Journal of Asian Scientific Research, 2014, 4(4):159-173

Journal of Asian Scientific Research

journal homepage: http://www.aessweb.com/journals/5003

SUPPORT VECTOR MACHINE BASED TOOL FOR PLANT SPECIES TAXONOMIC CLASSIFICATION

Manimekalai .K

M.Phil Research Scholar, PSGR Krishnammal College for Women, Coimbatore

Vijaya.MS

Associate professor, G.R. Govindarajulu School of Applied Computer Technology, Coimbatore

ABSTRACT

Plant species are living things and are generally categorized in terms of Domain, Kingdom, Phylum, Class, Order, Family, Genus and name of Species in a hierarchical fashion. This paper formulates the taxonomic leaf categorization problem as the hierarchical classification task and provides a suitable solution using a supervised learning technique namely support vector machine. Features are extracted from scanned images of plant leaves and trained using SVM. Only class, order, family of plants and species are considered for hierarchical classification in this research work. The trained models corresponding to class - Magnoliopsida, orders - Brassicales, Rosales, families - Cariaceae, Brassicacea, Rosaceae, Rhamnaceae have been used to develop a three level hierarchical classification model for hierarchical classification of plant species and the results are analyzed. A flat multiclass classifier has been built for species which is at last level of hierarchy and the results are analyzed. A user interactive taxonomic tool has developed for displaying taxonomic ranks of species.

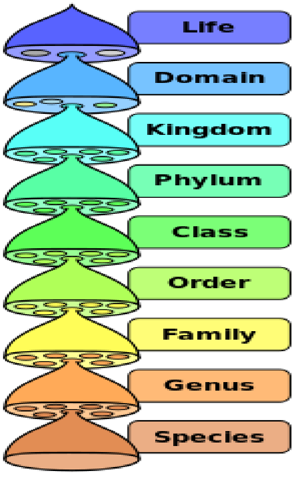
Keywords: Pre-processing, Feature extraction, Hierarchical classification, Multi-class classification, Plant species taxonomy, Support vector machine.

1. INTRODUCTION

Plant identification is the process of matching a specimen plant to a known taxon [1] and it is an fundamentally challenging. The herbal plays a vital role in siddha medicals and the prediction of plants helps to identify whether the plant is herbal or not. The species prediction makes ease of knowing the counting of living organisms. Natural resource management refers to the management of natural resources such as land, water, soil, plants and animals using species prediction process with a particular focus on how management affects the quality of life for both present and future generations. The field of ecology and geography uses the process of identifying of species. The taxonomic ranks of plant species are shown in Fig.1.

`

Fig-1. Taxonomic Classification



1.1. Kingdom

In biology, kingdom is a taxonomic rank that is composed of smaller groups called phyla. The five kingdom taxonomic classifications are Animalia, Plantae, Fungi, Protista, and Monera.

1.2. Phylum

A group of classes with similar distinctive characteristics is called phylum. The phylum is a biological classification, especially of plants which is second level of taxonomic rank.

1.3. Class

Class is a group or set with common characteristics, attributes, qualities or traits. It is further divided into one or more orders. It is below Phyla and above Order.

1.4. Orders

A taxonomic rank, order is used in classifying organisms, generally below the class, and comprised of families sharing a set of similar nature or character

.

1.5. Families

A taxonomic rank, family is placed in the classification of organisms between genus and order and it is a collection of things or entities grouped by their common attributes, It is a taxonomic group of one or more genera, especially sharing a common attribute.

1.6. Genus

In biology, a genus is a low level taxonomic rank used in the biological classification of living and fossil organisms. The term comes from the Latin genus meaning descent, family, type and gender. The standards for genus classification are not strictly codified, so different taxonomist often produce different classifications for genera. Generally the genus of plants should start by capital letter.

`

1.7. Species

Species are the lowest taxonomic rank, and the most basic unit or category of biological classification. An individual belonging to a group of organisms having common characteristics and are capable of matching with one another to produce fertile offspring.

A lot of research works have been carried out in plant classification. In Yan, et al. [2] a new supervised LLE (Locally Linear Embedding) method based on the fisher projection was proposed and combined it with a new classification algorithm based on manifold learning to realize the recognition of the plant leaves. The Fisher projection distance was replaced instead of samples geodesic distance, and a new supervised LLE algorithm was obtained.

In Ashit [3], a content based image retrieval system have been used for plant identification. The features were extracted from the common plant region that is segmented from environment using the max-flow min-cut technique. Image data have been compared using intensity and geometry. Color features which were used is relatively robust to background complication and independent of image size and orientation. The two division of shape features namely boundary based and region based was used as features.

An incorporated approach for an ontology based leaf classification system was projected in Stephen and Forrest [4], wherein machine learning techniques play a central role for the automation. A scaled CCD (charge coupled device) code system was proposed for categorizing the basic shape and margin type of leaf by using the taxonomy principle adopted by the botanists. Then a trained neural network was employed for recognizing the complete tooth patterns. The measurement on an unlobed leaf has been measured involuntarily according to the technique used in botany. Thresholding and neural network approach have been carried out to obtain more vein information of a leaf.

In Phoenix and Bastian [5] a hierarchical approach has been done to recognize live fish underwater videos. A tree has been constructed to organize all classes hierarchically in flat classifier. Flat classifier was developed as a multi-class classifier and that classifier used one-vs-one and one-vs-rest approaches to create multi class from binary class. The divide and conquer tactic strategy was used in hierarchical classification. Top to bottom approach has been used to display hierarchical information like Kingdom, Phylum, Class, Order, Family, Genus and Species of fish species.

In Valliammal1 and Geethalakshmi [6] the plant recognition system was developed using the preferential image segmentation (PIS) method. The PIS method has been used to segment an object of interest from original image. During the matching process, a probabilistic curve evolution method have been used with particle filters to measure the similarity between the shapes. Boundaries of objects have been detected using edge detection method. Euclidean distance was utilized to compare the feature vectors.

In this paper features like Margin, Veins, Color, Perimeter, Area, Aspect ratio, Rectangularity, Circularity, Sphericity, Lobes, Sub vein and Diameter have been used. Only Class - Magnolipsida, Order Brassicales and Rosales, Family - Caricacea, Brassicacea, Rosaceae, Rhamnacea and Species Carica papaya, Manihot esculenda, Coriandrum sativum,

`

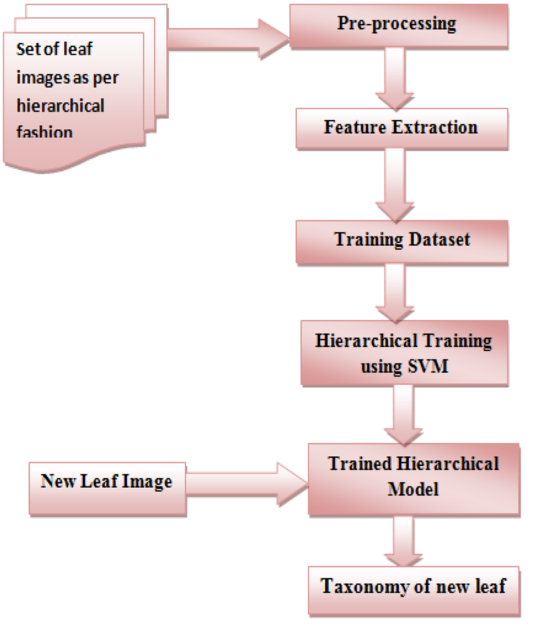
Brassica juncea, Rosa xathina, Rosa blanda, Rhamnus crocea, Rhamnus frangula of the plant taxonomy is taken into consideration for data modelling. The customized binary classifiers are trained with specific features at each node. Top to bottom approach is used for identifying taxonomic information like Class, Order, Family and Species name of plants. The last level of taxonomic ranks – species have been taken for Multi-class learning with species as class labels.

2. PLANT SPECIES PREDICTION MODEL

Plant species prediction is a procedure of identifying the Kingdom, Division, Class, Order, Family, Genus and specimen name of plant species. The plant species taxonomic classification is the theory and practice of grouping individuals into species, arranging species into larger groups and those groups names thus producing a classification.

The plant species prediction model building comprises of image acquisition, preprocessing, feature extraction, and hierarchical classification. The process flow of proposed leaf taxonomic classification system is shown in Fig.2.

Fig-2. Proposed Plant Taxonomic Model



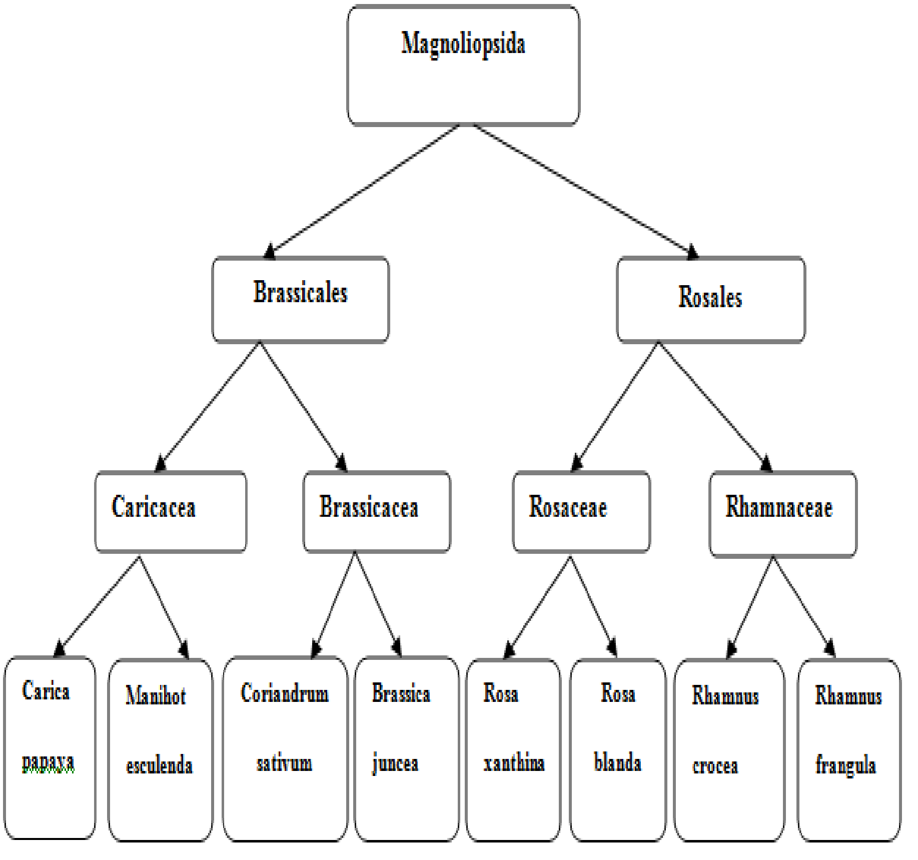
2.1. Data Acquisition

The images of various kinds of leave which are in JPEG format were captured using multi spectral CCD camera. Magnoliopsida class category consists of 400 images and that class is splitted up into two orders namely Brassicales and Rosales with 200 images respectively. The Caricacea and Brassicacea are the families that holds each 100 images and comes under the Brassicales order. The Rosaceae and Rhamnaceae are the families which come up to Rosales order and each contains 100 images.

`

The species such as Carica papaya, Manihot esculenda, Coriandrum sativum, Brassica juncea, Rosa xathina, Rosa blanda, Rhamnus crocea and Rhamnus frangula are taken into consideration and each species contains 50 images.

Fig-3. Plant Taxonomic Structure



2.2. Pre-Processing

The RGB image is converted into a grayscale image. The formula used to convert RGB value of a pixel into its grayscale value is

$gray = 0.2989 * R + 0.5870 * G + 0.1140 * B$

where R, G, B correspond to the red, green and blue color of the pixel respectively. Then the leaf image is transformed into threshold image. After thresholding binarized leaf image is obtained. The Fig.4 shows the original and binary image.

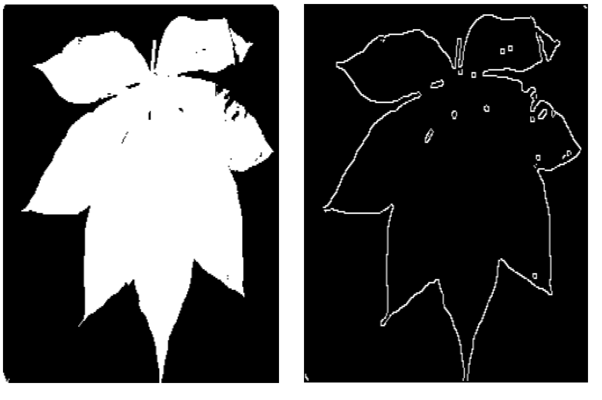
Fig.4. Original image and Binary image

`

2.2.1. Canny edge detection method

Edges in the binary image are detected using canny method. The canny edge detector is an edge detection operator that uses a multi-stage algorithm to detect a wide range of edges in images. To remove dubious objects, filtering is performed based on both the area of each object and based on the density of each object. The canny edge detector uses a gaussian filter to detect the edges of leaves. The Fig.5 shows the thresholding and edge detected image.

Fig.5. Thresholding and Edge detected image



2.3. Feature Extraction method

In pattern recognition and in image processing, feature extraction is a special form of dimensionality reduction. Transforming the input data into the set of features is called feature extraction. The descriptive features of leaf that are described below have been used in the proposed work.

2.3.1. Veins

Vein detection is performed on the gray scale image with flat, disk-shaped structuring element of radious are subtracted remained images by the margin based on that vein. The Gaussian filter methodology is used to identify veins on leaf. The following line of code is used to measure vein feature.

`

```
gauss = fspecial ( 'gaussian' , [ 5 5 ] ) ;
Vein = imfilter ( I, gauss ) ;
```

2.3.2. Diameter

The diameter is defined as the longest distance between any two points on the margin of the leaf. In geometry, the diameter of a circle is any straight line segment that passes through the center of the circle and whose endpoints lie on the circle. The following predefined function is used to compute this feature.

```
Diameter = regionprops (labeledImage,'EquivDiameter');
```

2.3.3. Aspect ratio

The aspect ratio is a ratio between the maximum length (width) and the minimum length (height) of the minimum bounding rectangle (MBR) and is computed using the following formula,

$$A_R = \frac{\text{dmin}}{\text{dmax}}$$

2.3.4. Rectangularity

The standard approach for measuring rectangularity is to use the ratio of the area of the region to the area of its minimum bounding rectangle. Rectangularity is the ratio of area and minimal bounding box Area. The rectangularity is measured using the following formula,

Rectangularity = L*W / A

where L is length, W is width and A is an area.

2.3.5. Margin

The leaf margin is the boundary area extending along the edge of the leaf. The margin is identified using Canny edge detection method. The following predefined code is used to compute this feature.

```
Margin = edge (im2bw(I, graythresh(I)),'canny');
```

2.3.6. Lobes

A leaf lobe is a partial rounded portion of a leaf margin, separated from the whole by a more or less deeply indentation that does not break the continuity of the structure and that may have toothed serrate. Area and solidity are used to measure the lobes area of a leaf. The following line of code is used to calculate the value of this feature.

```
Lobes = regionprops(im2bw(I, graythresh (I)) ,'Area', 'Solidity' );
```

2.3.7. Sphericity

Sphericity is a measure of the roundness of a shape. A sphere is the most compact solid, so the more compact an object is, the more closely it resembles a sphere. Sphericity is a ratio and therefore a dimensionless number. The following line of code is used to calculate this feature.

`

```
grainsBW = im2bw ( rgb2gray (grains), 126/255);
revim = imcomplement ( grainsBW );
grainsBW(L == 0) = 0;
L = mean (mean (L) )
```

2.3.8 Circularity

Circularity reflects the ratio of area to the perimeter and its computed using the following line of code,

```
circularity {k}= ceil (P^2 / (4 * pi * A));
```

where P is the perimeter and A is an area of leaf.

2.3.9. Perimeter

The perimeter of leaf is calculated by counting the number of pixels consisting of leaf margin, where the perimeter is P. The perimeter value is measured using the following code

```
Perimeter = regionprops ( im2bw ( I, graythresh (I) ) , 'Perimeter');
```

2.3.10. Area

The length and width of the leaf is measured and the leaf area is also measured using the following formula,

Area = L* W

where L is a length of leaf and W is a width of leaf. The distance between the two terminals of the main vein of the leaf is defined as the length. Drawing a line passing through the two terminals of the main vein, one can plot infinite lines orthogonal to that line. The longest distance between points of those intersection pairs is defined as the width.

2.3.11. Color

Features involved in color moments are mean, standard deviation skewness and kurtosis. The mean value of color is calculated using following line of code

```
Color = mean ( mean ( img ));
```

Finally a dataset with 400 instances have been created using the above mentioned features. Along with features the class name, order name, family name and name of the species are also added in order to facilitate hierarchical learning and training data set has been developed. The same dataset with species as class labels has been used for multi-class learning and training.

3. SUPPORT VECTOR MACHINE AND HIERARCHICAL CLASSIFICATION

Support vector machine is a training algorithm for learning classification and regression rules from data. The machine is presented with a set of training examples, (xi,yi) where the xi are the real world data instances and the yi are the labels indicating which class the instance belongs to. For the

`

two class pattern recognition problem, yi = +1 or yi = -1. A training example (xi,yi) is called positive if yi = +1 and negative otherwise.

SVMs construct a hyperplane that separates two classes and tries to achieve maximum separation between the classes. The simplest model of SVM called Maximal Margin classifier, constructs a linear separator given by w T x - γ = 0 between two classes of examples. The free parameters are a vector of weights w, which is orthogonal to the hyperplane and a threshold value γ. These parameters are obtained by solving the following optimization problem using Lagrangian duality.

Minimize = $\frac{1}{2}$ $||w||^2$

Subject to D_{ii} ($W^T X_i - \gamma) \geq 1$, i=1,….. l.

In this formulation the contribution to the objective function of margin maximization and training errors can be balanced through the use of regularization parameter c. The following decision rule is used to correctly predict the class of new instance with a minimum error.

$f(X) = \text{sgn}\,[W^T X - \gamma]$

The advantage of the dual formulation is that it permits an efficient learning of non–linear SVM separators, by introducing kernel functions. Technically, a kernel function calculates a dot product between two vectors that have been (non linearly) mapped into a high dimensional feature space. The parameters are obtained by solving the following non-linear SVM formulation (in Matrix form),

Minimize LD (u) $= \frac{1}{2} \quad u^T\ Q\,u \quad - e^T u$

$d^T u=0$, $0 \leq u \leq Ce$

Where Q = DKD and K - the Kernel Matrix. The kernel function K (AAT) (polynomial or Gaussian) is used to construct hyperplane in the feature space, which separates two classes linearly, by performing computations in the input space. The decision function is given by

$f(X) = \text{sgn}\,(\,K\,(\,x,\ x_i^T\,) * \ u - \gamma$ where, u - the Lagrangian multipliers.

When the number of class labels is more than two, the binary SVM can be extended to multi class SVM. One of the indirect methods for multiclass SVM is one versus rest method. For each class a binary SVM classifier is constructed, discriminating the data points of that class against the rest. Thus in case of N classes, N binary SVM classifiers are built. The comparison between the decision values produced by different SVMs is still valid because the training parameters and the dataset remain the same.

Hierarchical classification partitions all classes into multiple subsets and leaves similar classes for later stage. Flat classifier chooses a feature set based on the average accuracy over all classes whereas the hierarchical method applies a customized set of features to classify specific classes. So that hierarchical classification achieves good performance on similar classes. The hierarchical classification has the correlations between classes and finds out the related groupings [5]. The

`

algorithm for hierarchical classification developed in this research work as per Fig.3 is specified below:

Step 1: The dataset contains 400 leaf images with four class labels which are corresponding to various types of class, order, family and species.

Step2: Class – Magnoliopsida is taken as a root node and it is splitted up into two nodes order - Brassicales and Rosales with unique node number 1 and 2.

Step3: Order – Brassicales is partitioned into two families – Caricacea, Brassicacea, and those nodes are represented as node number 3 and 4.

Step 4: Order – Rosales is partitioned into two nodes (families – Rosaceae, Rhamnacea) with node number 5 and 6.

Step 5: Family – Caricacea is partitioned into two species - Carica papaya and Manihot esculenda with node number 7 and 8.

Step 6: Family – Brassicacea is divided into two species – Coriandrum sativum, Brassica juncea with node number 9 and 10.

Step 7: Family – Rosaceae is divided into two species Rosa xanthina, Rosa blanda with node number 11 and 12.

Step 8: Family – Rhamnacea is splitted up into two species – Rhamnus crocea, Rhamnus frangula with node number 13 and 14.

Step 9: Create binary models – Magnoliopsida, Brassicales, Rosales, Caricacea, Brassicacea, Rosacea and Rhamnacea. Finally all the seven models are integrated in a top-down fashion to develop hierarchical model.

Step 10: The common name with the taxonomic rank like Class, Order, Family and Species name of plant species are predicted for a given leaf image using hierarchical model.

Step 11: Exit.

4. EXPERIMENTS AND RESULTS

Two independent experiments have been carried out for plant species classification. In the first experiment, a flat multiclass classifier has been built using SVM^{light} for plant species classification and in the second experiment hierarchical classification model has been built using the same features

A. Multiclass Classification

A training dataset containing class labels corresponding to species - Carica papaya, Manihot esculenda, Coriandrum sativum, Brassica juncea, Rosa xanthina, Rosa blanda, Rhamnus crocea and Rhamnus frangula is used for constructing the classifier. The dataset is normalized using min-max normalization and the normalized dataset is trained using SVM^{light} tool with Linear, Polynomial and RBF kernels with different values for regularization parameter C. The parameters d and g are coupled with polynomial kernel and RBF kernel respectively. The accuracy and learning time of SVM classifiers are obtained for identifying the performance of classifiers. The accuracy and time

`

consuming for training and learning with Linear SVM, Polynomial SVM and RBF SVM are shown in Table 1, Table 2, and Table 3 respectively.

Table-1. Results of Linear SVM

c	Prediction Accuracy (%)	Time Taken (seconds)
10	72.60	0.00
20	66.43	0.00
30	71.39	0.00
40	71.37	0.00
50	72.84	0.00
60	77.54	0.00
70	73.84	0.00
80	76.54	0.00

Table-2. Results of Polynomial SVM

c	D	Prediction Accuracy (%)	Time Taken (Sec)
10	1	72.60	0.03
180	5	76.07	0.01
220	1	75.07	0.01
500	5	77.13	0.04
741	1	81.75	0.04
1750	5	81.01	0.03
2740	1	81.25	0.02
3100	5	81.25	0.01
6000	1	83.72	0.01

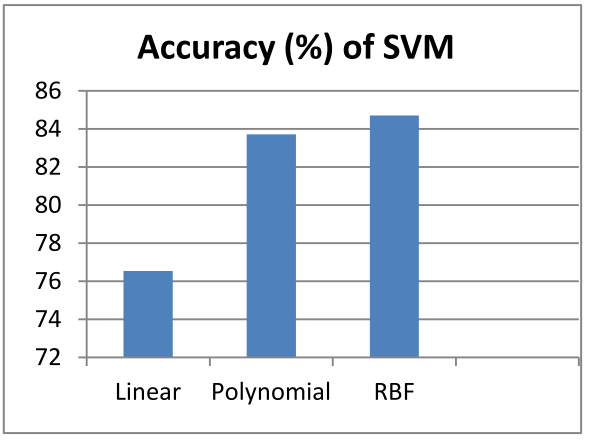
Table-3. Results of RBF SVM

c	G	Prediction Accuracy (%)	Time Taken (Sec)
60	0.5	73.5	0.01
159	1.5	73.81	0.01
300	0.5	72.51	0.02
1800	1.5	78.50	0.01
1810	0.5	78.81	0.01
2720	1.5	80.13	0.03
2730	0.5	80.46	0.00
2740	1.5	80.75	0.00
6000	0.5	84.72	0.00

The accuracy results of Linear, Polynomial, RBF kernels are compared and it proves that the RBF kernel has achieved better accuracy than other two kernels and the results are shown in Fig.6.

`

Fig-6. SVM Light Results



B. Hierarchical Classification

The hierarchical classification model for plant taxonomic classification has been developed through SVM learning in MATLAB. The training dataset consisting of 400 leaf images corresponding to class - Magnoliopsida, orders - Brassicales, Rosales, families - Cariaceae, Brassicacae, Rosaceae and Rhamnaceae has been employed to develop the hierarchical classification model.

Considering the hierarchical structures shown in Fig.3 since each node in the structure is binary, binary SVM classification model is built to construct hierarchical classification model. At the root level whole training dataset has been considered to create a binary classification model for Magnoliopsida (class name) with two class labels Brassicales and Rosales (order names) each consisting of 200 instances.

In the next level two binary SVM models for Brassicales and Rosales have been constructed by splitting the dataset into two datasets consisting of 200 instances. These two datasets have been trained with class labels Caricacea and Brassicacea (family names) each consisting of 100 instances for binary model corresponding to Brassicales and with class labels Rosacea and Rhamnacea (family names) each consisting of 100 instances for binary model corresponding to Rosales.

At the third level of hierarchy four binary SVM models have been generated for four nodes namely Caricacea, Brassicacea, Rosacea, Rhamnacea (family names). The training dataset is further partitioned into four datasets with 100 instances each corresponding to above four families. These training datasets have been trained with class labels Carica papaya, Manihot esculenda (species name) with 50 instances each for binary model corresponding to Caricacea, Coriandrum sativum, Brassica juncea (species name) with 50 instances each for binary model corresponding to Brassiacea, Rosa xanthina, Rosa blanda (species name) with 50 instances each for binary model corresponding to Rosacea and Rhamnus crocea, Rhamnus frangula (species name) with 50 instances each for binary model corresponding to Rhamnacea. Finally the binary SVM models generated at each level and at each node of the hierarchy are integrated in a hierarchical fashion to develop a hierarchical classification model.

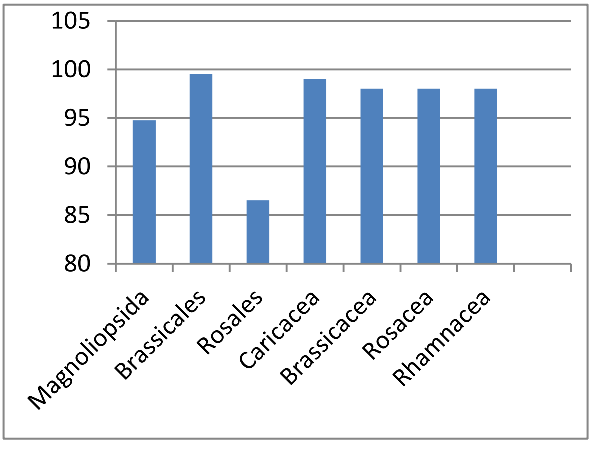
`

The performance of the trained models is evaluated using 10 – fold cross validation for its predictive accuracy. The models trained at each node of the hierarchy is evaluated individually and their results are analyzed. The cross validation results of the seven binary models namely class Magnoliopsida, orders Brassicales, Rosales, families Cariaceae, Brassicacea, Rosaceae and Rhamnaceae are shown in Table 4 and Fig.7.

Table-4. Ten Fold Cross Validation Results

Models	**Predictive Accuracy (%)**
Magnoliopsida	94.75
Brassicales	99.5
Rosales	86.5
Caricacea	99
Brassicacea	98
Rosacea	98
Rhamnacea	98

Fig-7. Ten Fold Cross Validation Results



5. PLANT SPECIES TAXONOMIC TOOL

An interactive tool for predicting the taxonomy of leaf has been developed using GUI in MATLAB. The Plant Taxonomic Tool begins by browsing a single leaf. The selected leaf image is displayed using grid. The margin, veins, color, perimeter, area, aspect ratio, rectangularity, circularity, sphericity, lobes, sub veins and diameter are the 12 kind of features which are used for plant taxonomy prediction. These features are extracted for the selected leaf and the values of features are displayed in list box. The record for that selected leaf is shown in command window and the corresponding feature vector given as input to hierarchical classification model.

Hierarchical classification of plant species was done through binary models, namely Class Magnoliopsida, Order Brassicales, Rosales, Families Caricacea, Brassicacea, Rosaceae and Rhamnaceae. The top down strategy is used for taxonomic classification. Finally the taxonomic

`

ranks of plant specimen namely class, order, family and species along with its common name are predicted and displayed. The screenshots of the GUI are shown in Fig.8 to Fig.10.

Fig-8. Browse a Leaf

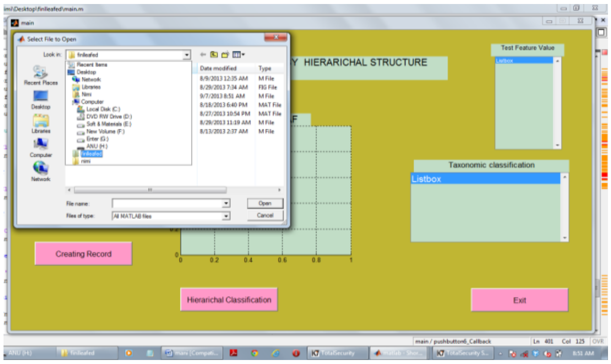


Fig-9. Feature Extraction

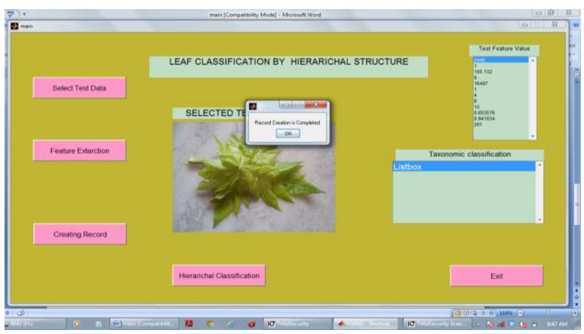
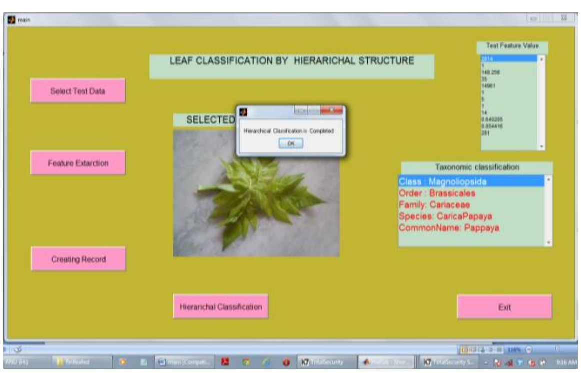


Fig-10. Hierarchical Classification



6. CONCLUSION

In this research work the multiclass classification has been done with last level of taxonomic ranks – species as class labels using SVM^{light}. The accuracy and learning time has been analyzed in Linear, Polynomial and RBF kernels with different values. The hierarchical classification also has

`

been done using seven binary models namely class - Magnoliopsida, orders - Brassicales, Rosales, families - Cariaceae, Brassicacae, Rosaceae, Rhamnaceae with different types of class, orders, families and species as class labels. The performance of hierarchical model using Support Vector Machine has also been evaluated. This paper proves that the performance of hierarchical classification is better than multiclass classification.

REFERENCES

[1] A. F. Budd, "Taxonomic classification of the reef coral family mussidae," *Zoological Journal of the Linnean Society,* vol. 166, pp. 465-529.

[2] Q. Yan, D. Liang, and J. Zhang, "Research of plant-leaves classification algorithm based on supervised LLE," *TELKOMNIKA,* vol. 11, pp. 3265 - 3270, 2013.

[3] G. Ashit, "Content-based image retrieval: Plant species identification," Master's Thesis, Oregon State University, 2002.

[4] G. W. Stephen and S. B. Forrest, "A leaf recognition algorithm for plant classification using probabilistic neural network, ARXIV:0707.4289v/ [cs.AI]," 2007.

[5] X. H. Phoenix and J. B. Bastian, "Hierarchical classification for live fish recognition."

[6] N. Valliammal1 and D. S. N. Geethalakshmi, "Automatic recognition system using preferential image segmentation for leaf and flower images," *An International Journal (CSEIJ),* vol. 1, 2011.

BIBLIOGRAPHY

[1] E. Franz, M. R. Gebhardt, and K. B. Unklesbay, "The use of local spectral properties of leaves as an aid for identifying weed seedlings," *Digital Images Transactions Of ASAE,* vol. 34, pp. 682-687, 1991.

[2] D. E. Guyer, G. E. Miles, D. L. Gaultney, and M. M. Schreiber, "Application of machine vision to shape analysis in leaf and plant identification," *Transactions Of ASAE,* vol. 36, pp. 163-171, 1993.

[3] K. Saranya and M. S. Vijaya, "Text dependent writer identification using support vector machine," *International Journal of Computer Applications,* vol. 65, pp. 6-11, 2013.

`