International Journal of Computer Trends and Technology
ISSN: 2231–2803 / https://doi.org/10.14445/22312803/IJCTT-V73I4P123

Volume 73 Issue 4, 167-172, April 2025

Original Article

Identification of Medicinal Plants Using Machine Learning Algorithms

P. Sandeep Kumar[1], M. Rajeshwar[2], Giridhar Paida[3], Kodi Satheesh[4]

[1,2,3,4] *Department of ET, Hyderabad Institute of Technology and Management, Telangana, India.*

[1]*Corresponding Author : sandeepkumarpappula@gmail.com*

Received: 25 March 2025 Revised: 19 April 2025 Accepted: 24 April 2025 Published: 30 April 2025

Abstract *- Medicinal plants have long been a cornerstone of the pharmaceutical and healthcare industries due to their natural healing properties. However, conventional plant identification techniques are laborious and ineffective because they depend on specialized knowledge and are prone to mistakes. Automation using computer vision and Artificial Intelligence (AI) has become necessary as a result of this difficulty. Researchers have long struggled to identify and categorize plant species using subtle visual characteristics. By utilizing a Convolutional Neural Network (CNN) with the VGG16 architecture, this study seeks to automate the Identification of medicinal plants through machine learning. The Plants Type Dataset, which includes 30,000 plant photos divided into 30 plant classes spanning seven plant types—crops, fruits, industrial, medicinal, nuts, tubers, and vegetables—was utilized in this project. There are 1,000 photos in each class featuring common plants like bananas, coconuts, and pineapples and less common ones like galangal and bilimbi. The VGG16 model is the study's most successful machine learning algorithm because it can automatically extract visual features from the plant images, significantly improving classification performance. The findings show that deep learning methods—specifically, VGG16—can provide quicker and more precise plant identification. This strategy may transform drug development and botanical research while opening new avenues for discovering therapeutic plants. The results highlight how machine learning has the potential to revolutionize plant identification procedures, advancing scientific understanding and the use of medicinal plants.*

Keywords *- Convolutional Neural Networks (CNNs), Machine Learning, Plants Type Dataset, Support Vector Machines (SVMs),VGG16.*

1. Introduction

The Plants Type Dataset , curated by [1], the 30,000 plant images in the Plants Type Dataset provide a complete collection spanning seven plant types: crops, fruits, industrial, medicinal, nut, tubers, and vegetables broken into thirty plant classes. Training machine learning models can benefit much from this dataset, which also helps build automated plant identification systems. Regarding image classification, the VGG16 design has shown quite good performance. [2] presented VGG16, a deep convolutional neural network distinguished by depth and 3x3 convolution filter use. Their work underlined network depth's need to improve model accuracy and produced state-of-the-art ImageNet dataset results. With the release of AlexNet, a deep convolutional neural network that won first place in the ImageNet competition, [3] transformed image classification before VGG16. Their research opened the door for later developments in the field by proving the effectiveness of deep learning models in challenging image recognition tasks. As suggested by [4], Support Vector Machines (SVMs) provide a reliable method for handling classification issues. SVMs ensure high generalization ability by identifying the hyperplane that best divides data points of various classes. They are a mainstay in machine learning applications due to their efficacy and theoretical underpinnings. To increase classification accuracy, ensemble techniques like Random Forests, which were first presented by [5], combine several decision trees. Random Forests are appropriate for challenging classification tasks because they improve model robustness and minimize overfitting by averaging the predictions of individual trees. This literature review overviews past studies on machine learning-based plant recognition. The section on the Existing System emphasizes the current approaches and their drawbacks. The proposed system introduces a CNN-based framework and compares it with more traditional models such as Random Forest and SVM. The methodology explains the pre-processing, model training, and evaluation stages. The Results section compares the model's performance using standard metrics, and the conclusion presents the key findings. All cited sources are listed in the references.

2. Literature Survey

Deep learning has transformed image recognition over the last ten years, propelling notable advancements in various fields, including the automated Identification of plant species.

These developments have made it possible to create incredibly accurate models that can perform intricate visual tasks previously thought too complex for conventional machine learning algorithms. Deep learning techniques were revolutionized by the groundbreaking work of [6]. By tackling the vanishing gradient problem, one of the fundamental drawbacks of deep neural networks, their introduction of ResNet (Residual Networks) signaled a sea change. ResNet's use of residual connections made it possible to successfully train networks with hundreds of layers, greatly enhancing performance on image classification tasks and establishing new standards. Building on this framework, depthwise separable convolutions were introduced as a more effective substitute for conventional convolutional layers in [7] Xception architecture.

This design increased the network's ability to learn fine-grained visual features while increasing computational efficiency. Since Xception outperformed other models on a number of image classification tests, it has gained popularity in computer vision research, including for plant recognition applications.[8] carried out a comparative analysis utilizing a number of conventional machine learning approaches in the context of medicinal plant identification. Their research showed that when combined with carefully extracted features, algorithms such as Support Vector Machines (SVM) and Decision Trees could achieve a respectably high classification accuracy. However, their findings also demonstrated that deep learning models, which automatically, usually perform better than these traditional techniques when learning features from raw image data, especially in large and varied datasets. [9] thoroughly investigated the function of manually engineered features in plant classification, concentrating on leaf shape and texture descriptors.

Their method required a great deal of domain expertise and frequently struggled with changes in image conditions, such as lighting, background, and leaf orientation, even though it was only moderately successful in identifying plant species. These drawbacks emphasize the need for more flexible and data-driven techniques, like deep learning, which can automatically recognize intricate visual patterns without human assistance. [10] suggested an ensemble of Convolutional Neural Networks (CNNs) to address classification issues in the biomedical field. They increased accuracy and robustness by merging several CNNs trained on various data subsets or architectural variations. Its potential applicability to plant image classification tasks is suggested by the fact that this ensemble approach proved especially advantageous in domains with complex image data. Transfer learning was investigated by [11] to overcome the lack of data in plant recognition applications. They greatly enhanced model performance on small, domain-specific datasets by optimizing deep CNNs already trained on massive datasets like ImageNet. Their research demonstrated how transfer learning can achieve high accuracy even with little training data, which is a common problem in the field. [12] presented an early but fundamental method by creating a leaf classification algorithm based on Probabilistic Neural Networks (PNNs). Although more recent CNN-based models have overtaken the accuracy and scalability of this approach, it served as a vital foundation for studies on computational models for automated plant recognition.

Finally, [13] used a hybrid dataset augmentation technique to present a deep learning framework for classifying plant diseases. To increase the diversity and representativeness of training data—two factors essential for enhancing deep models' capacity for generalization their methodology integrated a number of augmentation techniques. Their research emphasizes the value of data preprocessing and augmentation, especially when datasets are sparse, unbalanced, or impacted by environmental variability, which frequently arises.

3. Existing System

Using manually created features like leaf shape and texture, traditional plant identification systems frequently rely on manual techniques or traditional machine learning models like Support Vector Machines (SVMs) and Decision Trees. Even though these methods offer a respectable level of accuracy, they necessitate specialized knowledge and have trouble generalizing across different datasets because of variations in lighting, background, and orientation. Models like AlexNet and VGG16, which automatically learn features from photos, have enhanced plant classification since deep learning became popular. However, these models frequently require high computational resources and sizable labeled datasets. Furthermore, most current systems are not designed with medicinal plant identification in mind, and they frequently have problems like overfitting and subpar performance on unseen data. Notwithstanding these developments, many current systems have a number of drawbacks:

- *Limited Dataset Diversity***:** Most models' capacity to generalize across various plant species and environmental conditions is restricted because they are trained on comparatively small or homogeneous datasets.
- *Overfitting:* Deep learning models trained on sparse data frequently exhibit overfitting, which causes them to perform well on training data but poorly on unseen samples.
- *Absence of Attention to Medicinal Plants:* Although many systems are designed to recognize general plant species, very few are tailored for medicinal plants, which frequently need more precise classification because of minute visual variations.
- *Single-Model Dependency:* Many systems only use one CNN architecture, which might not be strong enough to manage intricate differences in leaf photos.

4. Proposed System

A Convolutional Neural Network (CNN)-based method for automatically identifying medicinal plants from leaf photos is presented by the suggested system. In contrast to conventional techniques that depend on manually created features, this system uses deep learning to automatically extract intricate patterns, improving generalization and accuracy. To assess performance, we use and contrast a number of CNN architectures, such as VGG16, ResNet, and custom CNN models. Additionally, as baselines for comparison, conventional classifiers such as Random Forest and Support Vector Machine (SVM) are included.
The system goes through these crucial stages:

- *Preprocessing:* To improve dataset diversity and lessen overfitting, images are resized, normalized, and enhanced (flipping, zooming, and rotation).
- *Model Training:* A labelled dataset of 30,000 plant photos in 30 classes is used to train CNN models, concentrating on seven plant species, particularly medicinal ones.
- *Evaluation:* Common metrics like accuracy, precision, recall, and F1-score are used to assess models.

By fusing deep learning with careful data preparation and thorough assessment, the suggested system seeks to increase the precision and resilience of plant identification, especially for medicinal plants.

5. Methodology

Identifying medicinal plants through machine learning involves a series of well-structured steps, starting with data collection and ending with the deployment of the final model. Each phase was critical in ensuring that the system performed well and efficiently. The methodology used in this study was as follows.

5.1. Data Collection

A carefully selected dataset of 30,000 high-resolution plant photos arranged into 30 different classes forms the basis of the model. Crops, Fruits, Industrial, Medicinal, Nuts, Tubers, and Vegetables are the seven plant types that further subdivide these classes. The 1,000 labelled images in each plant class were taken in various settings, including lighting conditions, backgrounds, and viewpoints. In order to replicate real-world situations and increase the model's resilience, this diversity has been purposefully added.

5.2. Data Preprocessing

Improving model accuracy and generalization requires efficient preprocessing. The following methods are used:

- *Resizing:* Every image has been resized to 224 x 224 pixels, a standard size that works with the input layer of well-known Convolutional Neural Networks (CNNs), including VGG16 and ResNet50.
- *Normalization* is scaling image pixel values initially in the [0,255] range to the [0,1] range. During backpropagation, this normalization enhances gradient flow and speeds up training.
- *Data Augmentation:* Augmentation techniques decrease overfitting and increase the dataset's adequate size and variability. These techniques include:
  Flipping both horizontally and vertically,
  Rotations at random,
  Zooming in and out,
  Adjustments for contrast brightness and
  Cropping at random.

  These additions improve the model's ability to generalize to new data.

5.3. Model Selection and Training

Three different model types are created and contrasted in order to identify the top-performing architecture:

5.3.1. CNN-based Deep Learning Models

Using VGG16 and ResNet50 for Transfer Learning: The plant dataset is used to refine ImageNet's pre-trained models. The models maintain general visual characteristics while adjusting to patterns unique to plants by freezing the early layers and retraining the deeper layers.

Custom CNN Architecture: From the ground up, a lightweight CNN with fewer layers is created and trained. Without significantly sacrificing accuracy, this model is tuned for performance in resource-constrained settings, like mobile apps.

5.3.2. Conventional Models for Machine Learning

Random Forest (RF) and Support Vector Machine (SVM) classifiers are trained as a baseline. But these models need:
Techniques for manually extracting features to obtain descriptors like:
Histograms of colors,
Features of texture (such as Local Binary Patterns) and
Descriptors of shape.
To compare the SVM and RF models' performance with deep learning models, these features are fed into the models.

- *Configuration for Training:* Loss Categorical cross-entropy is a useful tool for multi-class classification.
- *Adam (Adaptive Moment Estimation)* is an optimizer renowned for quick convergence.
- *Learning Rate:* Adjusted through experimentation.
- *Data Split:* 20% for testing and 80% for training.

5.4. Model Evaluation

The models are thoroughly assessed using a range of performance metrics after training:

- *Accuracy:* Indicates the proportion of accurately predicted pictures.

- *Precision, Recall, and F1-Score:* Evaluated per class to handle any class imbalance and give deeper insight into model behaviour.
- *Confusion Matrix:* Provides a thorough overview of all classes' true positives, false positives, and false negatives while emphasizing any incorrect classifications.
- In order to guarantee reliability, K-fold cross-validation, also known as 5-fold, is an optional procedure in which the dataset is split up into k subsets, and the model is trained and validated k times, using a different validation subset each time. By doing this, bias from a single train-test split is lessened.

5.5. Model Deployment

Regarding accuracy, dependability, and speed, the CNN-based model performs better after comparison, especially the transfer learning model that uses ResNet50. The deployment of this final model is chosen.

Choices for Integration:

- *Web application:* Users can get immediate identification results by uploading plant photos from a browser.
- *Mobile Application:* Using a smartphone's camera, users can take and upload pictures of flowers or plant leaves thanks to an easy-to-use interface.
- *Prediction in Real-Time Pipeline:* Input of images (from the camera or upload), Resizing and normalizing images and Model forecasting.

Plant class, confidence level, and a brief description of the medication are shown in the output.

The suggested system offers a practical, user-friendly tool for automated medicinal plant recognition by utilizing deep learning and an extensive dataset, which may benefit researchers, farmers, herbal practitioners, and plant enthusiasts.

5.6. Algorithm

Step 1 : Images are entered.
Step 2 : A scanner with the highest resolution was used to scan the leaves' front and back sides. A leaf-image dataset contained the pictures.
Step 3 : Preprocessing was done on the photos. The dataset's image dimensions are set to the necessary size.
Step 4 : Training and testing datasets were separated from the preprocessed dataset.
Step 5 : The Convolutional Neural Network was fed the Training dataset.
Step 6 : The testing dataset and the CNN layer's output are supplied as inputs for performance evaluation. The model's and the validation set's accuracy and loss were considered in this step, and a confusion matrix was used to plot the accuracy and loss graphs.
Step 7 : The output layer of the convolutional neural network displays the image.

5.7. Flowchart of the Algorithm

This flowchart uses a Convolutional Neural Network (CNN) to show how images are classified. Preprocessing steps are taken to prepare the data after image properties are analyzed and the dataset is gathered. The CNN is trained on the training set and assessed using the validation set after the dataset has been divided into training and validation sets. In order to guarantee precise image classification, performance is evaluated at the end.

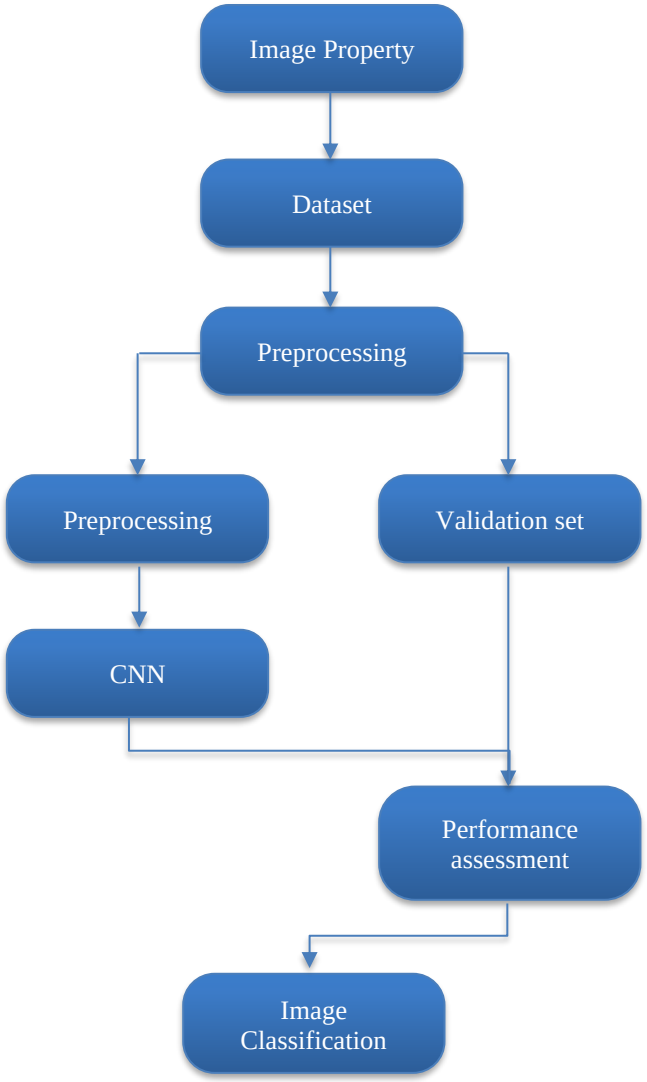

Fig. 1 Flowchart of the algorithm

6. Results

6.1. Training and Validation Accuracy Graph

Following model training, the relationship between training accuracy, validation accuracy, and the number of epochs (30) is plotted on a graph. Training accuracy is represented by the red line in the graph, and the blue line represents validation accuracy. The following formula is used to determine the accuracy. The training and validation graphs are displayed below.

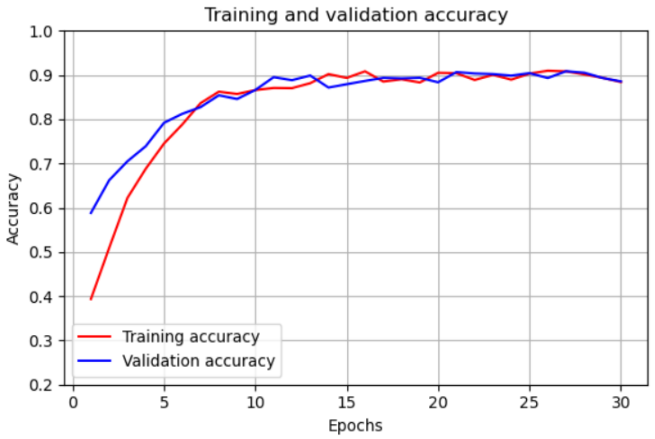

Fig. 2 Training and validation accuracy

The above-mentioned table displays the training and validation accuracy, in which the epoch is been shown with a scale of 5, i.e., 1-30.

Epoch	Training Accuracy	Validation Accuracy
1	0.35	0.45
5	0.80	0.86
10	0.91	0.91
15	0.92	0.91
20	0.92	0.91
25	0.92	0.91
30	0.92	0.90

Fig. 3 Training and Validation accuracy table

Formula to calculate accuracy

Accuracy = Number of correct predictions / Total number of predictions

6.2. Training and Validation Loss Graph

Fig. 4 Training and validation loss graph

Epoch	Training Loss	Validation Loss
0	2.4	1.9
5	1.0	0.7
10	0.4	0.3
15	0.2	0.1
20	0.1	0.05
25	0.05	0.02
30	0.03	0.01

Fig. 5 Training and Validation Loss Table

After training the model, a graph shows the relationship between training accuracy, validation accuracy, and the number of epochs (30). In the graph, the blue line represents validation accuracy, while the red line represents training accuracy. The accuracy is calculated using the following formula. The following graph shows the training and validation loss graph, and the following table shows the values of the loss graph. The above-mentioned table displays the training and validation loss, in which the epoch is been shown with a scale of 5, i.e., 1-30.

Formula to calculate loss

$$\textbf{Loss} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

Yi- The actual output
Yi^- The predicted output
n- Number of inputs
i- Iteration

6.3. Output

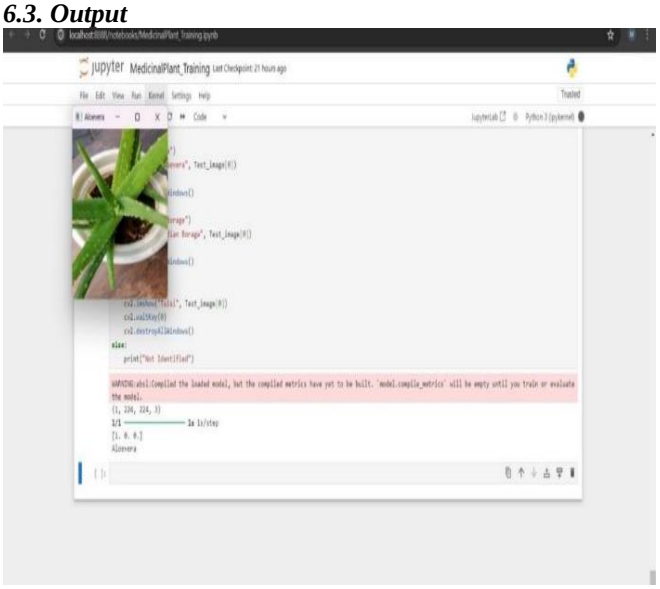

Fig. 6 Example output of Aloe vera

7. Conclusion

This project shows how well deep learning methods, particularly Convolutional Neural Networks or CNNs, work for automatically identifying medicinal plants. The system attains high classification accuracy and robustness by utilizing a vast and varied dataset of 30,000 photos spanning 30 plant classes and seven plant types. It was discovered that CNN-based models perform noticeably better than classical methods in accuracy and generalization after extensive testing using pretrained models like VGG16 and ResNet50, a lightweight custom CNN, and conventional machine learning algorithms like SVM and Random Forest. Applying transfer learning, data augmentation, and standard evaluation metrics further enhances the system's dependability. Researchers, farmers, students, and herbal practitioners can now access the solution thanks to the trained model's successful integration into an intuitive web or mobile application. Just uploading or taking a picture enables real-time plant identification, speeding up decision-making and raising awareness of medicinal plants. In conclusion, this work offers a valuable and scalable AI-powered plant recognition tool that may eventually be extended to detect plant diseases, provide multilingual support, and integrate with agricultural advisory systems.

Acknowledgments

We express our heartfelt gratitude to all those who contributed to successfully completing the project, Identification of medicinal plants using machine learning algorithms. We express our sincere appreciation to the Hyderabad Institute of Technology and Management for providing an enriching academic environment and the necessary resources to undertake this project.

Our special thanks go to Mrs. Krishna Jyothi, Project Coordinator, for her invaluable guidance, meticulous oversight, and continuous encouragement throughout the project. We are deeply grateful to our Internal Guide, Dr. M. Rajeshwar, Associate Professor, for his unwavering support, technical expertise, and insightful feedback, which were instrumental in shaping the project's direction and ensuring its quality.

We also acknowledge the collaborative efforts and dedication of the project team members, P. Sandeep Kumar, Giridhar Paida, and Kodi Satheesh. Their hard work, innovative ideas, and commitment to excellence were pivotal in bringing this vision to fruition.

References

[1] Yudha Islami Sulistya, *Plants Type Datasets*, Kaggle, 2021. [Online]. Available: https://www.kaggle.com/datasets/yudhaislamisulistya/plants-type-datasets

[2] Karen Simonyan, and Andrew Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *arXiv*, pp. 1-14, 2014. [CrossRef] [Google Scholar] [Publisher Link]

[3] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84-90, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[4] Corinna Cortes, and Vladimir Vapnik, "Support-Vector Networks," *Machine Learning*, vol. 20, no. 3, pp. 273-297, 1995. [CrossRef] [Google Scholar] [Publisher Link]

[5] Leo Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5-32, 2001. [CrossRef] [Google Scholar] [Publisher Link]

[6] Kaiming He et al., "Deep Residual Learning for Image Recognition," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770-778, 2016. [Google Scholar] [Publisher Link]

[7] Francois Chollet, "Xception: Deep Learning with Depthwise Separable Convolutions," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1251-1258, 2017. [Google Scholar] [Publisher Link

[8] Nilesh S. Bhelkar, and Avinash Sharma, "Identification and Classification of Medicinal Plants using Leaf with Deep Convolutional Neural Networks," *International Journal of Health Sciences*, vol. 6, no. S6, pp. 11596-11605, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[9] Chengzhuan Yang et al., "Plant Leaf Recognition by Integrating Shape and Texture Features," *Pattern Recognition*, vol. 112, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[10] Loris Nanni, Stefano Ghidoni, and Sheryl Brahnam, "Ensemble of Convolutional Neural Networks for Bioimage Classification," *Applied Computing and Informatics*, vol. 17, no. 1, pp. 19-35, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[11] Angie K. Reyes, Juan C. Caicedo, and Jorge E. Camargo, "Fine-Tuning Deep Convolutional Networks for Plant Recognition," *Working Notes of CLEF 2015 - Conference and Labs of the Evaluation Forum*, Toulouse, France, vol. 1391, pp. 1-9, 2015. [Google Scholar] [Publisher Link]

[12] Stephen Gang Wu et al., "A Leaf Recognition Algorithm for Plant Classification Using Probabilistic Neural Network," *IEEE International Symposium on Signal Processing and Information Technology*, Giza, Egypt, pp. 11-16, 2007. [CrossRef] [Google Scholar] [Publisher Link]

[13] J. Arun Pandian, G. Geetharamani, and B. Annette, "Data Augmentation on Plant Leaf Disease Image Dataset Using Image Manipulation and Deep Learning Techniques," *2019 IEEE 9th International Conference on Advanced Computing (IACC)*, Tiruchirappalli, India, pp. 199-204, 2019. [CrossRef] [Google Scholar] [Publisher Link]