Feature Selection for Malaysian Medicinal Plant Leaf Shape Identification and Classification

Mohd Shamrie Sainin
School of Computing
Universiti Utara Malaysia
Sintok, Kedah, Malaysia
shamrie@uum.edu.my

Rayner Alfred
Faculty of Computing and Informatics
Universiti Malaysia Sabah
Kota Kinabalu, Sabah
ralfred@ums.edu.my

Abstract— **Malaysian medicinal plants may be abundant natural resources but there has not been much research done on preserving the knowledge of these medicinal plants which enables general public to know the leaf using computing capability. Therefore, in this preliminary study, a novel framework in order to identify and classify tropical medicinal plants in Malaysia based on the extracted patterns from the leaf is presented. The extracted patterns from medicinal plant leaf are obtained based on several angle features. However, the extracted features create quite large number of attributes (features), thus degrade the performance most of the classifiers. Thus, a feature selection is applied to leaf data and to investigate whether the performance of a classifier can be improved. Wrapper based genetic algorithm (GA) feature selection is used to select the features and the ensemble classifier called Direct Ensemble Classifier for Imbalanced Multiclass Learning (DECIML) is used as a classifier. The performance of the feature selection is compared with two feature selections from Weka. In the experiment, five species of Malaysian medicinal plants are identified and classified in which will be represented by using 65 images. This study is important in order to assist local community to utilize the knowledge and application of Malaysian medicinal plants for future generation.**

Malaysian medicinal plant, leaf, shape, identification, classification, image processing; feature selection, wrapper

I. Introduction

Since in the early 1990s, the efforts to identify plant from images have attracted various studies on different techniques for image processing, feature extraction and identification. Most of the studies are concentrating on full scale leaf features and still open the research for different approaches. Prior to this study, leaf identification can be categorized into three types, which includes shape-based, venation-based and combination of both approach.

In this paper, a study is conducted to illustrate the full scale leaf shape identification technique in order to identify leaf species. Leaf shape features can generate large number of features and pose a challenge for any classifiers to correctly identify the leaf class. Thus, dimensionality reduction method using feature selection will selects a subset of features from the available features to increase the performance of a classifier [1].

This paper is organized as follows. Section 2 briefly discusses the medicinal plant identification and classification, plant leaf shape based features, image processing and feature extraction for plant leaf, feature selection and classifier for medicinal leaf images. The experimental setup is discussed in section 3 and Section 4 discusses the results of the feature selection methods using DECIML as the classifier. Finally section 5 concludes this paper.

II. Medicinal Plant Leaf Identification and Classification

The apparent tasks in any automated identification and classification has been presented in previous section. This section brings several works and important attention in the focus of this research which is Malaysian medicinal plant leaf species identification and classification.

A. Related Works in Medicinal Plant Leaf Identification and Classification

Few attempts were done mainly in Unites States, India, China, Thailand and Indonesia. Thai herb leaf image recognition system has presented by [2] which consists of four main components: 1) image acquisition, 2) image preprocessing, 3) recognition and 4) display of results. The system applied several image-processing techniques and extracted 13 features from the leaf image and uses a k-nearest neighbor (k-NN) algorithm for recognition process. The experiment involved 32 species of Thai herbs, with more than 1,000 leaf images and they reported that the classification performance is 93.29%.

Reference [3] have studied and discussed an automatic recognition system of medicinal plants. With the leaf image recognition of medicinal plants as its core, the system applied the up-to-date technologies of image processing and neural network. There was no reported performance of the automatic recognition in their paper.

New method for Indonesian medicinal plants identification using the combination of some leaf features, i.e. texture, shape, and color was proposed in [4]. The feature extraction was performed based on the Local Binary Pattern Variance and the classification was performed by using a Probabilistic Neural Network classifier. In this research, they reported that the average accuracy of medicinal plant identification was 72.16%. The data used comprises of 51 different species of Indonesian medicinal plants, in which 48 different images were used to represent each species.

Furthermore, [5] proposed a new mobile application based on Android operating system for identifying Indonesian medicinal plant images based on texture and color features of digital leaf images. The accuracy performance of the identification process was reported to be 74.51% considering 51 different species of Indonesian medicinal plants with 48 images used to represent each species.

The development of Indonesian leaf recognition system is further studied by [6]. They have proposed a system called MedLeaf as a new mobile application for medicinal plants identification based on leaf image texture. Previous methods described in [4] were applied for the development in which 30 species of Indonesian medicinal plants with 48 leaf images each were used in the experiment. They reported that the accuracy performance of the medicinal plant identification process, based on leaf texture, was 56.33%.

Meanwhile in India, [7] presented an automated system that was able to recognize and classify medicinal plant leaves. This automated system comprises of 250 different leaf images obtained for five different species. The types of leaf features obtained include grey textures, grey tone spatial dependency matrices (GTSDM), and Local Binary Pattern (LBP) operators which generates statistical values. It was reported that the accuracy performance of the proposed automated system based on 70% training and 30% testing set was 94.7%.

In [8], a recognition approach using a MATLAB based Neural Network algorithm as a classifier that identifies the shape and texture features of the medicinal leaves was proposed. They discussed the identification and classification comparison of several leaf data that includes Hibiscus, Betel, Castor and Manathakali leaves. However, the number of leaves in the data used in their experiments was not presented.

Leaf-shape based plant identification was one of the earliest and most popular approached. It was considered as an important feature used to perform the leaf identification and classification process. Previous works specifically addressed leaf based identification process using different approaches.

B. *Plant Leaf Shape based Features*

Leaf image features are extracted mainly from shape information. Other features that can be extracted are vein patterns, colour and textures. Most of the previous and current leaf identification literatures utilize the whole leaf for feature extraction and to be used in the leaf identification process. Shape-based is the most popular approach for feature extraction as many of the researches show that this approach provides not only speed-up image processing but low cost and its conveniences [9].

Shape-based is one of the popular approaches used for feature extraction as it provides rich information for classification [10]. Efficient shape feature extraction should present several essential properties such as identifiability, scalability, affinity and occultation invariance, noise resistance, statistically independent and reliable [11]. The earliest work in leaf shape-based automated identification on specific leaf was started by [12], which involved the task of extracting the shape of the leaf (represented as grid) using a neural network algorithm for identification purposes.

Accelerated from the early shape-based, researchers began to introduce other techniques such as shape and centroid contour distance [13-15]. One of the most successful leaf identifications to date which was able to produce systematic leaf identification was designed based on the shape-based leaf identification. This approach used the inner-distance shape context approach [16-19]. Some of the refined works related to this approach have been conducted in [20]. Moving Median Center Hypersphere (MMC) was also introduced in the plant leaf identification technique [21, 22]. Recently, leaf identification based on shape and MMC has been used in solving leaf identification problem with a complicated background as described in [23]. Author in [24], discussed a plant leaf recognition using shape based features which provides accuracies ranging from 90%-100%. Recent study on shape-based leaf images recognition is presented by [25] and [26].

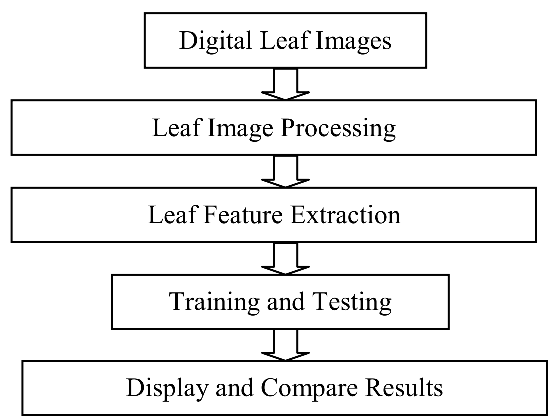

Figure 1. Flow diagram of the leaf image processing scheme.

C. *Image Processing and Feature Extraction for Plant Leaf*

Image processing is considered as one of the important steps in the digital leaf image processing scheme (Figure 1), in which it involves edge detection and thinning processes [9]. These techniques used in image processing are employed as it is much simpler due to less number of features extracted and processed.

There are two types of features described as the result of feature extraction step, shape tokens, which determine the number of features to be extracted, and the angle feature for the final data construction. The important part of this phase is the process of extracting tokens from the boundary line of the leaf image. Tokens are assigned to the boundary line of the leaf image based on the predefined distance between tokens. The shorter the distance between the tokens are, the more tokens will be assigned to the boundary of the leaf image as shown in Figure 2.

Angles of the tokens are the features where the values for cosines and sinus are computed according to the direction of the angle. As shown in the portion of the processed leaf image in Figure 3, the two adjacent tokens (P1 and P2) are used to define angles based on the direction of hypotenuse from both tokens. According to Figure 3, P1 and P2 are tokens and θ is the hypotenuse and C is the leaf centroid.

In this paper, features from five species with 65 leaves were extracted to produce 564 statistical values as feature (attributes) according to token distance at 1.

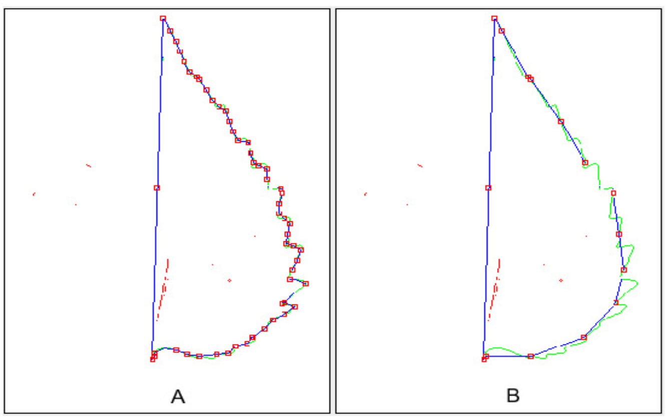

Figure 2. Different distance effects on the number of tokens assigned to the boundary of the leaf image. (A) More tokens if distance is 1 (configured as 10) and (B) less tokens if distance is 3 (configured as 30).

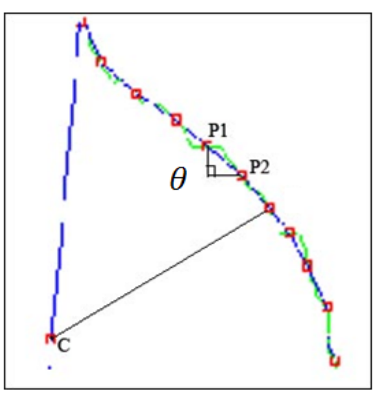

Figure 3. Portion of the processed leaf image and representation of feature extraction, θ, hypotenuse angle for token P1 and P2.

D. *Classifier for Medicinal Leaf Image Datas*

Several classifiers for leaf classification have been described in previous section, ranging from neural networks and k-nearest neighbor. In this paper, several ensemble classifiers from WEKA [27] were used to compare their classification accuracy over the leaf data. The selected classifiers obtained from WEKA includes AdaBoostM1 and Bagging with SMO and Random Forest as base classifiers, each using WEKA default settings.

In addition to that, an ensemble based classifier called a Direct Ensemble Classifier for Imbalanced Multiclass Learning (DECIML) is used to investigate the classifier performance. The classifier was proposed by [28] specifically to address the problem of multiclass classification with imbalance data. An imbalanced data with multiclass labels refers to a dataset with target class, which is skewed in distribution and poses a significant effect to classifier performance [29]. This is due to the fact that medicinal leaf data is not often balanced for all collected species, where some of the leaf samples are limited for specimen purpose. Thus, imbalance data in term of the number of samples for some leaves will exist. The researchers reported that the average accuracy using the DECIML on 16 imbalanced multiclass benchmark data was higher than the other tested single classifiers.

TABLE I. GA PROCESS FOR FEATURE SELECTION

Wrapper_GA(D, T, a_n, $Gain(S, A)$, method, ig, Max, p, $cross\_rate$, mut_rate)

1. A_{FS} ← initial attribute
2. Run feature selection optimization with GA
   - IF method = 1 THEN generate Pop population with p size using random features from D
   - ELSE IF method = 2, Call InitGen(ig, a_n) to generate population Pop with p size using features such that $gain(A, T{:}\,S) >= ig$ from D
   - FOR $iteration = 1$ to Max
     - FOR i= to population size, p
       ε ← Classify $t_i \in T$ using DECIML
       assign fitness $f(Pop[i]) = 1 - \varepsilon$
     - Pop' ← do selection, Pop' ← do crossover, Pop' ← do mutation, $Pop = Pop'$
     - IF convergence THEN, $A_{FS} = pop_i \in Pop \leftarrow \max_{i=1} f(Pop[i])$
       exit
     - ELSE $iteration = iteration + 1$
3. Return A_{FS}

E. *Feature Selection for Leaf Images Data*

Feature selection can be addressed in three general schemes; filter methods, wrapper methods and embedded methods. These broad schemes are best described by [30] and [1]. Wrapper methods associate the hypotheses search with the inductive classifiers to get the feedback whether the model selection is good. In this method, various combinations of feature subsets are generated and evaluated in order to improve the classification performance.

The study in this paper is using backward random selection with wrapper based feature selection methods. The reasons are 1) GA is working with all features and randomly selects (reduce) a set of features 2) GA will randomly add or remove any features from the feature set until an optimized subset is achieved. Thus, the effectiveness of applying the genetic-based wrapper feature selection methods for DECIML (DECIMLFS) classifier is investigated. Wrapper approach is developed and compared with two other feature selection methods available in Weka (CFsSubsetEval and Filteredsubseteval). The GA based wrapper feature selection algorithm (Table I) is adopted to optimize the possible combination of a certain number of attributes that best describe the dataset while maintaining higher classification rates. In Table I, D is set of d training examples with a_n features, T is set of t testing examples with a_n features, a_n is list of features in D, ig is the threshold from $Gain(S, A)$, $Gain(S, A)$ is information gain for each attribute a in D, method 1 = random; 2 = using ig, Max is maximum iteration for GA, p is population size, $cross\_rate$ is crossover rate, mut_rate represents mutation rate and A_{FS} is the set of optimal feature subset.

TABLE II. SELECTED LEAF SPECIES FOR THE EXPERIMENTAL DATA

Class	Example	Name	Train	Test
1		Cemumar	11	4
2		Kapal Terbang	12	4
3		Kemumur Itik	11	4
4		Lakom	5	4
5		Mengkudu	6	4

III. EXPERIMENTS DESIGN

The dataset for the experiment is obtained from villages situated in Perlis state where, 65 leaf samples are randomly selected from specified leaf species for the experimental data. The leaf sample size is selected in this preliminary study due to enormous time required to process the images without specific automated image processing. Table II is the list of leaf species selected in this paper.

A. Medicinal Leaf Images Data

In order to create the preliminary experimental data, basic image processing and feature extraction as described in previous section were implemented. Full leaf features based on token angles (cosines and sinus) is extracted. Initially, the image processing and feature extraction processes will produce different number of features. Thus, further data pre-processing is applied to find the largest number of features and leaf with fewer features will be filled with several dummy values on the remaining features (-1). Table III shows the description of the experimental data. Based on Table III, #Examples is the number of leaf image, #Attribute in Description column shows the largest number of leaf features using full leaf image processing, #Training and #Testing are number of leaf images as training and testing. The #Majority and #Minority represent the imbalance for the data.

TABLE III. SELECTED LEAF SPECIES FOR THE EXPERIMENTAL DATA

Description	Value#
#Examples	65
#Attributes	564
#Training	45
#Testing	20
#Majority	12
#Minority	5

Fundamentally, the data is small however it depicts a fairly high dimension where it poses the challenge of possible problem such as: 1) not enough data, 2) the "curse of dimensionality". Thus, this data can be used to investigate the effects of using imbalance data on the accuracy performances of all the classifiers used in this experiment. Therefore, the dataset constructed in this study is designed to show how the classifiers work on the available experimental data. Advanced tasks on image processing and feature extraction optimization will be left out for future work in this domain.

B. Comparing Feature Selection Methods

In order to provide comparisons, two feature selection methods from Weka are applied with its default parameters, namely CFsSubsetEval and FilteredSubsetEval. CFsSubsetEval evaluates the worth of a subset of attributes by considering the individual predictive ability of each feature along with the degree of redundancy between them. Thus, subsets of features that are highly correlated with the class while having low inter-correlation are preferred. FilteredSubsetEval provides two meta attribute selection evaluators that can apply an arbitrary filter to the input data before executing the actual attribute selection scheme. The methods are selected for their comparable performance as stated in [31].

C. Wrapper based Feature Selection with Random Population Generation

This method is also known as GA-based wrapper with random feature selection method. In this feature selection approach, all attributes are given to GA and feature subset is randomly generated. Optimization process takes place until a stopping condition is reached. DECIML is used as the induction algorithm to provide fitness to each subset in a generation. Minimum error (highest classification performance) will be selected as a feature selection subset and creates new dataset with selected features. Selected features will be evaluated using five classifiers mentioned above.

D. Wrapper based Feature Selection with ig Threshold Population Generation

This method is also known as GA-based wrapper with ig threshold feature selection method. Threshold ig is applied to select expected "good" attributes to minimize the problem of random population generation. Therefore, GA will only optimize a set of feature subset with $\mathrm{gain}(\mathrm{A}, \mathrm{T}: \mathrm{S}) >= ig$. Again, DECIML is used as the induction algorithm.

E. Genetic Algorithm Settings

Based on several feature selections using GA such as [32-33], range of values for settings are as follows: population size [10-1000], crossover rate [0.6-1.0], mutation rate [0.001-0.25] and the number of generations [20-50]. In the experiment, the population size is set to 50, the mutation rate is set to 0.001, crossover rate is set to 0.6, and number of generation is 100. Moreover, the genetic operators are roulette wheel selection, one-point crossover and standard mutation.

IV. RESULTS

A. Classification performance of Medicinal Leaf Data without Feature Selection

Table IV shows the performance of DECIML classifier compared to several ensemble method from WEKA (AdaBoostM1(AB) using Support Vector (SMO) and Random Forest (RF), and Boosting (BO) using SMO and RF). The ensemble classifier DECIML (DE) is the target classifier in this paper.

TABLE IV. AVERAGE PERFORMANCE OF THE CLASSIFIERS

Classifier	**AB (SMO)**	**AB (RF)**	**BO (SMO)**	**BO (RF)**	**DE**
Accuracy %	50	45	55	50	**65**

The results shown in Table IV indicate that the experimental data is challenging and the extracted features from the leaf are not enough to describe the domain problem. In addition, the data has not been further preprocessed (discretization, etc.). However, the objective of the study is to create an experimental data based on leaf shape features is achieved. The DECIML classifier which is an ensemble approach to classify multiclass data with imbalance performs slightly better than the other classifiers, while the ensemble methods using AdaboostM1 and Bagging almost perform similar using all features.

B. Classification performance of Medicinal Leaf Data using Feature Selection

Wrapper based feature selection using GA and DECIML as the induction algorithm is applied to investigate if the classification of the data can be improved. In this method, GA is estimated to produce the combination of optimized feature subset selection. Two experiments conducted to compare the performance of the random population generation (DECIMLFS.WR) and ig based threshold population generation (DECIMLFS.WIG). Another two experiments are using CFsSubsetEval and FilteredSubsetEval of Weka.

TABLE V. CLASSIFICATION PERFORMANCE AND REDUCTION OF ATTRIBUTE (#ATT.FS) USING WRAPPER BASED FEATURE SELECTION (DECIMLFS.WR, DECIMLFS.WIG), CFSSUBSETEVAL AND FILTEREDSUBSETEVAL.

Feature Selection	**#ATT**	**#ATT.FS**	**Accuracy (%)**
DECIMLFS.WR	564	350	80
DECIMLFS.WIG	564	181	85
CFsSubsetEval	564	8	65
FilteredSubsetEval	564	8	65

The results in Table V indicate that the experimental data is challenging and in another point of view (leaf classification and identification community), the samples and extracted features from the leaf (angle of the leaf shape tokens) may not enough to describe the domain problem. However, the objective of this study to create an experimental data based on leaf shape features is achieved in testing the performance of the wrapper feature selection (DECIMLFS). The DECIMLFS.WIG managed to produce better classification (85%) with 181 selected features. Although that the two feature selection methods from Weka produce fewer features, classification performance is low. It shows that by specifying the value of ig in $\mathrm{gain}(\mathrm{A}, \mathrm{T}: \mathrm{S}) >= ig$ provides better feature set for optimization. Best ig in the study is set to 0.417.

V. CONCLUSION

Malaysian medicinal leaf image recognition and classification is important for its knowledge. Future generation could easily forget the plants and benefits from these traditional remedies. This paper will motivate more research in this domain in order to preserve the knowledge with the help of computing technologies. In this paper, basic image processing and feature extraction methods have been performed to prepare the Malaysian medicinal leaf data. Several classifiers including the DECIML have been applied to the data. However, the classification performance is still considered as low. It is due to the small data and large feature. Feature selection was applied and the results are promising. In future works, more Malaysian medicinal plant leaves will be collected to create pool of leaf data. Advanced image processing and other feature extraction approach will also be investigated. Improvement of the DECIML classifier and DECIMLFS feature selection method will be studied to match with Malaysian medicinal leaf images identification and classification research.

REFERENCES

[1] Ladha, L., and Deepa, T.: 'Feature selection methods and algorithms', International Journal on Computer Science and Engineering (IJCSE), 2011, 3, pp. 1787-1797.

[2] Pornpanomchai, C., Rimdusit, S., Tanasap, P., and Chaiyod, C.: 'Thai herb leaf image recognition system (THLIRS)', Nat. Sci., 2011, 45, (551-562).

[3] Gao, L., and Lin, X.: 'A study on the automatic recognition system of medicinal plants'. Proc. 2nd International Conference on Consumer Electronics, Communications and Networks (CECNet), Yichang. 2012.

[4] Herdiyeni, Y., and Santoni, M.M.: 'Combination of morphological, local binary pattern variance and color moments features for Indonesian medicinal plants identification'. Proc. International Conference onAdvanced Computer Science and Information Systems (ICACSIS2012), Depok, 2012.

[5] Herdiyeni, Y., and Wahyuni, N.K.S.: 'Mobile application for Indonesian medicinal plants identification using fuzzy local binary pattern and fuzzy color histogram'. Proc. International Conference onAdvanced Computer Science and Information Systems (ICACSIS2012), Depok, 2012.

[6] Prasvita, D.S., and Herdiyeni, Y.: 'Medleaf mobile application for medicinal plant identification based on leaf image', Advanced Science Enginering Information Technology, 2013, 2, (2), pp. 5-8.

[7] Arun, C.H., and Emmanuel, W.R.S.: 'Texture feature extraction for identification of medicinal plants and comparison of different classifiers', International Journal of Computer Applications, 2013, 62, (12).

[8] Ananthi, C., Periasamy, A., and Muruganand, S.: 'Pattern recognition of medicinal leaves using image processing techniques ', Journal of NanoScience and NanoTechnology, 2014, 2, (2), pp. 214-218.

[9] Langner, J. (2006). Neuronal Network based recognition system of leaf images. Retrieved 24 Febuary 2009, 2009, from http://www.jens-langner.de/lrecog/.

[10] Xiao, X.-Y., Hu, R., Zhang, S.-W., and Wang, X.-F.: 'HOG-Based approach for leaf classification', in Huang, D.-S., Zhang, X., Reyes García, C., and Zhang, L. (Eds.): 'Advanced Intelligent Computing Theories and Applications. With Aspects of Artificial Intelligence' (Springer Berlin / Heidelberg, 2010), pp. 149-155.

[11] Mingqiang, Y., Kidiyo, K., and Joseph, R.: 'A survey of shape feature extraction technique', Pattern Recognition, 2008, 43, (90).

[12] Heymans, B.C., Onema, J.P., and Kuti, J.O.: 'A neural network for opuntia leaf-form recognition'. Proc. IEEE International Joint Conference on Neural Networks, 1991.

[13] Hong, F., Zheru, C., Dagan, F., and Jiatao, S.: 'Machine learning techniques for ontology-based leaf classification', in Editor (Ed.)^(Eds.): 'Book Machine learning techniques for ontology-based leaf classification' (2004, edn.), pp. 681-686 Vol. 681.

[14] Wang, Z., Chi, Z., and Feng, D.: 'Shape based leaf image retrieval', Vision, Image and Signal Processing, IEE Proceedings -, 2003, 150, (1), pp. 34-43.

[15] Zhiyong, W., Zheru, C., and Dagan, F.: 'Fuzzy integral for leaf image retrieval', in Editor (Ed.)^(Eds.): 'Book Fuzzy integral for leaf image retrieval' (2002, edn.), pp. 372-377.

[16] Agarwal, G., Ling, H., Jacobs, D., Shirdhonkar, S., Kress, W.J., Russell, R., Belhumeur, P., Dixit, A., Feiner, S., Mahajan, D., Sunkavalli, K., Ramamoorthi, R., and White, S.: 'First steps toward an electronic field guide for plants', Taxon, 2006, 55, (3), pp. 597-610.

[17] Ling, H., and Jacobs, D.W.: 'Using the inner-distance for classification of articulated shapes', in Editor (Ed.)^(Eds.): 'Book Using the inner-distance for classification of articulated shapes' (2005, edn.), pp. 719-726 ,vol. 712.

[18] White, S., Marino, D., and S., F.: 'LeafView: A User Interface for Automated Botanical Species Identification and Data Collection'. Proc. ACM UIST 2006 Conference Companion, Montreux, Switzerland, October 15-18, 2006.

[19] White, S.M., Marino, D., and Feiner, S.: 'Designing a mobile user interface for automated species identification'. Proc. Proceedings of the SIGCHI conference on Human factors in computing systems, San Jose, California, USA, 2007.

[20] Belhumeur, P.N., Chen, D., Feiner, S., Jacobs, D.W., Kress, W.J., Ling, H., Lopez, I., Ramamoorthi, R., Sheorey, S., SeanWhite, and Zhang, L.: 'Searching the world's herbaria: a system for visual identification of plant species', (Springer-Verlag, 2008, edn.), pp. 116–129.

[21] Du, J.-X., Wang, X.-F., and Zhang, G.-J.: 'Leaf shape based plant species recognition', Applied Mathematics and Computation, 2007, 185, (2), pp. 883-893.

[22] Guo-Jun, Z., Xiao-Feng, W., De-Shuang, H., Zheru, C., Yiu-Ming, C., Ji-Xiang, D., and Yuan-Yuan, W.: 'A hypersphere method for plant leaves classification', in Editor (Ed.)^(Eds.): 'Book A hypersphere method for plant leaves classification' (2004, edn.), pp. 165-168.

[23] Wang, X.-F., Huang, D.-S., Du, J.-X., Xu, H., and Heutte, L.: 'Classification of plant leaf images with complicated background', Applied Mathematics and Computation, 2008, 205, (2), pp. 916-926.

[24] Chaki, J., and Parekh, R.: 'Plant Leaf Recognition using Shape based Features and Neural Network classifiers ', International Journal of Advanced Computer Science and Applications, 2011, 2, (10), pp. 41-47.

[25] Mouine, S., Yahiaoui, I., and Verroust-Blondet, A.: 'Advanced shape context for plant species identification using leaf image retrieval'. Proc. Proceedings of the 2nd ACM International Conference on Multimedia Retrieval, Hong Kong, China, 2012.

[26] Mouine, S., Yahiaoui, I., and Verroust-Blondet, A.: 'A Shape-based Approach for Leaf Classification using Multiscale Triangular Representation'. Proc. ICMR'13, Dallas, Texas, USA, 2013.

[27] Witten, I., and Frank, E.: 'Data mining: practical machine learning tools and techniques with java implementations' (Morgan Kaufmann, 2000).

[28] Sainin, M.S., and Alfred, R.: 'A direct ensemble classifier for imbalanced multiclass learning'. Proc. Data Mining Optimization (DMO2012), Langkawi, 2012.

[29] Yang, Q., and Wu, X.: '10 challenging problems in data mining research', International Journal of Information Technology and Decision Making, 2006, 5, (4), pp. 597-604.

[30] Saeys, Y., Inza, I., and Larranaga, P.: 'A review of feature selection techniques in bioinformatics', Bioinformatics, 2007, 23, (19), pp. 2507-2517.

[31] Cuaya, G., Munoz, A., and Morales, E.: 'A minority class feature selection method', In Proc. of the 16th. Iberoamerican Congress on Pattern Recognition CIARP-2011, 2011.

[32] Jayaram, M.A., Karegowda, A.G., and Manjunath, A.S.: 'Feature subset selection problem using wrapper approach in supervised learning', International Journal of Computer Applications 2010, 1, (7), pp. 13-16.

[33] Tiwari, R., and Singh, M.P.: 'Correlation-based attribute selection using genetic algorithm', International Journal of Computer Applications, 2010, 4, (8), pp. 28-34.