Automated Detection and Classification of Philippine Medicinal Plants Using Convolutional Neural Networks: A Deep Learning Approach

*King Dheniel N. Concepcion
College of Computing and Information Technology
National University Manila, Philippines
concepcionkd@students.national-u.edu.ph

Brando Allen A.Donato
College of Computing and Information Technology
National University Manila, Philippines
donatoba@students.national-u.edu.ph

Lee-an Mar Anicken P. Carpio
College of Computing and Information Technology
National University Manila, Philippines
carpiolm@students.national-u.edu.ph

Angelo Gabriel D. Castillo
College of Computing and Information Technology
National University Manila, Philippines
castilload@students.national-u.edu.ph

Allen M. Siaton
College of Computing and Information Technology
National University Manila, Philippines
siatonam@students.national-u.edu.ph

Abstract—

Keywords:

I. Introduction

The rural Filipino marketplaces and countryside contain a wealth of plant-based medicines that have been curing diseases for centuries. The Philippines is endowed with an awesome array of medicinal plants – living medicine shops that have been the pillars of folk healing for generations. But with all this heritage, accurately identifying these potent plants is still amazingly challenging, particularly with the loss of traditional knowledge with each succeeding generation. When cultural knowledge can no longer be assuredly passed on about which leaf is used to treat fever and which one is possibly toxic, communities are truly in danger. Santos and colleagues [1] discovered something deeply alarming – individuals are suffering adverse health consequences, occasionally severe ones, due to inaccurate plant identification and use. This is not a theoretical issue; it is happening to actual people in markets and communities all over the Philippines.

The challenge of plant identification lies in distinguishing between similar-looking leaves without specialized botanical knowledge. Even experienced traditional healers sometimes struggle to differentiate between beneficial medicinal plants and their potentially harmful lookalikes, particularly when examining dried specimens or plants outside their natural growing environment. The task resembles identifying a subject from a blurry photograph – possible, but prone to errors. The challenge becomes even more daunting when considering that leaves with nearly identical appearances can have dramatically different effects when consumed. In fact, studies in Philippine herbal markets found that more than one in four medicinal plant specimens were misidentified – a concerning statistic considering the health implications involved.

This research addresses these challenges through the development of a machine learning-driven application that functions as a portable botanical expert. Technology enables users to point a mobile device at a leaf and instantly determine whether it is the intended medicinal plant or a potentially dangerous alternative. The project involves developing a machine learning application capable of detecting subtle differences between plant species that might be imperceptible to untrained observers. Utilizing the latest advances in computer vision and deep learning, the system will capture highly detailed images of leaves from multiple angles and process these images through sophisticated neural networks – essentially training computational systems to recognize plants with greater accuracy than most humans can achieve. Unlike previous projects with limited scope, this system builds upon a comprehensive collection of over 5,000 images covering 50 different medicinal plants native to the Philippines, ensuring the technology works effectively for the plants most relevant to Filipino communities.

The research objectives include: First, creating a machine learning system specifically designed to recognize medicinal plants found throughout the Philippines. Second, building a comprehensive library of plant images captured in various natural conditions – sunny, cloudy, partially shaded – to ensure the system works reliably in real-world settings. Third, rigorously test the system's accuracy across different plant species and growing conditions. And finally, packaging the technology into a user-friendly application accessible without specialized botanical knowledge. As Cruz and Mendoza [2] emphasized, bridging the gap between ancient herbal wisdom and modern technology isn't merely an academic exercise – it represents an effort to preserve cultural heritage, protect public health, and ensure that the remarkable healing potential of Philippine flora continues to benefit communities for generations to come.

II. Review Of Related Works

The Philippines has diverse botany with some 13,500 plant species, of which about 1,500 are known to have documented medicinal properties (Ong et al., 2017). These plants have been essential to traditional Philippine medical practices for many generations, providing remedies for ailments and health problems. Nonetheless, identifying these plants remains difficult, especially for non-experts, restraining their full benefits and proper utilization (Tan et al., 2020). The emergence of machine learning approaches, particularly deep learning techniques for image classification, presents promising opportunities for automated medicinal plant identification. These technologies could significantly improve accessibility to plant identification resources and help preserve traditional knowledge about Philippine medicinal plants.

2.1 **Unsupervised Machine Learning for the Market Segmentation and Targeting**

Convolutional Neural Networks (CNNs) now predominate in plant identification because they can learn pertinent visual features from pictures automatically. Sulc et al. (2020) established that CNN-based methods systematically dominate conventional computer vision methods in plant identification, with performances over 90% on general plant datasets. For medicinal plants alone, Nguyen et al. (2019) reached an accuracy of 91.7% to recognize 50 medicinally important species using transfer learning with deep CNNs. The success of CNNs for plant recognition lies in their capacity for learning hierarchical feature representations of different plant features, ranging from basic edges and textures to intricate patterns such as leaf venation and margin features (Lee et al., 2017). Such a capacity is highly useful in medicinal plant recognition, where slight visual distinctions can separate medicinal members from morphologically similar non-medicinal relatives.

2.2 Identifying and Analyzing Emerging Trends in Market Metrics and Purchasing Behavior

Understanding the purchasing pattern is demonstrated as a must-have for companies and those who are aiming to reassemble and transform their products based on customer preferences. Unsupervised learning methods, which are mainly the association rule method [6] and the topic modeling [5], utilize these methods to spot trends, hence enabling organizations to be quick and flexible to the altering requirements of markets (Ikhwanul Hakim, 2023). This method also allows the identification of trend developments through the analysis of large volumes of consumer data (w/o/with) predefined data. Cluster and association rules are the most important techniques that help in discovering new patterns; they assist companies in changing their marketing strategies in dynamic ways during the decoding.

The Apriori method is a famous technique used in market analysis to mine association rules. It was first presented by Agrawal et al. (1993) and has been in use in many research studies, finding relationships among the products that are frequently bought together. Employing this technique, businesses are to make the most of additional information by the implementation of an automatic storage system that will optimize the effectiveness of the website. Deservingly, Nayyar et al. (2021) talk about technology as the next enabler of e-commerce.

III. Methodology

A. Data Collection

The dataset utilized in this study is the "Philippine Medicinal Plant Leaf Dataset" created by Jayde Paolo Mirandilla of forty (40) different medicinal plant species found in the Philippines with 4,922 leaf images. All plant species are labeled and grouped according to their medicinal uses and botanical classification. They were photographed against natural light using a 48-MP Android phone camera. Each leaf was shot both front and back to ensure that it could be identified regardless of the position. Data collection was done between September 2022 and December 2022, where local medicinal plants from the Philippines were taken.

B. Data Pre- Processing

A comprehensive data pre-processing approach was implemented to enhance the efficiency of the leaf image analysis and classification system. The following pre-processing methods were employed:

1. Image Standardization: All the leaf images were resized to a uniform size (224×224 pixels) to have the same input data format for the machine learning models.
2. Background Removal: A background subtraction approach was utilized to isolate the leaf from the background, eliminating noise and focusing the analysis on the features of the leaf.
3. Image Enhancement: Techniques such as histogram equalization were applied in an attempt to improve contrast and highlight distinctive features of different leaf species. Data
4. Augmentation: To improve the robustness of the model and to counteract the effect of limited data, operations such as rotation, flip, zooming, and certain color modifications were used. These augmented the dataset size and improved the model's ability to generalize under various real-world scenarios.
5. Normalization: Normalized pixel values to

range [0,1] to standardize the data distribution and improve the stability of model training.

C. Data Features

The dataset consists of 4,922 leaf images across 40 Philippine medicinal plant species. The features extracted from these images include:

Table I. *Features of Dataset*

Feature Type	Description
Shape Features	*Leaf contour, aspect ratio, eccentricity, solidity*
Texture Features	*Surface patterns, venation structure, edge characteristics*
Color Features	*RGB color distribution, color moments, dominant colors*
Morphological Features	*Area, perimeter, compactness, circularity*
Venation Features	*Vein patterns, density, and distribution*

D. Experimental Setup

The study utilizes Python as the primary programming language and various deep learning frameworks for the environment, with several libraries for image processing, machine learning, and data visualization. The following tools and libraries were used:

1. TensorFlow/Keras: For building and training deep learning models.
2. OpenCV: Handles image pre-processing and feature extraction.
3. NumPy: Used for numerical computation and array manipulations.
4. Pandas: For data management and organization.
5. Matplotlib and Seaborn: Used for data visualization and result presentation.
6. Scikit-learn: Used for data preprocessing, model evaluation, and traditional machine learning techniques.

The experiments were conducted on a system with GPU acceleration to handle the computational demands of deep learning model training.

E. Machine Learning and Deep Learning Models

A. CNN Architecture

A specifically tailored convolutional neural network architecture was designed for the Philippine medicinal plant data set. The architecture begins with the input layer, which can accept 224×224×3 RGB images, and four convolution blocks with increasingly larger filter sizes (32, 64, 128, and 256) with 3×3 kernels followed by ReLU functions and then by 2×2 max pooling operations. After flattening, the network uses two dense layers of 512 and 256 neurons with ReLU activation and dropout regularization with 0.5 and 0.3 rates to prevent overfitting. The output layer has 40 neurons with softmax activation, each representing a plant species in the data set.

B. Support Vector Machine (SVM)

A traditional machine learning algorithm was implemented for comparison purposes. This approach utilizes HOG (Histogram of Oriented Gradients) features extracted from the preprocessed leaf images, which capture local gradient information and edge orientations. The SVM classifier employs an RBF kernel to handle the non-linear separation of data points in the feature space, implementing a one-vs-rest strategy for tackling the multi-class classification problem. Hyperparameter optimization was conducted through grid search, focusing on the regularization parameter C and the kernel coefficient gamma to achieve optimal performance.

This comparison between deep learning (CNN) and traditional machine learning (SVM) approaches provides insights into the relative performance of different classification paradigms for the medicinal plant identification task.

F. Training Procedure

Training commenced with the separation of the dataset into training (70%), validation (15%), and test sets (15%), with the equal representation of all 40 plant species. In the case of the CNN model, transfer learning strategies were followed by leveraging pre-trained models and replacing the last classification layer with a new one having 40 output nodes specific to every plant species. The models were subjected to two-step fine-tuning: the first training the sole custom classification layers and freezing the pre-trained weights, then fine-tuning after unfreezing the later convolutional layers at a smaller learning rate. Wide hyperparameter search was carried out via grid search over learning rates of 0.0001 to 0.01, batch sizes 16, 32, and 64, different configurations of Adam optimizers, and dropout between 0.3 and 0.5 for successful regularization. To reduce overfitting, early stopping was utilized with patience 10 epochs and tracking of validation loss to stop at the right time.

G. Evaluation Metrics

Each model's performance was measured based on the following metrics:

1. Accuracy: The percentage of correctly classified plant species.

2. Precision: The ratio of true positive predictions to the total number of positive predictions for each class.

3. Recall: The ratio of true positive predictions to the total actual positives for each class.

4. F1-Score: The harmonic mean of precision and recall, giving a balanced score.

5. Confusion Matrix: Exhaustive breakdown of classification performance by true positives, false positives, true negatives, and false negatives per class.

6. Top-5 Accuracy: Percentage of test samples for which the correct label is among the top five predicted labels.

7. Training and Inference Time: Computational efficiency metrics for deployment considerations.

The best-performing model was selected based on a weighted combination of these metrics, with particular emphasis on overall accuracy and F1-score across all plant species.

IV. Results and Discussion

V. Conclusion

References

[1] Santos, J., Garcia, P., & Lopez, M. (2021). Adverse health effects from misidentified medicinal plants: A five-year review of cases in Philippine rural communities. Toxicology Reports, 8, 1523-1537. https://doi.org/10.1016/j.toxrep.2021.09.007

[2] Cruz, M., & Mendoza, R. (2020). Challenges in medicinal plant identification and authentication in Philippine traditional medicine. Journal of Ethnopharmacology, 245, 112193. https://doi.org/10.1016/j.jep.2019.112193

[3] H. G. Ong, Y. D. Kim, and J. H. Lee, "Ethnomedicinal plants and traditional knowledge among three subgroups of Bukidnon in Mindanao, Philippines," Journal of Ethnobiology and Ethnomedicine, vol. 13, no. 1, pp. 1-14, 2017. https://doi.org/10.1186/s13002-017-0182-7

[4] J. Y. Tan, C. Tengco-Sarmiento, and A. Calibo, "Digital preservation of Philippine medicinal plants: Challenges and opportunities," Journal of Philippine Librarianship, vol. 40, no. 1, pp. 52-68, 2020. https://journals.upd.edu.ph/index.php/jpl/article/view/7844

[5] M. Sulc, L. Picek, J. Matas, T. Jeppesen, and J. Heilmann-Clausen, "Recognition of the host tree species of fungi based on convolutional neural networks," Ecological Informatics, vol. 61, article 101194, 2020. https://doi.org/10.1016/j.ecoinf.2020.101194

[6] T. T. Nguyen, T. L. Le, N. H. Pham, T. T. Tran, and T. T. Le, "Identification of medicinal plants by using deep learning with expanded multi-scale CNN," Journal of Computer Science and Cybernetics, vol. 35, no. 3, pp. 267-284, 2019. https://doi.org/10.15625/1813-9663/35/3/13930

[7] S. H. Lee, C. S. Chan, S. J. Mayo, and P. Remagnino, "How deep learning extracts and learns leaf features for plant classification," Pattern Recognition, vol. 71, pp. 1-13, 2017. https://doi.org/10.1016/j.patcog.2017.05.015

[8] Mirandilla, J. D. P. (2023). Philippine Medicinal Plant Leaf Dataset. Kaggle. https://www.kaggle.com/datasets/jaydepaolomirandilla/philippine-medicinal-plant-leaf-dataset

[9] D. S. Prasvita and Y. Herdiyeni, "MedLeaf: Mobile application for medicinal plant identification based on leaf image," International Journal on Advanced Science, Engineering and Information Technology, vol. 3, no. 2, pp. 5-8, 2013. http://dx.doi.org/10.18517/ijaseit.3.2.303

[10] C. R. Quinto, J. P. A. Samonte, and M. M. Magramo, "Philippine Herbal Plants Classification Using Deep Convolutional Neural Network," in 2020 IEEE 12th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment, and Management (HNICEM), pp. 1-6, 2020. https://doi.org/10.1109/HNICEM51456.2020.9400095