4], who integrated a Conditional Generative Adversarial Network (cGAN) for data augmentation, a Convolutional Neural Network (CNN) for feature extraction, and a Logistic Regression model for classification. This combination achieved an impressive average accuracy of 96.1%, with certain datasets reaching up to 100%. A key strength of this technique is its ability to automate feature learning, reducing dependence on manually designed features. Additionally, cGAN effectively addresses class imbalance by generating high-quality synthetic leaf images. However, the approach faces challenges, including high computational demands and the risk of overfitting with limited training data. While CNNs excel at feature extraction, their lack of interpretability complicates understanding the reasoning behind classification results. Future advancements could focus on improving model transparency and incorporating diverse plant characteristics, such as flowers and stems, to enhance classification accuracy.

M.S.I Musafayya et al. [5] introduced a deep learning approach for classifying Indonesian medicinal plants using transfer learning with models like ResNet, DenseNet, VGG, ConvNeXt, and Swin Transformer. ConvNeXt achieved the highest accuracy at 92.5%, proving effective in feature extraction. The method automates plant identification, reducing expert dependency and benefiting from pre-trained models to handle limited labeled data. However, the study was limited by a small dataset (100 species) and few model options. Future work could expand the dataset, explore more architecture, and enhance model interpretability.

Kavitha et al. [6] developed a deep learning method for real-time medicinal plant identification using the MobileNet model, trained on six species from a Kaggle dataset. After preprocessing and augmentation, MobileNet achieved 98.3% accuracy, offering fast, efficient classification suitable for mobile applications. The key strength is its real-time deployment via a cloud-integrated app. However, the limited dataset restricts broader applicability, and image quality or complex plant features may affect accuracy. Still, the study highlights deep learning” promise in automating plant recognition and supporting digital taxonomy.

G. Kiran Kumar et al. [7] proposed a machine learning-based approach using CNNs, specifically MobileNetV2 with transfer learning, to automate medicinal plant identification. Their model, enhanced with image augmentation techniques, showed high accuracy while reducing manual effort. They also developed a real-time web app using Streamlit for interactive classification. However, limited dataset diversity and external factors like lighting and plant similarities impacted performance. Still, the study provides a strong foundation for automated plant recognition, with future improvements suggested through larger datasets and model refinement.

In the realm of Ayurvedic plant identification, Amey Sunil Deshmukh et al. [8] explored the integration of artificial intelligence and image processing. Their research highlights the challenges in recognizing medicinal plants due to variations in leaf shape, texture, and color. They examined various machine learning techniques, including KNN, ANN,PNN, and SVM, to classify plant species based on leaf features. The study emphasizes the advantages of artificial intelligence in automating plant recognition, making it accessible for medical and botanical applications. However, despite the effectiveness of AI models, the approach has limitations such as computational complexity, dependency on high-quality datasets, and challenges in distinguishing visually similar plant species. Their work contributes to the growing field of Ayurvedic plant identification, offering a foundation for further research in automated plant classification systems.

Biplob Dey et al. [9] explored automated medicinal plant identification using deep convolutional neural networks (DCNN). They evaluated seven advanced models on a dataset of 5,878 images across 30 species, with DenseNet20 achieving the highest accuracy—99.64% on public data and 97% on real-world images. DCNNs proved effective in identifying species with high precision, even across different families. However, challenges included image quality, complex backgrounds, similar-looking species, and high computational demands. Despite these, the study highlights the potential of AI in advancing plant identification and biodiversity research.

A novel approach integrating two-dimensional correlation spectroscopy (2DCOS) with a residual neural network (ResNet) was proposed by Lian Li et al. [10], who focused on differentiating medicinal plant raw materials based on drying techniques and geographical origins. Their study, which examined Eucommia ulmoides leaves (EULs),demonstrated outstanding accuracy, achieving 100% classification across training, testing, and validation phases. The advantage of this method lies in its ability to enhance spectral resolution and extract deep features, making it a reliable plant differentiation technique. However, the requirement for extensive spectral data processing and high computational costs may limit its application in resource-constrained environments. Nevertheless, this research provides a significant breakthrough in automated medicinal plant identification, offering a promising alternative to traditional classification methods.

III. METHODOLOGY

The methodology employs 2 models: GAN with logistic regression(M1) , WGAN and DCGAN with logistic regression(M2). The GAN's were fundamentally built on adversarial learning which involves 2 networks: generator and discriminator that are trained in competitive adversarial settings. The generator attempts to create realistic data while discriminator classifies if the generated sample is real or fake. The generator and discriminator engage in a minimax game where the generator tries to minimize the discriminator 's ability to distinguish between real and fake generated samples and the discriminator tries to maximize its ability to do so. The following flowchart shown below are the proposed flow:

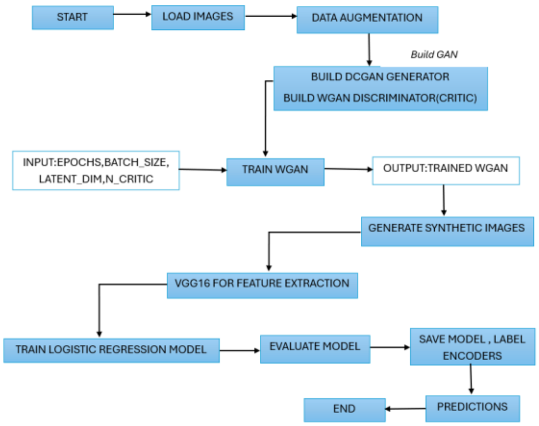

Fig 1:Proposed Workflow for Model 2

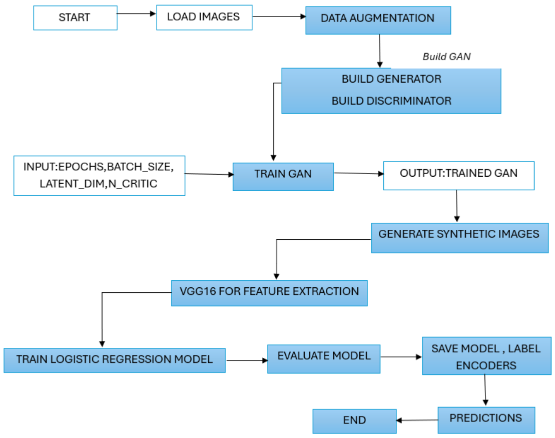

Fig 2:Proposed Workflow for Model 1

1. *Data collection*: The dataset titled "Segmented medicinal leaf images" comprises 30 different species of medicinal leaves. In our approach, the

model was trained on 20 species. Each species is represented by 50 samples which were utilized by GAN generator to produce an additional 250 samples. This augmentation of dataset was essential to enhance the overall size and diversity of data thereby supporting effective training of logistic regression.

2. *Data Augmentation*: It is a technique used in computer vision to artificially expand the size of the training dataset by creating modified versions of existing images. It is particularly useful when the available dataset is small, which helps the model ability to generalize new, unseen data.

Table 1:Techniques Used in Data Augmentation

Parameter	Description	Value
rotation_range	Creating new training images by rotating the existing ones	20
width_shift_range	It shifts images to left or right(horizontal shifts)	0.2
height_shift_range	It modifies the image height by shifting the pixels up or down within the image boundary	0.2
shear_range	It skews images along the x or y-axis to create different perspectives.	0.2
zoom_range	It allows the images to randomly zoom in or out of the specified value.	0.2

3. *Feature Extraction using VGG16*: The top classification layer is excluded when loading the VGG16 model from Keras. The images are pre-processed to meet the VGG16 input criteria, which include scaling to desired size and normalizing pixel values. Once the model has been imported and preprocessed, the features are retrieved by giving images to it. The result will be a collection of feature maps, which are multidimensional arrays representing the learnt features for each plant species. To employ feature maps in machine learning models, they must be reshaped into a 2D array with each row representing a flattened vector from an image.

4. *Training logistic regression*: The extracted features for each image in a feature vector serves as an input the logistic regression model. The target labels undergo label encoding to convert the categorical labels into a numerical format. The logistic regression model is fitted by passing the training features and labels. During each step the model learns the relationship between features and targets. The model evaluation for logistic regression will be discussed in the Results and discussion section.

5. *Predictions*: The input image is loaded and resized to the expected size(128,128)) to match the model input and converted to a numpy array. The pixel values are normalized by scaling from [0, 255] to [0, 1].The image dimensions are expanded by expanding the dimensions to match image shape as (1,128,128,3).The image tensor is fed to the VGG16 feature extractor which outputs a feature map that represents the important visual features of the input image. The feature map is flattened into a 1D vector which is used an input by the logistic regression which predicts a numerical value corresponding to the encoded class label.

IV. EXPERIMENTAL SETUP

1. For GAN with Logistic Regression(M1):This setup aims to train a class-conditional GAN to generate synthetic images for multiple classes, leveraging label conditioning for improved control over generated outputs.

Table 1:Model 1 Requirements

Component	Description
Generator input	Noise vector(latent_dim=100)+one-hot labels
Generator output	Synthetic Images(128x128x3 with tanh activation)
Discriminator input	Real or synthetic images
Discriminator output	Multi-class+real/fake classification(sigmoid)
Loss functions	Categorical cross-entropy for GAN, discriminator

Training steps	Alternate training of generator and discriminator
Batch size	32
Epochs	1000
Conditioning	Class label concatenation of conditional generation

2. For WGAN, DCGAN with Logistic Regression: This experimental setup effectively combines the strengths of WGAN's for generating synthetic images and logistic regression for classifying test images based on the extracted features.

Table 2:Model 2 Requirements

Component	Description
DCGAN Generator input	Input: 1. Latent vector+class labels 2. Dense reshape Layers to project latent space 3. Conv2D Transpose Layers for upsampling 4. Activation:ReLU,Batch Normalization for stability
DCGAN Generator Output	Synthetic image tanH activation
WGAN Discriminator input(Critic)	Input: 1. Images(real or fake) 2. Flatten layer followed by dense layers with LeakyReLU 3. Dropout for regularization
WGAN Discriminator output	Single score to indicate image quality(no activation)
Training process	1. Critic trained n_critic times per epoch to classify real and fake images. 2. Generator trained to fool the critic into classifying fake as real.
Epochs	1000
Batch Size	32

V. PERFORMANCE METRICS

Repeated Stratified K-fold cross validation is a robust technique used to evaluate logistic regression models, especially in multi-class classification scenarios with imbalanced class distributions. It involves splitting the data into k-folds ensuring each fold maintains the original class proportions and training and testing the model multiple times, averaging the results for a more reliable performance estimate. Model 1 and model 2 are evaluated based on the strategy mentioned above (refer section VII).

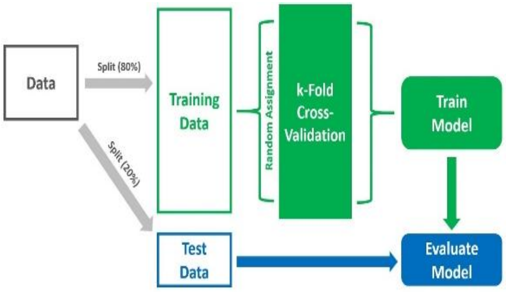

Fig 1: Repeated Stratified K-fold cross validation

VI. RESULTS AND DISCUSSIONS

The model is trained and validated 15 times(5 folds x 3 repeats).Each fold is used as validation set while the remaining is used as training sets. The accuracy scores for each fold were recorded(table 1), providing insights into the model's performance across different subsets of the data. The mean accuracy across all folds was calculated to summarize the model's overall performance. Model 2 outperformed model 1.

Table 3:Model Comparison based on Mean Accuracy

Model	Mean Accuracy	F1-Score
Model 1	0.9021	0.8923
Model 2	0.9885	0.9883

CONCLUSION

This project adequately illustrates the detection of various medicinal plants and raw materials by image processing with the aid of machine learning algorithms. Integrating data augmentation and synthetic image generation enhanced significantly the training data diversity and robustness, leading to its high performance for multi-class image classification tasks. With the employment of Generative Adversarial Networks (GAN, DCGAN, and WGAN) and logistic regression classifiers, the system has shown high performance for multi-class image classification. Of the two models that were tested, Model 2—WGAN and DCGAN with Logistic Regression—performed better than Model 1 (GAN with Logistic Regression), recording a staggering mean accuracy value of 0.99, while Model 1 was 0.94. These findings verify the effectiveness of GAN-based methods in enhancing classification accuracy and imply that the designed system has strong potential to be used for real-world applications in medicinal plant classification.

REFERENCES

[1] P. K. Sekharamantry, Dr. S. Rao, Y. Srinivas, and A. Uriti, "PSR-LeafNet: A Deep Learning Framework for Identifying Medicinal Plant Leaves Using Support Vector Machines" Big Data and Cognitive Computing, vol. 8, no. 12, p. 176, 2024.

[2] 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, p. 100987,2024.[Online].Available: https://doi.org/10.1016/j.jaim.2024.100987

[3] V. Swu, I. Kharir, and D. Bora, "Identification of Different Plants through Image Processing Using Different Machine Learning Algorithms" Sambodhi, vol. 43, pp. 172-179, 2020.

[4] P. S. Kanda, K. Xia, and O. H. Sanusi, "A deep learning-based recognition technique for plant leaf classification", IEEE Access, vol. 9, pp. 162590-162613, 2021.

[5] M. S. I. Musyaffa, N. Yudistira, M. A. Rahman, A. H. Basori, A. B. F. Mansur, and J. Batoro, "IndoHerb: Indonesia medicinal plants recognition using transfer learning and deep learning", Heliyon, vol. 10, no. 23, 2024.

[6] S. Kavitha, T. Satish Kumar, E. Naresh, V. H. Kalmani, K. D. Bamane, and P. K. Pareek, "Medicinal plant identification in real-time using deep learning model" SN Comput. Sci., vol. 5, no. 73, 2024. doi: 10.1007/s42979-023-02398-5.

[7] G. K. Kumar, A. Rama, D. Sri Charan Reddy, and R. Vinayak, "Medicinal plants identification through image processing and machine learning," 2025.

[8] S. Deshmukh, P. M. Mudhaliar, and S. Thorat,"Ayurvedic plant identification using image processing and artificial intelligence" Int. J. Sci. Res. Comput. Sci. Eng. Inf. Technol., vol. 7, pp. 212-218, 2021.

[9] 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. [Online].Available: https://doi.org/10.1016/j.heliyon.2023.e23655

[10] L. Li, Z. M. Li, and Y. Z. Wang, "A method of two-dimensional correlation spectroscopy combined with residual neural network for comparison and differentiation of medicinal plants raw materials superior to traditional machine learning: a case study on Eucommia ulmoides leaves" Plant Methods, vol. 18, no. 1, p. 102,.