Classification of philippine herbal plants via leaf using different machine learning algorithms

Robert G. De Luna, Marife A. Rosales, and Elmer P. Dadios

***Abstract*—Herbal plants have significant role in the field of medicine to cure some known diseases. Presently, the herbal plant identification is performed manually by an expert or person with enough knowledge regarding the plant. Sometimes, manual identification is prone to human error, resulting to incorrect usage of herbal plants. In this study, a machine vison-based herbal plant identification was implemented. An improvised image capturing system with 16 megapixels resolution camera was used with the aid of MATLAB installed in laptop to gather real images of twelve herbal plants. An intelligent system was developed by utilizing image processing, feature extraction, and machine learning (ML) algorithms using Python. The classification accuracy was used to select the best model. Moreover, F1 score metric was used to compare the performances of the default and optimized models in identifying all the herbal classes. Based on the results, SVM model showed the best performance in classifying the herbal plants with accuracy score of 94.50% and 93.30% for the optimized training and testing performances.**

Keywords*— *philippine herbal plants, machine learning, image processing, leaf classification

I. INTRODUCTION

Treatment using herbal plants is considered effective, safe, and inexpensive. Herbal plants are abundant everywhere, but people cannot maximize its usage since it is hard to identify the name of each species. In Philippines, there are 13,500 plant species, from which, around 1,500 are medicinal plants and more than 3,500 species are considered indigenous [1]. Among the registered medicinal plants, only 120 (12.5%) species are scientifically validated to use safely and effectively [1]. Presently, the herbal plant identification is performed manually by an expert [2] but the procedure is prone to human error, slow and time-consuming [3]. Some researchers used leaf, shape, texture, and flower to identify the herbal plants to optimize their applications and utilization. Development of a robust system capable of identifying herbal plant species, exploring their therapeutic applications and proper utilization is still a challenge [2] [3].

Machine-vision system is one of the promising technologies to develop a smart herbal plant identification system. It will alleviate and modernize the manual method which is inaccurate and slow. The objective of the study is to design and develop a robust and machine-vision based system to correctly identify the herbal plants. The system is composed of a capturing box for image acquisition, where MATLAB was used to perform the image processing and feature extraction. Sci-kit learn libraries using Python integrated development environment (IDE) was used to develop and to train the models using ML algorithms for classification such as support vector machine (SVM), Logistic Regression (LR) and K-Nearest Neighbor (KNN). These three models are the most common machine learning models used in classification task.

II. REVIEW OF RELATED WORKS

Researches on identifying herbal plants are done with different approaches. Some proponents used shape features [4], color feature and texture features [5] with application of artificial intelligence and machine learning. Leaf identification is one of common methods to identify the plant type but there should be a robust, accurate and efficient system to perform the tasks [2][6][7][8][9][10][11][12][13]. Gugol et al [13] on the other hand, used flower species utilizing color, texture and shape to identify medicinal plant using deep learning approach by means of convolutional neural networks (CNN). A novel fuzzy local binary pattern (LBP) model was used and developed to extract the texture features of herbal pants for effective herbal identification [14][15].

Robert G. De Luna, Department of Electronics Engineering, De La Salle Lipa, Philippines (email: robert.deluna@dlsl.edu.ph)

Marife A. Rosales, Department of Electronics Engineering, De La Salle Lipa, Philippines (email: marife.rosales@dlsl.edu.ph)

Elmer P. Dadios, Department of Manufacturing Engineering and Management, De La Salle University, Philippines (email: elmer.dadios@dlsu.edu.ph)

Image processing [16] is a method where different processes are performed in the images in order to enhance its quality and to extract some features for evaluation to get the desired result. Different image processing techniques were done to enhance the leaf images of plants for easier identification. Features of leaf can be classified as shape-based such as aspect ratio, area, and rectangularity; color-based such as standard deviation and mean; and venation. Singh and Bhamrah [17] used leaf characteristics such as solidity, major and minor axes, perimeter, area, aspect ratio and eccentricity. Fataniya et al. [3] used microscopic images and classify plants using the characteristic features. The features used in the study are ratio, major and minor axes length and area. After the feature extraction, the microscopic images of herbal plant were analyzed correctly using K-means data mining algorithm. A digital camera of 15 megapixels was used to capture the images of different leaves [8]. Significant features of leaf were identified in [18] such as shape, leaf margins, color and texture [19], and vein structure [20]. Feature selection was done to reduce redundant data to improve the learning rate and the training time [18]. Gopal et al [21] has successfully designed a system and used image processing, feature extraction, pattern recognition and classification to identify the medicinal plant leaf. The developed system has an accuracy of 92%.

The study of Sabu et all [11] used the Speed Up Robust Feature (SURF) [22] and Histogram of Oriented Gradients (HoG) to extract scale invariant features of plant leaf. After feature extraction, K-NN classifier was used which provided nearly 100% classification accuracy [11]. Near-infrared spectroscopy (NIR) spectra with principal component analysis (PCA) [8] was used also to classify correctly the different medicinal plant varieties via leaf [23].

Other method to perform the feature extraction is using the wrapper based genetic algorithm (GA) as presented in the study of Sainin & Alfred [20]. In this study, the identification and classification of medicinal plant via leaf shape using Direct Ensemble Classifier for imbalanced multiclass learning (DECIML) was used but the classification result was low due to small dataset. Gray-level co-occurrence matrix (GLCM) and back propagation multi-layer perceptron (BP-MLP) were utilized for feature extraction and plant classification in [24] but the study did not perform feature selection and optimization to further improve the classification rate. Moreover, the artificial neural network (ANN) was used for herbal plant leaf identification using shape feature with accuracy equal to 98.8% [17] while support vector machine (SVM) classifier was used in [25] but the trained model encountered some misclassification due to not normalized features. Other study utilized also the SVM classifier like in the study of Poudel et al [26] with good classification accuracy of 90%. Furthermore, SVM, Random Decision Forest, and ANN were used and compared in the study [27][28] to identify plant diseases via leaf which showed that Random Forest achieved has the highest F1-score. Additionally, Rahmani et al [29] used different supervised ML algorithms such as Decision Tree (DT), Naive Bayes, K-nearest neighbor (KNN) and ANN to classify plant leaves which gave best performance result for ANN [30]. Other works used and implemented a plant leaf identification using pre-trained model like VGG16 [31].

Previous studies used different approaches to identify herbal plants such image processing, feature extraction, and artificial intelligence. However, the development of a machine vision-based system with optimized models to further improve the learning performance and processing time was not implemented. In this study, the proponents developed a smart machine vision-based system to capture the images of herbal leaf and to classify it correctly and efficiently. The features of leaf such as aspect ratio [17], circularity[32], convexity[32], solidity [17], and rectangularity[16] were extracted to generate the needed dataset which will be fed to the machine learning algorithms. Comparison of learning performances using default and optimized features and parameters was performed to choose the best machine learning model.

III. METHODOLOGY

A. Dataset Description

The study is limited only to twelve herbal plants such as balbas-pusa, ampalaya, akapulko, malunggay, sambong, lagundi, tsaang-gubat, niyog-niyogan, oregano, bayabas, yerba Buena, and ulasimang-bato included in the approved list of the Department of Health (DOH) for medical applications. Fig. 1 shows the herbal plants sample images included in the study.

Images are acquired using laptop with MATLAB software connected in an improvised capturing box, built with 16 MP A4Tech web-camera having a resolution of 16 megapixels. The camera is mounted in the capturing box with 37 cm distance from the leaf samples placed on a white background. A total of 600 images were collected from an equal distribution of 50 images per herbal plant.

Gathered images are subjected to several image processing steps to generate physical features of the leaf which will be used for the creation of machine learning models.

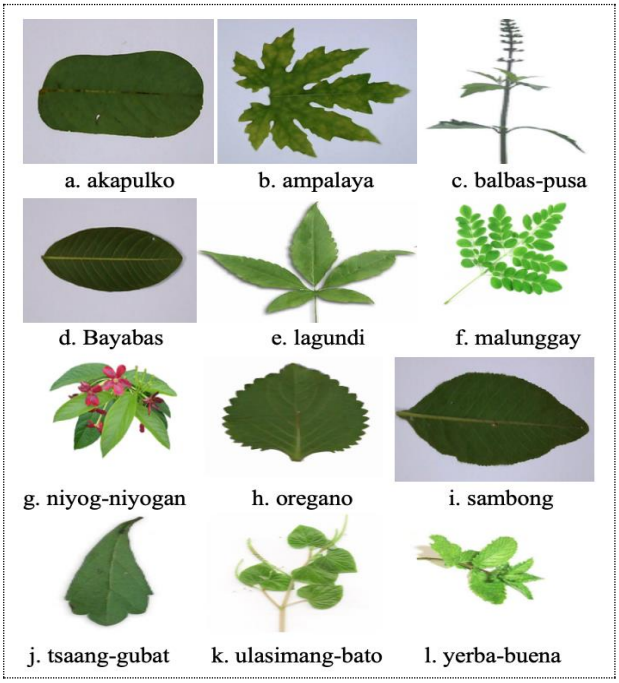


Fig 1. Leaf Images of the Herbal Plant

B. Proposed Work

Shown in Fig. 2 is the overview of the proposed work with the following stages; image acquisition, image processing, extraction of geometric features, machine learning modelling, and evaluation.

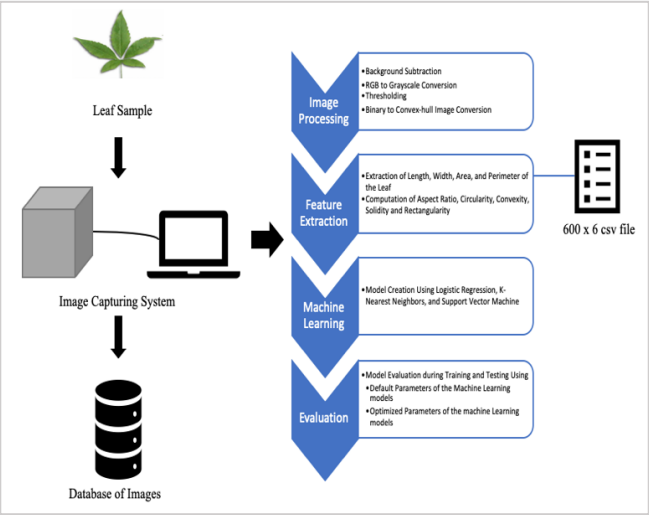


Fig 2. Overview of the Proposed Work

After gathering the 600 images, several image processing techniques were applied. The subject in the image is highlighted by removing the background. Its red-green-blue (RGB) component is converted to grayscale value before the process of thresholding which provided an image in black and white form. This binary image is then transformed to convex hull image to define edges of the subject where geometric features will be extracted.

Once the pre-conditioning of images was done, geometrical feature extraction followed the process. Primary features were extracted first that include area, width, length, and perimeter. Using the formula from the study of [22], secondary features like circularity, aspect ratio, rectangularity, convexity, and solidity, were computed. The generated values served as the dataset that were tabulated in a comma-separated value (csv) file format with 600 samples (rows) and 6 features (columns) including the "Herbal" class column.

After image processing and feature extraction, next step is the development of intelligent models using SVM, LR, and KNN algorithms. Models were created using the default parameters and then optimization was done for the improvement of the learning performances. The classification accuracy and F1-score metrics were used to evaluate and to choose the best model to be deployed. To make comparative analysis between models, the F1-score of each herbal class are also been tabulated. This provided a conclusion of what is the best machine learning model to be used in the identification of herbal plants and what herbal plants are difficult to recognized by the three models.

C. Development of the Intelligent Models

The proponents adapted three commonly used supervised ML algorithms such as SVM, LR, and KNN. SVM is one of the popular algorithms used for classification or regression that performs data transformation using kernel trick. After transformation, it finds optimal boundary or hyperplane to classify the data points. Another known supervised machine learning algorithm is KNN which is a model that classifies data points based on the points that are most similar to it. KNN is easy to use with efficient calculation time but mostly the accuracy depends on the quality of data. Moreover, like linear regression, logistic regression (LR) finds an equation that predicts the outcome. However, unlike linear regression, the response variable of logistic regression can be categorical or continuous and can be used to predict the probability of a certain class.

Python was utilized in this study to develop the intelligent models. The dataset was divided into 80% reserved for training set and 20% reserved for testing set. The training set is used to build the models while the test or validation set is used to validate the built models. To select the best ML model, classification accuracy was used as the performance metrics. On the other hand, to evaluate the model performance to correctly identify all the herbal plant classes, F1-score performance metric was used.

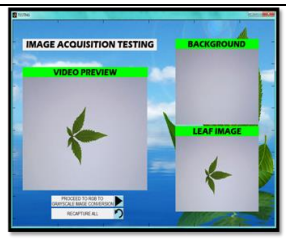
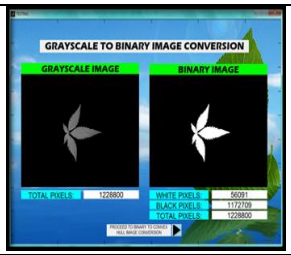
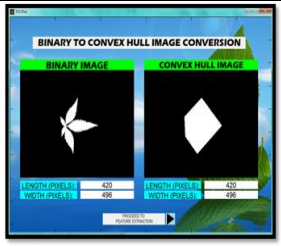
IV. RESULTS AND DISCUSSIONS

A. Dataset Generation for Machine Learning

Needed samples are captured with the aid of MATLAB installed in a laptop using the improvised

image capturing system with 16 megapixels resolution web-camera. The effect of varying illumination is lessened due to the controlled lighting inside the capturing box. The image processing toolbox of MATLAB was utilized in the preparation of images. Table 1 tabulated the sample results of every stages.

TABLE 1
SAMPLE RESULTS OF THE IMAGE PROCESSING AND FEATURE EXTRACTION

Process	GUI Result
Image Acquisition	
RGB to Grayscale Image Conversion	
Thresholding	
Binary to Convex Hull Image Conversion	
Geometrical Feature Extraction	

The processes are repeated for all the 600 images gathered. The extracted features are consolidated and saved in tabulated form as csv file. It contains 600 rows (samples) and 6 columns (features and class).

Table 2 summarized the descriptive statistics of the said dataset. As shown in the tabulated data, all features have complete number of sample count. Feature scaling using standardization was performed to assure that all feature will contribute during training. The 25%, 50%, and 75% only indicate the first, second, and third quartile respectively, reflecting boundaries value for features in every quartile.

TABLE 2
DESCRIPTIVE STATISTICS OF THE DATASET OF SELECTED HERBAL PLANTS

	Aspect Ratio	Circularity	Convexity	Solidity	Rectangularity
count	600	600	600	600	600
mean	0.6628	0.3911	1.1926	0.8919	1.6997
std	0.2622	0.1616	0.3098	0.1657	0.7591
min	0.3259	0.0592	0.9946	0.3490	1.1434
25 %	0.4286	0.3206	1.0185	0.9205	1.3528
50 %	0.6250	0.4312	1.0507	0.9628	1.4333
75 %	0.8786	0.5238	1.1632	0.9813	1.6538
max	1.2663	06238	2.2898	0.9916	5.1639

B. Model Performances for Herbal Plant Identification

SVM, KNN and LR are the three classifiers used in the training to develop the intelligent models for herbal plant identification. The dataset splitting is 80% and 20% for training set and testing set wherein the large portion of the dataset distribution is used for training. To get reliable accuracy and avoid overfitting and high variance, hold-out validation and 10-fold stratified cross-validation were applied during training. Moreover, GridSearchCV was used to perform the model optimization to get the optimal values of the hyperparameters. Table 3 and 4 summarized the accuracy and F1-score of all models using the default and optimized parameters.

TABLE 3
F1-SCORE AND ACCURACY USING DEFAULT PARAMETERS

Model	Cross-Validation			
	Accuracy		Hold-out Validation	
	Mean	Variance	Accuracy	F1-score
LR	0.7650	0.0712	0.7917	0.7800
KNN	0.8783	0.0658	0.8833	0.8900
SVM	0.8316	0.0992	0.4250	0.3700

TABLE 4
F1-SCORE AND ACCURACY USING OPTIMIZED PARAMETERS

Model	Cross-Validation		Hold-out Validation	
	Accuracy			
	Mean	Variance	Accuracy	F1-score
LR	0.9050	0.0506	0.9250	0.9300
KNN	0.8967	0.0547	0.9083	0.9100
SVM	0.9450	0.0269	0.9330	0.9300

As shown in Table 3, using default parameters, KNN has better performance compared to LR & SVM. On the other hand, as shown in Table 4, after optimization, there's an improvement in LR and SVM models as reflected having above 90 performances both in accuracy and F1-score.

Fig. 3 shows the comparative analysis of the models in terms of accuracy using hold-out validation and cross-validation. As shown in figure, SVM provided a good learning both in training and testing as compared to LR and KNN.

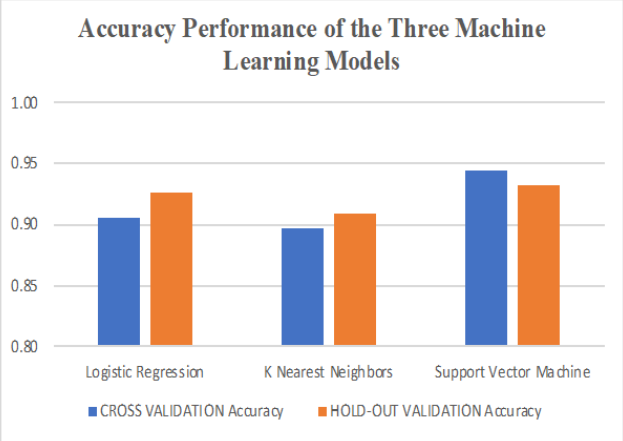


Fig 3. Training-Testing Validation Accuracy of the Three Optimized Models

Table 5 shows the summarized F1-score performance of each herbal plants in the three optimized ML models.

TABLE 5
F1-SCORE PERFORMANCE OF ALL HERBAL PLANTS USING THE THREE OPTIMIZED MACHINE LEARNING MODELS

Herbal Plant	LR	KNN	SVM
Akapulko	/	/	/
Ampalaya	/	/	/
Balbas-pusa	/	/	/
Bayabas	/	/	/
Lagundi	/	/	/
Malunggay	/	X	/
Niyog-Niyogan	X	X	X
Oregano	/	/	/
Sambong	/	/	/
Tsaang-gubat	X	X	X
Ulasimang-bato	/	/	/
Yerba Buena	/	/	/
No. of Items with F1-score Above 90 %	10	9	10
No. of Items with F1-score Below 90 %	2	3	2

Table 5 provided an insight of which among the herbal plants included in the study have difficulty for the models to classify. A / is placed in the table which indicate that the specific herbal plant attained an F1-score of 90 percent and above, and X otherwise.

For the three models, only the niyog-niyogan and the tsaang-gubat are the herbal plants with F1-score less than 90%. It means that these two plants have high degree of confusion due to their geometrical features. They are very similar to each other that is why they are interchangeably classified.

V. CONCLUSIONS

The researchers successfully developed a classification system for twelve herbal plants common in the country. These are balbas-pusa, ampalaya, akapulko, malunggay, sambong, lagundi, tsaang-gubat, niyog-niyogan, oregano, bayabas, yerba Buena, and ulasimang-bato. Image capturing box was developed with attached high-resolution camera and controlled lighting condition. MATLAB image processing toolbox was used in doing the image conditioning of the 600 captured images of herbal plants. Geometrical features of the leaf like circularity, aspect ratio, solidity, convexity, and rectangularity are extracted to established good dataset. The performance of the three models were evaluated during training and testing phase

using the default and optimized parameters. SVM model registered the best performance in classifying the herbal plants with optimized training and testing accuracy equal to 94.50% and 93.30%, respectively. With a criterion of above 90% F1-score, all the three optimized models failed to classify the niyog-niyogan and the tsaang-gubat. This is due to their similar geometrical features that made the models confused during classification.

ACKNOWLEDGEMENT

The authors would like to acknowledge the assistance of the Office of Research and Publication (ORP) of De La Salle Lipa (DLSL) to complete this research. The authors also recognize the members of the Electronics Engineering Department under the College of Information technology and Engineering for their support and help.

REFERENCES

[1] J. Z. Galvez, "The Best 100 Medicinal Plants Medicinal Plants : Philippines ' Natural Living Treasures The Philippine Herbal Market 1 The Philippine Herbal Market 2," pp. 1–16, 2014.

[2] E. S. Kumar and V. Talasila, "Leaf features based approach for automated identification of medicinal plants," *Int. Conf. Commun. Signal Process. ICCSP 2014 - Proc.*, pp. 210–214, 2014.

[3] B. Fataniya, P. M. Patel, T. Zaveri, and S. Acharya, "Microscopic image analysis method for identification of Indian Herbal Plants," *2014 Int. Conf. Devices, Circuits Commun. ICDCCom 2014 - Proc.*, vol. 1, pp. 0–4, 2014.

[4] N. Jamil, N. Aslina, C. Hussin, S. Nordin, and K. Awang, "Automatic Plant Identification : Is Shape the Key Feature ?," *Procedia - Procedia Comput. Sci.*, vol. 76, no. Iris, pp. 436–442, 2015.

[5] C. H. Arun, W. R. S. Emmanuel, and D. C. Durairaj, "Texture Feature Extraction for Identification of Medicinal Plants and Comparison of Different Classifiers," vol. 62, no. 12, 2013.

[6] T. Vijayashree and A. Gopal, "Database Formation for Authentication of Basil (Ocimum tenuiflorum) Leaf Using Image Processing Technique," *Gate to Comput. Vis. Pattern Recognit.*, vol. 1, no. 1, pp. 9–17, 2015.

[7] L. Gao and X. Lin, "A study on the automatic recognition system of medicinal plants," *2012 2nd Int. Conf. Consum. Electron. Commun. Networks, CECNet 2012 - Proc.*, pp. 101–103, 2012.

[8] G. Mukherjee, A. Chatterjee, and B. Tudu, "Morphological feature based maturity level identification of Kalmegh and Tulsi leaves," *Proc. - 2017 3rd IEEE Int. Conf. Res. Comput. Intell. Commun. Networks, ICRCICN 2017*, vol. 2017-Decem, pp. 1–5, 2017.

[9] A. Rahmad, Y. Herdiyeni, A. Buono, and S. Douady, "Multiscale fractal dimension modelling on leaf venation topology pattern of Indonesian medicinal plants," *Proc. - ICACSIS 2014 2014 Int. Conf. Adv. Comput. Sci. Inf. Syst.*, pp. 357–361, 2014.

[10] A. Salima, Y. Herdiyeni, and S. Douady, "Leaf vein segmentation of medicinal plant using Hessian matrix," *ICACSIS 2015 - 2015 Int. Conf. Adv. Comput. Sci. Inf. Syst. Proc.*, pp. 275–279, 2016.

[11] A. Sabu, K. Sreekumar, and R. R. Nair, "Recognition of ayurvedic medicinal plants from leaves: A computer vision approach," *2017 4th Int. Conf. Image Inf. Process. ICIIP 2017*, vol. 2018-Janua, pp. 574–578, 2018.

[12] A. Sabu, "Used in Leaf Based Plant Recognition Through Image Analysis Approach," no. Icicct, 2017.

[13] I. Gogul and V. S. Kumar, "Flower species recognition system using convolution neural networks and transfer learning," *2017 4th Int. Conf. Signal Process. Commun. Networking, ICSCN 2017*, pp. 1–6, 2017.

[14] "A novel Fuzzy LBP based Symbolic Representation technique for Classification of Medicinal Plants . Naresh Y G DoS in Computer Science Nagendraswamy H S DoS in Computer Science," pp. 524–528, 2015.

[15] N. N. K. Krisnawijaya, Y. Herdiyeni, and B. P. Silalahi, "Parallel Technique for Medicinal Plant Identification System using Fuzzy Local Binary Pattern," vol. 11, no. 1, pp. 77–90, 2017.

[16] H. P. Borase and B. K. Salunke, "Plant Extract : A Promising Biomatrix for Ecofriendly , Controlled Synthesis of Silver Nanoparticles," pp. 1–29, 2014.

[17] S. Singh and M. S. Bhamrah, "Leaf Identification Using Feature Extraction and Neural Network," vol. 10, no. 5, pp. 134–140, 2015.

[18] I. B. Pavaloiu, R. Ancuceanu, C. M. Enache, and A. Vasilateanu, "Important shape features for Romanian medicinal herb identification based on leaf image," *2017 E-Health Bioeng. Conf. EHB 2017*, pp. 599–602, 2017.

[19] T. Sathwik, R. Yasaswini, R. Venkatesh, and A. Gopal, "Classification of selected medicinal plant leaves using texture analysis," *2013 4th Int. Conf. Comput. Commun. Netw. Technol. ICCCNT 2013*, pp. 2–7, 2013.

[20] M. S. Sainin and R. Alfred, "Feature selection for Malaysian medicinal plant leaf shape identification and classification," *2014 Int. Conf. Comput. Sci. Technol. ICCST 2014*, vol. 2014, 2014.

[21] A. Gopal, S. Prudhveeswar Reddy, and V. Gayatri, "Classification of selected medicinal plants leaf using image processing," *2012 Int. Conf. Mach. Vis. Image Process. MVIP 2012*, pp. 5–8, 2012.

[22] D. Venkataraman and N. Mangayarkarasi, "Computer vision based feature extraction of leaves for identification of medicinal values of plants," *2016 IEEE Int. Conf. Comput. Intell. Comput. Res. ICCIC 2016*, 2017.

[23] P. K. Sahaya Rajesh, C. Kumaravelu, A. Gopal, and S. Suganthi, "Studies on identification of medicinal plant variety based on NIR spectroscopy using plant leaves," *2013 15th Int. Conf. Adv. Comput. Technol. ICACT 2013*, pp. 3–6, 2013.

[24] G. Mukherjee, A. Chatterjee, and B. Tudu, "Study on the potential of combined GLCM features towards medicinal plant classification," *2016 2nd Int. Conf. Control. Instrumentation, Energy Commun. CIEC 2016*, pp. 98–102, 2016.

[25] D. Venkataraman and N. Mangayarkarasi, "Support vector machine based classification of medicinal plants using leaf features," *2017 Int. Conf. Adv. Comput. Commun. Informatics, ICACCI 2017*, vol. 2017-Janua, pp. 793–798, 2017.

[26] P. Poudel, S. Kumar, V. S. Philip, and P. Kishore, "ROBUST RECOGNITION AND CLASSIFICATION OF HERBAL LEAVES," pp. 2319–2322, 2016.

[27] S. Communication, "A comparative analysis of machine learning approaches for plant disease identification," vol. 4, no. 4, pp. 120–126, 2017.

[28] "A REVIEW OF DIFFERENT CLASSIFICATION TECHNIQUES IN MACHINE," 2018.

[29] M. E. Rahmani, A. Amine, and R. M. Hamou, "Supervised Machine Learning for Plants Identification Based on Images of Their Leaves," vol. 7, no. 4, pp. 17–31, 2016.

[30] M. A. Rosales, A. A. Bandala, R. R. Vicerra, and E. P. Dadios, "Physiological-Based Smart Stress Detector using Machine Learning Algorithms," *2019 IEEE 11th Int. Conf. Humanoid, Nanotechnology, Inf. Technol. Commun. Control. Environ. Manag. HNICEM 2019*, 2019.

[31] S. Prasad and P. P. Singh, "Medicinal plant leaf information extraction using deep features," *IEEE Reg. 10 Annu. Int. Conf. Proceedings/TENCON*, vol. 2017-Decem, pp. 2722–2726, 2017.