Plant Leaf Recognition Using Competitive Based Learning Algorithm

E M Imah[1*], Y S Rahayu[2] and A Wintarti[1]

[1]Mathematics Department, Universitas Negeri Surabaya, Surabaya, East Java, Indonesia
[2]Biology Department, Universitas Negeri Surabaya, Surabaya, East Java, Indonesia

*ellymatul@unesa.ac.id

Abstract. Plant recognition based on digital leaf image has received as particular attention in computer vision and intelligence system, due its important implication in automatic plant identification. Plant species have the unique leaf characteristics such as the shape, texture, margin, and colour, which different each other. This study presents a novel method for automation plant recognition using Generalized Relevance Learning Vector Quantization (GRLVQ). GRLVQ is a competitive based learning algorithm which is integrating features extraction and classification phases. The experimental result shows that GRLVQ has better performance than the predecessor algorithm.

1. Introduction

Classification of plants is basis of Botany science; it is also foundation of plant genetics, plant ecology, plant medicine, and file science. Traditional methods plants classification methods are mainly depend on researcher's subject judgment [1]. Moreover, it is difficult to satisfy the need that people want to quickly identify the plants; therefore, automatic plant recognition has been needed.

Study on plants recognition based on image processing has been rapid development by collaborating the biologists especially botanist, and computer scientist. Many researchers were drawing their interest in computer's automatic plants identification. Different with traditional plant classification, this method is rapid and not depending on the person's subjective judgment. Using the stat-of-the art Machine Learning methods, this task will be more simple and fast, so it able to help people to knowing the plant quickly.

Automatic plant recognition is still challenges task, not only in classification phases, but also in feature extraction and image preprocessing phases. Some researchers have published their study that focus on feature extraction. B V Lakshmi and her team, studying plant leaf detection based on digital leaf image by using midpoint circle algorithm [2]. J Chaki et al also focus on feature extraction of automatic plant recognition, they used Ridge Rilter and Curvelet Transform to get the feature of images then classifying it using Neuro-Fuzzy Classifier [3]. Their study focuses on feature extraction phases. S Sladojevic et al study plant disease classification based on digital leaf image using deep learning algorithm [4]. Plant leaf has special character in texture, shape, and color features, and it challenges in digital image processing and pattern recognition study, how to get those feature automatically [5]. Ji-xiang Du et all use fractal feature in their study [6], X Wang using dual scale decomposition for feature extraction, it is based on spatial domain feature extraction [7].

The machine learning study in automatic plant classification also interesting and many researchers have paid interest in this task. N Ahmed has studied automatics plant recognition based on leaf images using Support Vector Machine (SVM), he's study present that the accuracy of 16 different plants species is 87% [8]. A Hasyim et all published their study on plant shape recognition using Probabilistic Neural Network (PNN), their study shows good accuracy, but it only classify four different shapes [9]. A Kadir used ANN for Filio plant classification [10], V Lakshmi use Kernel PSO and FRVM classifier for automatic plant detection with some feature extraction methods [5]. Based on the literatures in this study KNN show good performance but this algorithm is memory costly. SVM or kernel based algorithm is good performance but it is also costly and complex algorithm, many parameters that have to try, similarly with Backpropagation ANN.

Based on those literatures, this study we use Generalized Relevance Learning Vector Quantization (GRLVQ) as classification algorithm because this algorithm is rarely used for automatic plant recognition, although this algorithm is very powerful especially for multiclass classification problem. The result of GRLVQ will be compared to the other algorithm that has been used by the other researchers and report has good performance. GRLVQ is competitive based learning that using prototypes of each classes to classify. Prototype is determining in the training process from training dataset and capture the essential features of the data in the same space [9]. GRLVQ is a modification of Relevance Learning Vector Quantization (RLVQ) by using adaptive metric and very powerful to do task in classification. GRLVQ is proposed by Hammer et all and used stochastic gradient descent on an energy function [11]. This study used UCI leaf dataset which is collected by James Cope of Royal Botani Garden.

This paper organized as follow, section II describe plant leaf dataset and preprocessing methods that used in this study. Section III describes the basic concept of machine learning classification algorithm that used. Section IV present the experimental setup, result, and discuss of the study. Section V is presents the conclusion of the study.

2. Digital plant leaf dataset and pre-processing

2.1. Digital plant leaf dataset

UCI plant leaf dataset comprise one-hundred species of leaves, for each species there are sixteen distinct specimens. The original image is colour image on white background [12]. This dataset very challenging because contain one-hundred classes. Multiclass classification task is big task in machine learning, because many algorithms lacks to classify the dataset in big classes data. In order to preliminary study, this study used fifteen species only; the detail can be seen on Figure 1. Digital plant leaf images in Figure 1 are image which are after image pre-processing.

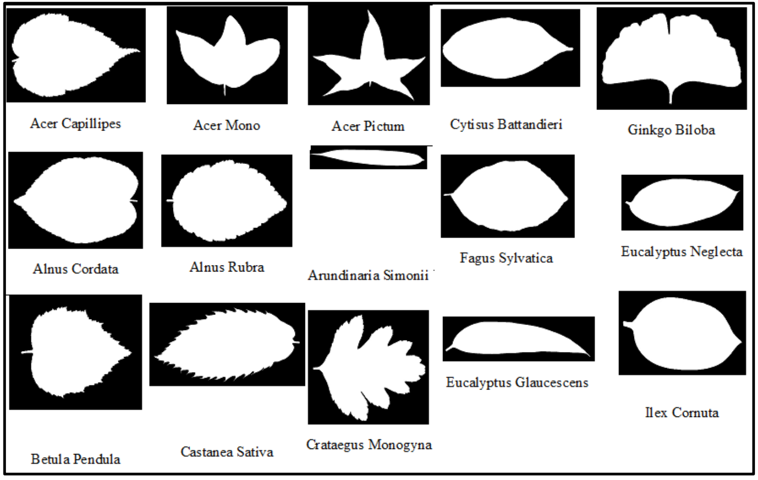


Figure 1. Fifteen plant species

2.2. Digital plant leaf pre-processing

Digital image pre-processing is an important phase in automatic plant leaf recognition because there are some noises in digital image data that may arise from improper illumination or some Gaussian white noises. Good noise reduction methods able to result better recognition accuracy. Digital image pre-processing that was used in this study can be seen in Figure 2. Original input image was converted to grey scale image, then get the edge of the leaf using edge detection methods. After edge detection, the image leaf was smoothing, then get the point of shape of leaf. The colour image was extracted using histogram image data. The point set of shape and histogram colour data was extracted and used them as extracted feature for classification phase to build the model of plant recognition system.

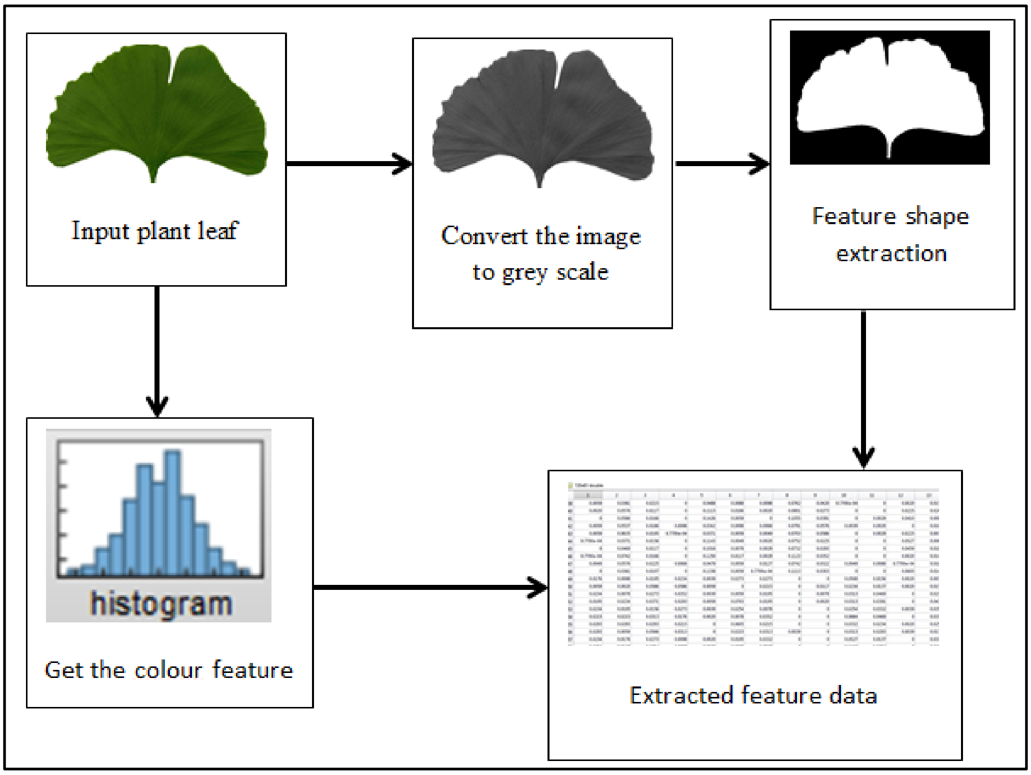


Figure 2. Pre-processing Phase

3. Machine learning classification algorithm

3.1. Random forest

Random forest algorithm is classification algorithm that basically based on random tree. In random forest, every input feature vector is compared to the one stored in the train dataset in order to find the best match. Growing an ensemble of random trees for recognition using a probabilistic scheme is called random forest of trees. Recognition accuracy is high as the trees vote for the most popular class. Trees drawn at random from a set of possible trees is called random tree. Random tree is a decision tree that considers *k* randomly chosen attributes at each node. The class probabilities on each node are based on back fitting with no pruning [13]. The steps involved in growing a random tree are as follow:

1. The training set for growing the tree is obtained by selecting *N* cases at random but with replacement from original dataset.
2. A random number of attributes m are chosen for each tree. The attributes from the nodes and leaves using standard tree building algorithms. The best split on *m* is used to split the nodes and *m* is held constant.
3. Each tree is growing to the fullest extent possible without pruning.

A new object is classified using its input vector down each of the trees in the forest. The forest chooses the class with the most vote, the new object input vector is classified.

3.2. Support Vector Machine (SVM)

The simplest version of a SVM is called Maximal Margin Classifier, which is applicable for linearly separable data. It is simple to understanding the basic ideas behind more sophisticated SVMs. Consider a linearly separable dataset {(X_i, d_i)}, where X_i is the input pattern for the i-th example and d_i is the corresponding desired output {-1, 1}. The assumption, ''the dataset is linearly separable", means there exist a hyper plane working as the decision surface. We can write:

$$W^T X_i + b \geq 0, \text{then} d_i = +1$$
$$W^T X_i + b \leq 0, \text{then} d_i = -1 \quad (1)$$

Where $W^T X_i + b,$ is the output function. The distance from the hyper plane to the closest point is called the geometric margin. The idea is, to have a good machine, so the geometric margin needs to be maximized. First, we introduce the marginal function $W^T X_i + b$ because the dataset is linearly separable we can rewrite as (2), as follow:

$$W^T X_i + b = +1$$
$$W^T X_i + b = -1 \quad (2)$$

Where $X^+(X^-)$ is the closest data point on the positive (negative) side of the hyperplane. Now it is straight forward to compute the geometric margin.

$$\begin{aligned} \gamma &= \frac{1}{2}\left(\frac{W^T X^+ + b}{|w|} - \frac{W^T X^- + b}{|w|}\right) \\ &= \frac{1}{2|w|})W^T X^+ + b - W^T X^- - b) \\ &= \frac{1}{2|w|}\big(1 - (-1)\big) = \frac{1}{|w|} \end{aligned} \quad (3)$$

Hence, equivalent to maximize the geometric margin is fixing the functional margin to one and minimizing the norm of the weight vector |w|. This can be formulated as a quadratic problem with inequality constraints

$$d(w^T x_i + b) \geq 1.$$

$$\text{min: } \tfrac{1}{2} W^T W \text{ (quadratic-problem)} \quad (4)$$

subject to: $d(w^T x_i + b) \geq 1$

By the use of Lagrange multipliers$\alpha_i \geq 0$ the original problem is transformed into the dual problem. From the Kuhan–Tuker theory we have the following condition:

$$\alpha_i [d_i (W^T x_i + b) - 1] = 0 \quad (5)$$

It means only the points with functional margin unity are contributing to the output function. These points are called the Support Vectors, which are supporting the separating hyper plane.

3.3. Generalized Relevance Learning Vector Quantization (GRLVQ)

GLRVQ is a competitive based learning classification algorithm that modified of GLVQ. GLVQ has proposed by A. Sato Yamada, using steepest descent method which minimizes a cost function to defined the codebook or prototype vectors update [14]. Relative distance difference is defined as (6) and cost function as (7).

$$\mu(x) = \frac{d_j - d_k}{d_j + d_k} \tag{6}$$

$$S = \sum_{i=1}^{N} f(\mu(x_i)) \tag{7}$$

Where N is number of input vector and f is a monotonically increasing function. GRLVQ the distance was modified using weighted distance between input vector x_i and a codebook vector w_j [15]:

$$D_{ij} = \sqrt{\sum_{k=1}^{N} \lambda_k (x_{ik} - w_{jk})^2} \tag{8}$$

Where $\sum_{k=1}^{N} \lambda_k = 1$. Modification of distance formula, Eq. (6) must be reformulated to minimized and objective function based on this modified distance as in (9).

$$\mu_\lambda(x_i) = \frac{D_{ij} - D_{ik}}{D_{ij} + D_{ik}} \tag{9}$$

Obtained a modified rule of GLVQ, which is the GRLVQ rule:

$$\Delta w_j = \pm \eta \lambda \mathrm{I} \frac{\partial f}{\partial \mu} \frac{D_{ij}}{(D_{ij} + D_{ijk})^2} (x_i - w_j) \tag{10}$$

If x_i and w_k are different classes, the sign of Δw_j is (+), and if different classes is (-).

The relevance is updating using Eq. 12.

$$\lambda^{(t+1)} = \lambda^{(t)} - \alpha \frac{1}{4\sigma^2} G(y_1 - y_2, 2\sigma^2 \mathrm{I}).(y_2 - y_1)\mathrm{I}.(x_1 - w_{j(1)} - x_2 + w_{j(2)}) \tag{11}$$

Update on-line both the relevance and feature ranks algorithm as follow:

1. Initialize , α, and relevance vector $\lambda_k = \frac{1}{n}, k = 1, \ldots, n$.
2. Initialize codebook vector.
3. Update codebook vector using Eq. (10).
4. Update the relevance vector using Eq. (11).
5. Normalize the relevance vector.
6. Compute the weight of each feature as an average of its before ordering position index in the input vector, for all previous steps.

Repeat step 3-6 for each training pattern.

4. Results and discussion

4.1Experimental setup

The dataset which are used in this experiment is UCI plant leaf dataset comprise one-hundred species of leaves, for each species there are sixteen distinct specimens. The original image is color image on white background [12]. This dataset very challenging because contain one-hundred classes. We only used fifteen classes, and the sample of digital image leaf can be seen on Figure 1. The dataset consist of 64 attributes, every class consist of 48 samples. The training and testing data ratio in this study are 2:1.

4.2Experimental result and discuss

The experiment in this study to do task for recognition fifteen plant leaf image GRLVQ in full feature condition and compare the result with the others classification algorithms. The detail of the accuracy, precision, and recall, can be seen on Table 1 for Random Forest, Table 2 for SVM, and Table 3 for GRLVQ.

Table 1. Accuracy of Random Forest for Automatic 15 different plant leaf species

	FP rate	Precision	Recall	F measure
Acer Capillipes	0.004	0.938	0.833	0.882
Acer Mono	0.013	0.833	0.938	0.882
Acer Pictum	0.004	0.938	1	0.968
Alnus Cordata	0.009	0.895	0.85	0.872
Alnus Rubra	0.009	0.875	1	0.933
Arundinaria Simonii	0	1	0.944	0.971
Betula Pendula	0.017	0.75	0.8	0.774
Castanea Sativa	0.009	0.867	0.929	0.897
Crataegus Monogyna	0.009	0.867	0.813	0.839
Cytisus Battandieri	0.013	0.842	1	0.914
Eucalyptus Glaucescens	0.004	0.947	0.9	0.923
Eucalyptus Neglecta	0.022	0.722	0.765	0.743
Fagus Sylvatica	0	1	0.895	0.944
Ginkgo Biloba	0.013	0.727	0.615	0.667
Ilex Cornuta	0	1	0.933	0.966
Weighted Avg.	**0.008**	**0.885**	**0.882**	**0.882**

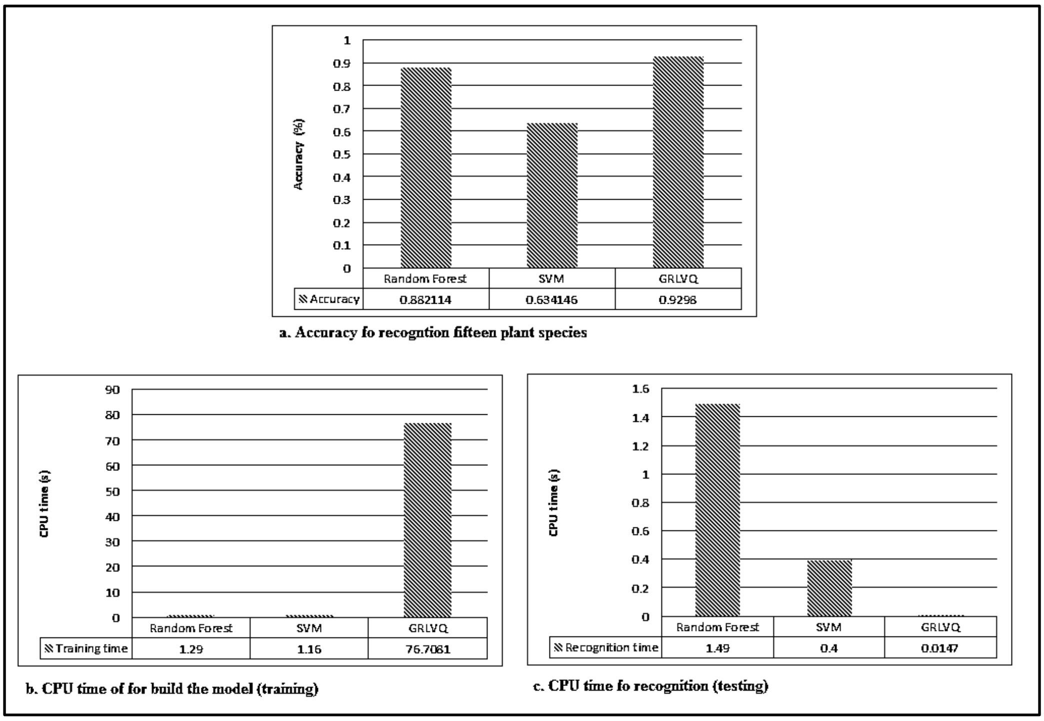
Table 2. Accuracy of SVM for Automatic 15 different plant leaf species

	FP rate	Precision	Recall	F measure
Acer Capillipes	0.009	0.846	0.611	0.71
Acer Mono	0	1	0.813	0.897
Acer Pictum	0	1	0.667	0.8
Alnus Cordata	0.004	0.929	0.65	0.765
Alnus Rubra	0	1	0.643	0.783
Arundinaria Simonii	0	1	0.5	0.667
Betula Pendula	0.004	0.909	0.667	0.769
Castanea Sativa	0.004	0.9	0.643	0.75
Crataegus Monogyna	0	1	0.5	0.667
Cytisus Battandieri	0.365	0.152	0.938	0.261
Eucalyptus Glaucescens	0	1	0.65	0.788
Eucalyptus Neglecta	0	1	0.706	0.828
Fagus Sylvatica	0	1	0.421	0.593
Ginkgo Biloba	0	1	0.385	0.556
Ilex Cornuta	0.004	0.917	0.733	0.815
Weighted Avg.	**0.026**	**0.911**	**0.634**	**0.71**

Table 3. Accuracy of GRLVQ for Automatic 15 different plant leaf species

	FP rate	**Precision**	**Recall**	**F measure**
Acer Capillipes	0.0038	0.9412	0.8421	0.8889
Acer Mono	0	1	1	1
Acer Pictum	0	1	1	1
Alnus Cordata	0.0038	0.95	1	0.9744
Alnus Rubra	0.0075	0.90	0.9474	0.9231
Arundinaria Simonii	0	1	1	1
Betula Pendula	0	1	0.9474	0.9730
Castanea Sativa	0.0113	0.8500	0.8947	0.8718
Crataegus Monogyna	0.0075	0.9048	1.0000	0.95
Cytisus Battandieri	0	1	0.9474	0.9730
Eucalyptus Glaucescens	0	1	0.7895	0.8824
Eucalyptus Neglecta	0.0075	0.8824	0.7895	0.8333
Fagus Sylvatica	0	1	0.7895	0.8824
Ginkgo Biloba	0.0301	0.7037	1	0.8261
Ilex Cornuta	0.0038	0.95	1	0.9744
Weighted Avg.	**0.0051**	**0.9380**	**0.9298**	**0.9270**

The result shows that GRLVQ has good performance better than the others. The precision and recall of GRLVQ are 0.9380 and 0.9298; GRLVQ also has a good performance in all classes. Random forest precision and recall are 0.885 and 0.882, and then SVM precision and recall are 0.911 and 0.634. The detail of experiment result can be seen on Table 1, Table 2, and Table 3.



a. Accuracy fo recogntion fifteen plant species

b. CPU time of for build the model (training)

c. CPU time fo recognition (testing)

Figure 3. Accuracy and CPU time

Overall accuracy and CPU time can be seen on Figure 3. Overall accuracy of recognition that present on figure 3.a can be seen that GRLVQ has the best accuracy among 92.98%, higher almost 5% than random Forest and more than 25% than SVM. Figure 3.b. shows that Random Forest able to build the models fastest, it is better than the others classification methods that was compared in this study. SVM is the longest in building the models. The fastest algorithm in recognition is GRLVQ. The testing time or recognition time of GRLVQ is less than 0.1 seconds; it is the fastest than SVM, Random Forest, or Backpropagation, because GRLVQ the simplest algorithm that only used minimum distance of data to prototype. The details can be seen on Figure 3.c. GRLVQ has the best performance is in full feature condition than extracted features. GRLVQ used weighted distance that generally is proposed for selected feature, therefore if feature of the input data have been extracted or reduced some information are lose for selection by relevance factor in GRLVQ.

5. Conclusion

This study examined various classification algorithm for automatically classifying fifteen classes of plant species based on digital image leaf. The experiments are using full feature attributes than classify using classification algorithms: Random Forest, SVM, and GRLVQ. The training and testing data ratio in this study are 2:1. The best performance in recognizing the five classes in EEG epileptic seizure dataset is GRLVQ, with the accuracy among 92.98%, higher almost 5% than random Forest and more than 25% than SVM. Moreover, GRLVQ is also has fastest CPU Time for recognition the plant species.

References

[1] Wang Z, Li H, Zhu Y, Xu T 2016 Review of Plant Identification Based on Image Processing. Arch Comput Methods Eng [Internet]. Available from: http://link.springer.com/10.1007/s11831-016-9181-4

[2] Vijaya Lakshmi B, Mohan V 2017 Plant leaf image detection method using a midpoint circle algorithm for shape-based feature extraction *J Mod Appl Stat Methods* [Internet]. 2017;16(1):461–80. Available from: http://digitalcommons.wayne.edu/jmasm/vol16/iss1/26

[3] Chaki J, Parekh R, Bhattacharya S. Plant 2017 Leaf Recognition Using Ridge Filter and Curvelet Transform with Neuro-Fuzzy Classifier. In Springer, New Delhi; 2016 [cited 2017 Aug 2]. p. 37–44. Available from: http://link.springer.com/10.1007/978-81-322-2538-6_5

[4] Sladojevic S, Arsenovic M, Anderla A, Culibrk D, Stefanovic D 2016 Deep Neural Networks Based Recognition of Plant Diseases by Leaf Image Classification. Comput Intell Neurosci [Internet]. 2016 Jun 22 [cited 2017 Aug 2];2016:1–11. Available from: http://www.hindawi.com/journals/cin/2016/3289801/

[5] VijayaLakshmi B, Mohan V 2016 Kernel-based PSO and FRVM: An automatic plant leaf type detection using texture, shape, and color features. Comput Electron Agric [Internet]. 2016;125:99–112. Available from: http://dx.doi.org/10.1016/j.compag.2016.04.033

[6] Du J, Zhai C-M, Wang Q-P 2013 Recognition of plant leaf image based on fractal dimension features. Neurocomputing [Internet]. 2013 Sep [cited 2017 Aug 2];116:150–6. Available from: http://linkinghub.elsevier.com/retrieve/pii/S0925231212007163

[7] Wang X, Liang J, Guo F 2014 Feature extraction algorithm based on dual-scale decomposition and local binary descriptors for plant leaf recognition. Digit Signal Process [Internet]. 2014 Nov [cited 2017 Aug 2];34:101–7. Available from: http://linkinghub.elsevier.com/retrieve/pii/ S105120041400253X

[8] Ahmed N, Khan UG, Asif S, Lahore T 2016 an Automatic Leaf Based Plant Identification

System. **28** 1 427–30

[9] Hasim A, Herdiyeni Y, Douady S 2016 Leaf Shape Recognition using Centroid Contour Distance. *IOP Conf Ser Earth Environ Sci* Available from: http://stacks.iop.org/1755-1315/31/i=1/a=012002?key=crossref.762da034f4c85b67bafd70e8b8462763

[10] Kadir A, Nugroho LE, Susanto A, Santosa PI 2013 Neural Network Application on Foliage Plant Identification. [cited 2017 Aug 2]; Available from: http://arxiv.org/abs/1311.5829

[11] Kästner M, Hammer B, Biehl M, Villmann T 2012 Functional relevance learning in generalized learning vector quantization *In: Neurocomputing* 1–22.

[12] James O C, Mallah J C 2013 Plant leaf classification using probabilistic integration of shape, texture and margin features *Signal Process Pattern Recogni- tion Appl.* 3–5

[13] Belgacem N. ECG 2012 Based Human Authentication using Wavelets and Random Forests *Int. J. Cryptogr Inf. Secur.* **2** 2 1–11

[14] Sato A, Yamada K 1996 Generalized Learning Vector Quantization. In: Touretzky DS, Mozer MC, Hasselmo ME, editors. *Advances in Neural Information Processing Systems 8 Proceedings of the 1995 Conference. MIT Press* 423–9

[15] Caţaron A, Andonie R 2004 Energy generalized LVQ with relevance factors *IEEE Int Conf Neural Networks - Conf Proc.* 2 1421–6