An Approach to Improve Classification Accuracy of Leaf Images using Dorsal and Ventral Features

Arun Kumar
Dept. of Computer Science & Engineering
Sir Padampat Singhania University
Udaipur, India

Deepak Khazanachi
School of Information Technology
University of Nebraska
Omaha, USA

Vinod Patidar
Department of Physics
Sir Padampat Singhania University
Udaipur, India

Poonam Saini
Dept. of Computer Science & Engineering
Sir Padampat Singhania University
Udaipur, India

***Abstract*—This paper proposes to improve the classification accuracy of the leaf images by extracting texture and statistical features by utilizing the presence of striking features on the dorsal and ventral sides of the leaves, which on other types of objects may not be that prominent. The texture features have been extracted from dorsal, ventral and a combination of dorsal-ventral sides of leaf images using Gray level co-occurrence matrix. In addition to this, this work also uses certain general statistical features for discriminating them into various classes. The feature selection work has been performed separately for the dorsal, ventral and combined data sets (for both texture and statistical features) using the most common feature selection algorithms. After selecting the relevant features, the classification has been done using the classification algorithms: K-Nearest Neighbor, J48, Naïve Bayes, Partial Least Square (PLS), Classification and Regression Tree (CART), Classification Tree(CT). The classification accuracy has been calculated and compared to find which side of the leaf image (dorsal or ventral) gives better results with which type of features(texture or statistical). This study reveals that the ventral leaf features can be another alternative in discriminating the leaf images into various classes.**

Keywords—Leaf image; Leaf classification; Texture features; Statistical features; Dorsal and ventral sides of leaves; Gray level co-occurrence matrix

I. INTRODUCTION

The plants play an integral role in the ecological balance by providing shelter, improving the atmosphere, providing medicinal values etc. Therefore, there is a dire necessity to preserve and conserve them. Plants have also been studied for increasing food production, bringing forth new varieties of fruits, flowers and plant species. Several attempts have been made to classify the plants on the basis of flowers, arrangement of leaves on the plants, shapes, color and texture, to name a few. Such studies are essential for the ecological balance as some of the plants are on the verge of the extinction. For a layman, the characteristics features of a digital image are texture, shape, color and size. But, for a computer system there must be a computer recognizable feature set which could be stored, refined and analysed for appropriate classification. The human quest for finding the image textural features dates back to 1970's when Haralick [1], Rosenfeld and Troy [2] have obtained textural coarseness of digital images by finding the difference of the gray values of the adjacent pixels and then performing autocorrelation of the image values. The texture based properties of digital images have also been used in medical images [3] and in tomography based images [4], analysis of ultrasound images [5] and classification of food items like Italian pasta and plum cakes [6,7].

Some common approaches for plant leaf classification using digital images are based on geometrical properties [8], texture and shape based features [9] and color features [10].

Nature has given two faces to the leaves: the dorsal (or the face up side) and the ventral (or the back side facing the substratum). The dorsal sides are generally smooth with texture, absorbing sunlight whereas the ventral sides have prominent vein structure. The fine line present on the leaf is called the mid rib or the prominent vein and other hair like lines are called secondary veins. The pattern of leaf venation is an important characteristic for the identification of a plant.

The texture is an integral property of every surface: patterns of tiles, wood, fabric or crops in the field. A texture contains important information regarding structural arrangement of surface and its relationships with the surroundings. The human eyes can interpret texture features for a surface which is fine, coarse, rough or smooth, rippled or irregular. In the case of digital images, the texture represents the arrangement of pixels and their distribution, which is very helpful in classifying the images into various categories. A digital image data structure is represented through pixels expressing the relative brightness values. All the pixels in a digital image form the population and in statistical jargon, a sample is a subset of values taken out from a digital image to draw appropriate conclusions about the characteristic properties exhibited by the population. A statistical sample drawn from a large population can be represented through frequency distribution or correlation curves can be drawn, through which detailed statistical analysis can be performed.

Therefore, the univariate image statistics like mean, mode, median or standard deviation etc. can be utilized in studying and discriminating one class of digital image from the other.

The present study proposes to improve the classification accuracy of the leaf images by extracting texture and statistical features by utilizing the presence of striking features on the dorsal and ventral sides of the leaves, which on other types of objects may not be that prominent. The texture features have been extracted from dorsal, ventral and a combination of dorsal-ventral sides of leaf images using Gray level co-occurrence matrix. In addition to this, certain general statistical properties [11] like Mean, Median, Integrated Density, Skewness, Kurtosis, Minimum value, Standard Deviation, Raw-Integrated Density, XM and YM of leaf images for discriminating them into various classes have been used. The most important task for achieving higher degree of classification accuracy is to extract relevant features which can improve the overall accuracy of the classifier. Hence, the selection of an appropriate set of features is very important in pattern recognition problems. The feature extraction algorithms help in reducing the storage space requirement for the data, the visualization of the small dataset improves, the features and their relation can be better understood, and further, the training phase is greatly reduced. This work performs the feature selection task for the dorsal, ventral and combined data sets (for both texture and statistical features) using the most common feature selection algorithms. After selecting the relevant features, the leaf images are classified using the classification algorithms: K-Nearest Neighbor, J48, Naïve Bayes, Partial Least Square (PLS), Classification and Regression Tree (CART), Classification Tree (CT) and then the classification accuracy for each algorithm with each feature data set is calculated. One of the objectives behind the study of dorsal and ventral sides of leaf images using texture and statistical features is to find which side (dorsal or ventral or dorsal-ventral) gives best classification results and with which type of classification approach: texture or statistical. The rest of the paper has been divided into four sections, the Section II highlights the proposed methodology (Creation of colored leaf Image data set, Preprocessing of the Digital images, Generation of texture features, Generation of statistical features, Feature Selection process in different data sets, Application of classification algorithms), the Section III describes the results obtained through the proposed methodology and their comparison with the similar recent work and in Section IV, the conclusion follows.

II. Proposed Methodology

This work proposes to use the dorsal and ventral sides of the leaf images using texture and statistical approaches for leaf discrimination into classes. The proposed approach involves the following steps:

A. Creation of colored leaf Image data set

The leaf image data set is available from several sources (Data Banks) including that of [12, 13, 14]. But the images that are stored in the data set are that of dorsal leaf images. But, this work proposes to utilize both the dorsal and the ventral faces of the leaf images. This necessitated the creation of a new data set with both the faces of the leaf images. For the purpose of creating the required database, the 24-bit RGB images of dorsal and ventral faces of leaves of the Helianthus annuus L.(Sunflower), Psidium guajava (Guava) and Alcia rosea (Hollyhock) have been captured as shown in Fig. 1 using Sony Cybershot HX200V with 18.2MP "Exmor R™," CMOS Sensor with extra high sensitivity technology, 30x optical zoom. The captured images include 100 dorsal side and 100 ventral side images for each of the above mentioned leaf categories totaling a sample size of 600 images with a pixel size of 1080 X 920.

Fig. 1. Colored sample of dorsal and ventral leaf images

B. Preprocessing of the Digital images

The leaf images were extracted using background removal technique. In order to find out the texture features and to reduce the computational complexity, all the colored images were converted to 8-bit gray level and reduced to the pixel size of 256X256. All the image processing tasks have been performed through ImageJ (Version 1.44) [11]. The gray stack of the slices of the dorsal, ventral and dorsal-ventral combined images have been prepared for further feature extraction using Gray Level Co-occurrence Matrix and statistical techniques [11] in a batch processing mode in ImageJ.

C. Generation of texture features

For batch processing, the gray stacks of the slices of the leaf images are processed through the texture extraction techniques given by Haralick [1] which provides the probability of gray level *i* occurring in the neighborhood of gray level *j* given distance *d* and angle Θ and total number of gray levels N (in the present case 256). The gray level co-occurrence matrix *GM* can be expressed mathematically as follows:

$$GM = \Pr(i, j \mid d, \theta, N) \quad (1)$$

In order to reduce the complexity, the inter pixel distance *d* is kept unity. Normally, in the case of Gray level co-occurrence matrix based methods, the calculations for feature extraction are carried out at unit pixel distance with Θ =0° and 45°, but this work has gone further in extracting the image texture features for the dorsal and ventral sides of the leaf images using Gray level co-occurrence matrix based method at unit pixel distance with angular pixel positions at Θ =0°, 45°, 90° and 135° independently and then combining them together using ImageJ software. The remaining angular positions of 225°, 270° and 315° are just the mirror images, therefore not considered.

Haralick [1], has described 14 texture properties out of which, the study uses the following 11 texture properties: Angular Second Moment(TF_1), Inverse Difference

Moment(TF_2), Contrast(TF_3), Energy(TF_4), Entropy(TF_5), Homogeneity(TF_6), Variance(TF_7), Shade(TF_8), Prominence(TF_9), Inertia(TF_{10}), Correlation(TF_{11}) [1,11,23] for preparing the texture feature values at different angular values of Θ for dorsal, ventral and dorsal-ventral combined leaf images.

The texture features dataset at the angular pixel position Θ is represented in the following manner:

$$TFD_\theta = \left(TF_1, TF_2, ..., TF_{11}\right)_\theta \quad (2)$$

Here $TF_1, TF_2, ..., TF_{11}$ indicate that all the 11 different values of texture features (mentioned above) measured at a particular value of Θ which is one of the values 0°, 45°, 90° and 135°.

The following three texture feature datasets have been prepared in this study: Image Texture Dorsal Dataset (ITDD), Image Texture Ventral Dataset (ITVD) and Combined Image Texture Dorsal-Ventral Dataset (CITDVD) using the equations (3), (4) and (5) respectively.

$$ITDD = \left\{TFD_{0^0}, TFD_{45^0}, TFD_{90^0}, TFD_{135^0}\right\}_{Dorsal} \quad (3)$$

$$ITVD = \left\{TFD_{0^0}, TFD_{45^0}, TFD_{90^0}, TFD_{135^0}\right\}_{Ventral} \quad (4)$$

$$CITDVD = \left\{TFD_{0^0}, TFD_{45^0}, TFD_{90^0}, TFD_{135^0}\right\}_{Dorsal\&Ventral} \quad (5)$$

D. Generation of statistical features

The 10 statistical features extracted from the leaf image dataset are Mean Gray value(SF_1), Median value(SF_2), Integrated Density(SF_3), Standard Deviation(SF_4), Minimum value(SF_5), XM(SF_6), YM(SF_7), Skewness(SF_8), Kurtosis(SF_9), Raw Integrated Density(SF_{10}).

The scale of calibration has been set to millimeter (mm). The statistical features datasets have also been prepared using ImageJ software [11].

The statistical features dataset is represented in the following manner:

$$SFD = \left\{SF_1, SF_2, ..., SF_{10}\right\} \quad (6)$$

Here $SF_1, SF_2, ..., SF_{10}$ indicate that all the 10 different values of statistical features (mentioned above).

The following three statistical feature datasets have been prepared: Image Statistical Dorsal Dataset (ISDD), Image Statistical Ventral Dataset (ISVD) and Combined Image Statistical Dorsal-Ventral Dataset (CISDVD) using the equations (7), (8) and (9) respectively:

$$ISDD = SFD_{Dorsal} \quad (7)$$

$$ISVD = SFD_{Ventral} \quad (8)$$

$$CISDVD = SFD_{Dorsal\&Ventral} \quad (9)$$

E. Feature Selection process in different data sets

Feature selection process involves selecting those features in the data set that are most useful and in simpler words most relevant and which shall provide better predictive accuracy and remove redundancy from the dataset. In addition to that, feature selection process also provides better understanding of the features.

In this study for the feature selection process, following seven feature selection algorithms have been used: Best First Search (BFS), Correlation Based Feature Selection (CFS), Chi-square (Chisq), OneR, Randomforest (RForest), ReliefF and Hill Climbing (HC). In addition to this one more method which includes all the features extracted from the image set i.e. No Feature Selection Algorithm (No Algo. Used) has also been used.

The algorithms mentioned above have been applied on the texture feature data sets: ITDD, ITVD, CITDVD and statistical feature data sets: ISDD, ISVD, CISDVD which generates a total of 48 different data sets comprising of 24 texture based and 24 statistical based data sets. The Tables I, II and III describe the number and names of texture features selected by each of the feature selection algorithm used in the present analysis. Similar results are given in Tables IV, V and VI for the statistical features.

TABLE I. FEATURE SELECTION ALGORITHM AND FEATURES SELECTED ON ITDD

S. No.	Feature Selection Algorithm	No. of Features Extracted	Name of the Features Extracted
1	BFS	4	TF_1, TF_5, TF_7, TF_8
2	CFS	4	TF_1, TF_3, TF_6, TF_7
3	Chisq	5	TF_7, TF_2, TF_9, TF_3, TF_{10}
4	OneR	5	TF_1, TF_4, TF_5, TF_6, TF_2
5	RForest	5	TF_9, TF_8, TF_7, TF_5, TF_{11}
6	ReliefF	2	TF_3, TF_{10}
7	HC	5	TF_1, TF_4, TF_5, TF_7, TF_8
8	No Algo. Used	11	TF_1, ..., TF_{11}

TABLE II. FEATURE SELECTION ALGORITHM AND FEATURES SELECTED ON ITVD

S. No.	Feature Selection Algorithm	No. of Features Extracted	Name of the Features Extracted
1	BFS	4	TF_1, TF_6, TF_7 ,TF_8
2	CFS	6	TF_2, TF_3,TF_5, TF_6, TF_7, TF_9
3	Chisq	3	TF_9, TF_5, TF_6
4	OneR	5	TF_7, TF_6, TF_2, TF_5 , TF_{11}
5	RForest	5	TF_9, TF_8, TF_7, TF_5 TF_{11}
6	ReliefF	2	TF_2,TF_1
7	HC	4	TF_2, TF_6, TF_9 , TF_{11}
8	No Algo. Used	11	TF_1, ..., TF_{11}

F. Application of classification algorithms

To discriminate the features obtained in section 2.5 into various classes (using 48 different data sets), the following six classification algorithms have been used: K-Nearest Neighbor (KNN), J48, Naïve Bayes, Partial Least Square (PLS), Classification and Regression Trees (CART), Classification Tree (CT) using "caret" package under RStudio [15]. Each data set was split into two groups (Training and Test sets) in the ratio 75:25. The training data set contains the class labels, whereas the testing dataset does not contain the class labels.

The preprocessing of the data involved centering and the scaling of the data matrix. In the classification procedure, a 10-fold cross validation technique has been applied which is repeated three times for validating any predictive model. Predictive accuracy and kappa values have been adopted as a measurable parameter for the classification process. Kappa is defined as the degree of right predictions of a model. This is originally a measure of agreement between two classifiers and is calculated as:

$$\kappa = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)} \quad (10)$$

In broad terms a kappa below 0.2 indicates poor agreement and a kappa above 0.8 indicates very good agreement or beyond chance [17, 18].

TABLE III. FEATURE SELECTION ALGORITHM AND FEATURES SELECTED ON CITDVD

S. No.	Feature Selection Algorithm	No. of Features Extracted	Name of the Features Extracted
1	BFS	8	TF_3, TF_4, TF_5, TF_6, TF_8, TF_9, TF_{10}, TF_{11}
2	CFS	4	TF_1, TF_3, TF_6, TF_7
3	Chisq	5	TF_7, TF_5, TF_8, TF_6, TF_7
4	OneR	5	TF_1, TF_4, TF_5, TF_6, TF_3
5	RForest	5	TF_9, TF_8, TF_7, TF_5, TF_{11}
6	ReliefF	2	TF_6, TF_2
7	HC	8	TF_1, TF_3, TF_5, TF_6, TF_7, TF_8, TF_9, TF_{11}
8	No Algo. Used	11	TF_1, ……, TF_{11}

TABLE IV. FEATURE SELECTION ALGORITHM AND FEATURES SELECTED ON ISDD

S. No.	Feature Selection Algorithm	No. of Features Extracted	Name of the Features Extracted
1	BFS	3	SF_5, SF_7, SF_{10}
2	CFS	4	SF_5, SF_6, SF_7, SF_1
3	Chisq	5	SF_7, SF_6, SF_1, SF_3, SF_{10}
4	OneR	5	SF_4, SF_1, SF_3, SF_{10}, SF_8
5	RForest	5	SF_7, SF_5, SF_6, SF_4,SF_{10}
6	ReliefF	2	SF_7, SF_5
7	HC	5	SF_4, SF_5, SF_6, SF_7, SF_3
8	No Algo. Used	10	SF_1,……., SF_{10}

TABLE V. FEATURE SELECTION ALGORITHM AND FEATURES SELECTED ON ISVD

S. No.	Feature Selection Algorithm	No. of Features Extracted	Name of the Features Extracted
1	BFS	4	SF_4 , SF_5 , SF_7 , SF_9
2	CFS	6	SF_1 , SF_4 , SF_5 , SF_6 , SF_7, SF_8
3	Chisq	3	SF_7 , SF_8 , SF_1
4	OneR	5	SF_9 , SF_8 , SF_1 , SF_3 , SF_{10}
5	RForest	5	SF_7 , SF_5 , SF_4 , SF_6 , SF_8
6	ReliefF	2	SF_7 , SF_6
7	HC	7	SF_1 , SF_4 , SF_6 , SF_7 ,SF_3 , SF_2 , SF_{10}
8	No Algo. Used	10	SF_1,……., SF_{10}

TABLE VI. FEATURE SELECTION ALGORITHM AND FEATURES SELECTED ON CISDVD

S. No.	Feature Selection Algorithm	No. of Features Extracted	Name of the Features Extracted
1	BFS	5	SF_4, SF_5, SF_6, SF_7, SF_9
2	CFS	4	SF_5, SF_6, SF_7, SF_8
3	Chisq	5	SF_7, SF_8, SF_6, SF_5, SF_4
4	OneR	5	SF_4, SF_8, SF_2, SF_7, SF_9
5	RForest	5	SF_7, SF_5, SF_6, SF_4, SF_9
6	ReliefF	2	SF_7, SF_5
7	HC	7	SF_1, SF_5, SF_6, SF_7, SF_3, SF_2, SF_9
8	No Algo. Used	10	SF_1,……., SF_{10}

TABLE VII. CLASSIFICATION ACCURACY FOR ITDD

Classification Algorithms	Feature Selection Algorithms								
	BFS	*CFS*	*Chisq*	*OneR*	*RForest*	*ReliefF*	*HC*	*No Algo. Used*	*Average Accuracy Across*
KNN	93.11	84.29	84.88	85.07	91.96	61.29	93.62	86.14	85.05
J48	96.00	90.92	91.00	87.25	94.00	59.96	95.18	93.55	88.48
Naïve Bayes	73.03	76.77	78.74	61.70	76.37	59.59	71.11	77.77	71.89
PLS	73.55	71.55	71.74	75.07	71.92	54.44	77.85	88.55	73.08
CART	89.85	85.48	86.25	81.03	87.51	61.81	90.37	88.11	83.80
CT	91.00	82.77	84.59	80.96	87.81	59.92	90.55	86.51	83.01
Average	86.09	81.96	82.87	78.51	84.93	59.50	86.45	86.77	80.89

TABLE VIII. CLASSIFICATION ACCURACY FOR ITVD

Classification Algorithms	Feature Selection Algorithms								
	BFS	*CFS*	*Chisq*	*OneR*	*RForest*	*ReliefF*	*HC*	*No Algo. Used*	*Average Accuracy Across*
KNN	93.81	85.22	90.51	83.92	95.37	76.74	87.88	85.66	87.39
J48	97.18	94.77	95.07	93.37	95.88	80.22	94.85	95.29	93.33
Naïve Bayes	74.74	78.03	69.67	73.81	76.00	62.66	74.48	79.88	73.66
PLS	71.14	84.14	63.14	81.55	64.62	58.22	69.00	89.88	72.71
CART	93.22	90.40	90.07	87.66	89.07	76.59	90.18	90.18	88.42
CT	86.37	87.18	85.88	80.92	90.85	76.44	86.81	89.03	85.44
Average	86.08	86.62	82.39	83.54	85.30	71.81	83.87	88.32	83.49

TABLE IX. CLASSIFICATION ACCURACY FOR CITDVD

Classification Algorithms	Feature Selection Algorithms								
	BFS	*CFS*	*Chisq*	*OneR*	*RForest*	*ReliefF*	*HC*	*No Algo. Used*	*Average Accuracy Across*
KNN	84.35	80.16	90.39	82.18	81.61	66.90	83.77	84.27	81.70
J48	94.37	90.85	93.87	88.42	94.42	66.77	94.25	94.85	89.73
Naïve Bayes	75.96	72.12	70.20	69.92	76.64	60.25	75.50	75.16	71.97
PLS	83.20	69.11	79.44	76.92	70.90	59.68	85.18	85.20	76.20
CART	89.37	83.83	86.81	82.27	87.62	66.74	86.96	89.70	84.16
CT	85.92	81.75	87.27	81.5	84.90	66.66	85.64	86.70	82.54
Average	85.53	79.64	84.66	80.20	82.68	64.50	85.22	85.98	81.05

III. Results and Discussion

The quantitative results, obtained by following the methodology proposed in Section II, for the predictive accuracy for texture feature data sets: ITDD, ITVD, CITDVD and statistical feature data sets: ISDD, ISVD, CISDVD are given in Tables VII, VIII, IX and Tables X, XI, XII respectively. However the pictorial representations of the kappa values for ITDD, ITVD, CITDVD texture feature data sets and ISDD, ISVD, CISDVD statistical feature data sets have been represented in Fig. 2((a),(b),(c)) and 3((a),(b),(c)) respectively.

TABLE X. Classification Accuracy for Isdd

Classification Algorithms	Feature Selection Algorithms								
	BFS	*CFS*	*Chisq*	*OneR*	*RForest*	*ReliefF*	*HC*	*No Algo. Used*	*Average Accuracy Across*
KNN	89.47	95.13	86.51	68.54	93.48	87.35	93.48	90.15	88.01
J48	89.88	92.16	85.91	72.99	91.86	86.31	91.86	92.04	87.88
Naïve Bayes	88.03	94.80	83.80	58.97	94.64	83.29	94.64	90.33	86.06
PLS	87.09	93.06	84.75	65.42	92.39	64.73	92.39	93.61	84.18
CART	92.58	90.20	83.57	53.91	89.31	90.45	89.31	89.06	84.80
CT	92.39	99.00	85.62	70.89	90.10	87.92	90.10	89.50	88.19
Average	89.91	94.06	85.03	65.12	91.96	83.34	91.96	90.78	86.52

TABLE XI. Classification Accuracy for Isvd

Classification Algorithms	Feature Selection Algorithms								
	BFS	*CFS*	*Chisq*	*OneR*	*RForest*	*ReliefF*	*HC*	*No Algo. Used*	*Average Accuracy Across*
KNN	85.54	87.72	84.53	70.79	88.87	80.58	79.19	85.13	82.79
J48	87.24	85.02	79.52	72.07	87.25	82.45	86.52	87.49	83.45
Naïve Bayes	83.06	84.93	79.68	53.39	87.96	78.84	78.60	82.86	78.67
PLS	73.69	79.77	75.00	53.17	80.09	61.77	76.31	84.95	73.09
CART	87.08	85.28	79.95	65.00	85.65	79.40	83.99	86.07	81.55
CT	83.28	85.98	75.91	58.67	84.85	81.28	84.17	85.83	80.00
Average	83.32	84.78	79.10	62.18	85.78	77.39	81.46	85.39	79.92

TABLE XII. Classification Accuracy for Cisdvd

Classification Algorithms	Feature Selection Algorithms								
	BFS	*CFS*	*Chisq*	*OneR*	*RForest*	*ReliefF*	*HC*	*No Algo. Used*	*Average Accuracy Across*
KNN	91.40	89.11	91.77	81.48	91.40	77.55	89.037	88.22	87.50
J48	89.40	86.37	89.48	80.44	89.40	76.44	86.518	89.11	85.89
Naïve Bayes	89.62	84.59	88.29	79.25	89.62	77.92	85.55	84.88	84.97
PLS	87.03	83.18	85.7	71.55	87.03	63.55	86.81	87.40	81.53
CART	85.48	86.29	86.81	78.00	85.48	63.03	83.037	85.77	81.74
CT	84.37	83.40	86.00	78.29	84.37	78.14	82.29	84.00	82.61
Average	87.88	85.49	88.01	78.17	87.88	72.77	85.54	86.56	84.04

A. Analysis on the basis of Texture feature data sets

It has been observed from the values for predictive accuracy for the texture feature dataset, the ITVD feature data model has the highest value for the average predictive accuracy (83.49%) amongst all the texture based data models (CITDVD(81.05%),ITDD(80.89%)) studied in this work. The comparison of the present results with the results of [9] are not directly comparable due to the differences in the datasets used. Despite of this fact this work compares its results with the results of [9]. In [9] two classification algorithms Neuro Fuzzy Controller(NFC) and Multi-Layer Perceptron(MLP) have been used for the texture based model with only dorsal side images and the average predictive accuracy achieved is 81.6% and 87% respectively. In the proposed ITVD model, ventral based texture feature model provides average predictive accuracy value of 83.49% and J48 classification algorithm gives 97.18% accuracy value using Best First Search(BFS) algorithm for feature selection, which is comparable with the results of [9] as shown in the Fig. 4.

While observing the accuracy values for texture feature based models, ITDD model provides accuracy value of 96% using Best First Search(BFS) algorithm for feature selection applied to dorsal leaf images and J48 as the classification algorithm. When CITDVD model is used, an accuracy value of 94.85% for J48 algorithm has been observed when all the textures features are used (No feature selection algo. used) as shown in the Tables VII, VIII and IX. On comparing results with textures segmentation model [19], which used Brodatz album (each image size 256 X 256) prepared the gray level co-occurrence matrix at unit pixel distance with angular pixel positions at Θ =0°, 45, has achieved predictive accuracy as high as 90% (approx.). However this work has prepared the gray level co-occurrence matrix at unit pixel distance with angular pixel positions at Θ =0°, 45°, 90° and 135° and have achieved better predictive accuracy in all the texture based (ITDD, ITVD, CITDVD) models as shown in Table XIII.

B. Analysis on the basis of Statistical feature data Sets

On observing the values for predictive accuracy for the statistical feature model, the ISDD feature model has the highest value for the average predictive accuracy (86.52%) amongst all the statistical based feature data models (CISDVD(84.04%), ISVD(79.92%)) studied in this work, as shown in the Fig. 4.

On comparing the results for statistical based feature models proposed in this work with [9], which is based on the dorsal based image sets only, two of the proposed statistical based feature models (ISDD and CISDVD) have fared better by giving more values for average predictive accuracy. Now, on comparing the texture based (ITDD, ITVD, CITDVD) and statistical based feature models (ISDD, ISVD, CISDVD) proposed in this work, the statistical based model ISDD fares the best amongst all the models proposed in this work in achieving the average predictive accuracy, as shown in the Fig. 5. While observing the classification using statistical features, by using K-Nearest Neighbor algorithm with correlation based feature selection algorithm has given the highest accuracy value of 95.13%. The ISVD model has achieved highest accuracy value of 88.87% with K-Nearest Neighbor algorithm with Random Forest based feature selection algorithm. On combining the dorsal and ventral images together and it has been observed that the predictive accuracy achieved is 91.77% with K-Nearest Neighbor algorithm with chi-square as the feature selection algorithm as shown in the Tables X, XI and XII.

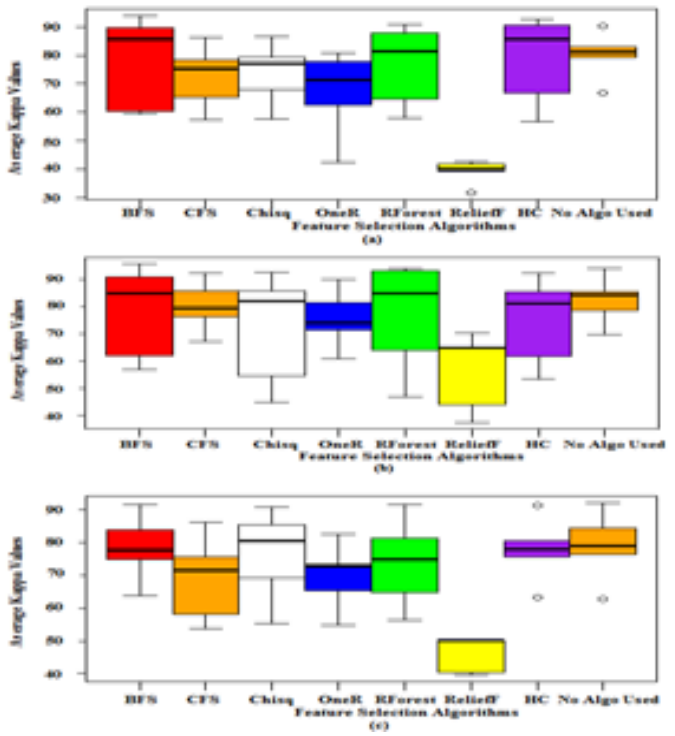

Fig. 2. Average kappa values versus feature selection algorithms for (a)ITDD (b) ITVD (c) CITDVD respectively

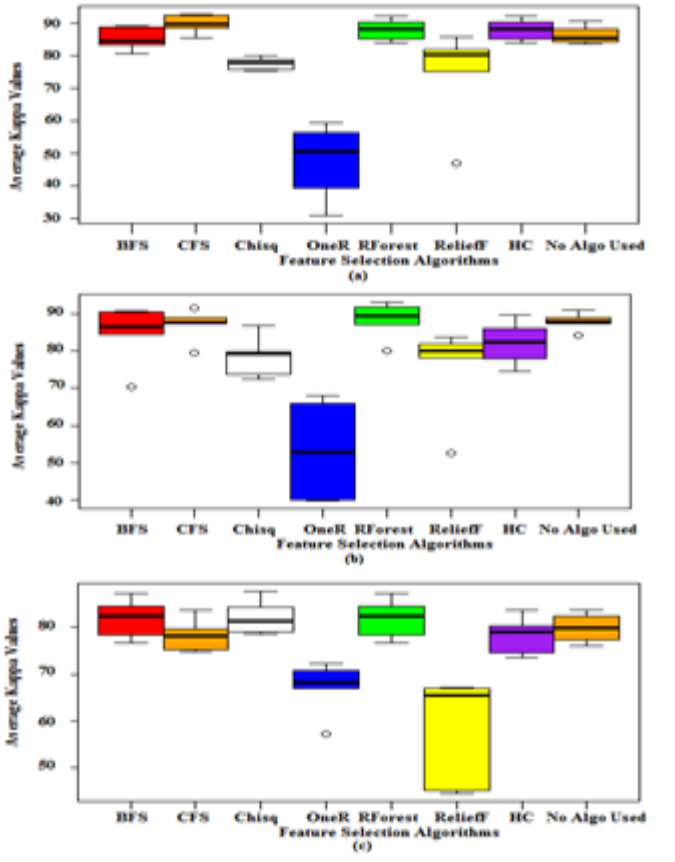

Fig. 3. Average kappa values versus feature selection algorithms for (a) ISDD (b) ISVD (c) CISDVD respectively

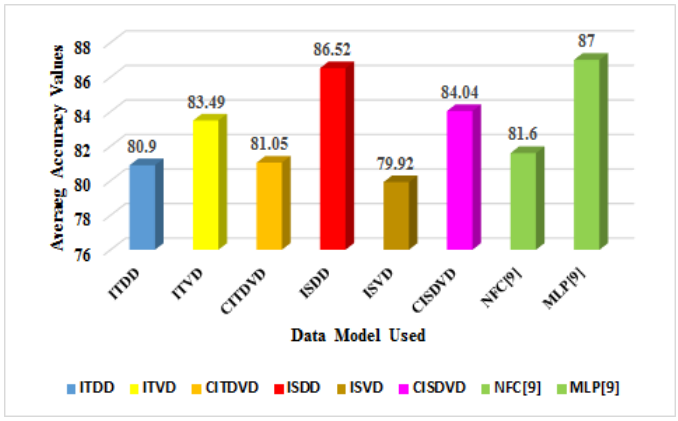

Fig. 4. Comparison of average accuracy of different data models with [9]

TABLE XIII. SUMMARY CHART FOR THE COMPLETE WORK CARRIED OUT

Model Type	Model Name	Average Predictive Accuracy (%)	Best Classification Algorithm on the basis of Predictive Accuracy Values	Feature Selection Algorithm used	Number of Features Used	Predictive Accuracy Values for Best Classification Algorithm (%)
Dorsal Leaf Image Models	ITDD	80.89	J48	BFS	4	96
	ISDD	86.52	KNN	CFS	4	95.13
Ventral Leaf Image Model	ITVD	83.49	J48	BFS	4	97.18
	ISVD	79.92	KNN	RForest	5	88.87
Combined Dorsal& Ventral Leaf Image Model	CITDVD	81.05	J48	No Algo. Used	11	94.85
	CISDVD	84.04	KNN	CFS	4	91.77

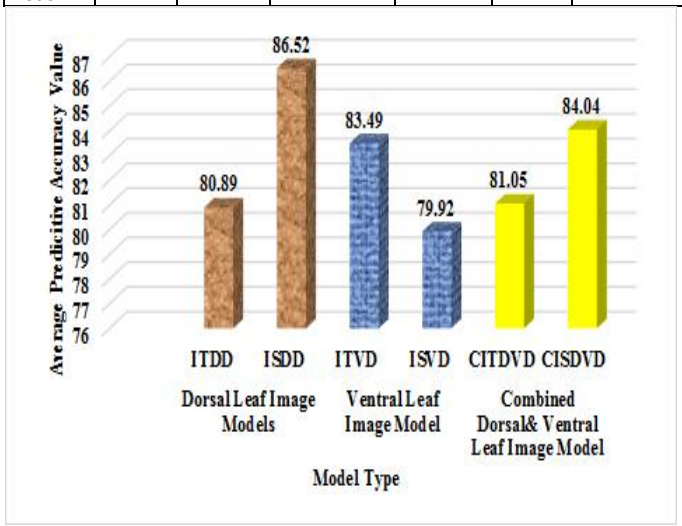

Fig. 5. Comparison of average predictive accuracy of different data models

C. Analysis on the basis of number of texture features used and the Average Misclassification results

In ITDD based models, when features are selected using Hill Climb algorithm (5 features selected) the average

misclassification rate is the 13.55% and when no feature selection algorithm is used (all the 11 features used), the average misclassification rate is the 13.23% as shown in the Fig. 6(a). In ITVD based model, when the features are selected using Correlation based feature selection algorithm (6 features selected), the average misclassification rate is the 13.88%, and when no feature selection algorithm is used (all the 11 features used), the average misclassification rate is the 11.68% as shown in the Fig. 6(b). In CITDVD based model, when the features are selected using Best First Search algorithm (8 features selected) the average misclassification rate is the 14.47%, and when no feature selection algorithm is used (all the 11 features used), the average misclassification rate is the 14.02% as shown in the Fig. 6(c).

D. *Analysis on the basis of number of statistical features used and the Average Misclassification results*

In ISDD based models, when features are selected using Correlation based algorithm (4 features selected) , the average misclassification rate is the 5.94% and when Hill Climb and Random Forest based algorithms are used with 5 features, the average misclassification rate is the 8.04% as shown in the Fig. 7(a). In ISVD based model, when the features are selected using Random Forest algorithm (5 features selected), the average misclassification rate is the 14.22%, as shown in the Fig. 7(b). In CISDVD based model, when the features are selected using Chi-square (5 features selected) the average misclassification rate is the 11.59%, as shown in the Fig. 7(c).

E. *Analysis on the basis of Number of features selected for classification*

On comparing the results of this work with the [19, 20] as shown in Fig. 8, [19] has the highest predictive accuracy of 90% with 32 features and the highest predictive accuracy achieved is 93.29% for 10 features on Lung Cancer Data [20], whereas in the present study, when 10 features are used, the accuracy achieved is 95.13% using Correlation based feature selection (CFS) for ISDD based model. In the case of ISDD model proposed in this study has achieved the highest accuracy values for 10 features.

The feature selection and misclassification method is not directly comparable due to different datasets used, but with the 10 features selection as the criteria for classification, this work has compared its results with [19,20].

F. *Analysis on the basis of dorsal and ventral features*

The summary of the results, presented quantitatively in Table XIII and graphically in Fig. 5, clearly demonstrate the supremacy of ventral features over the dorsal features. The highest predictive accuracy (97.18%) is achievable through classification algorithm J48 using Best First Search algorithm applied over texture features obtained from ventral side leaf images. The statistical features are giving the best average predictive accuracy (86.52%) amongst all the models proposed in this work.

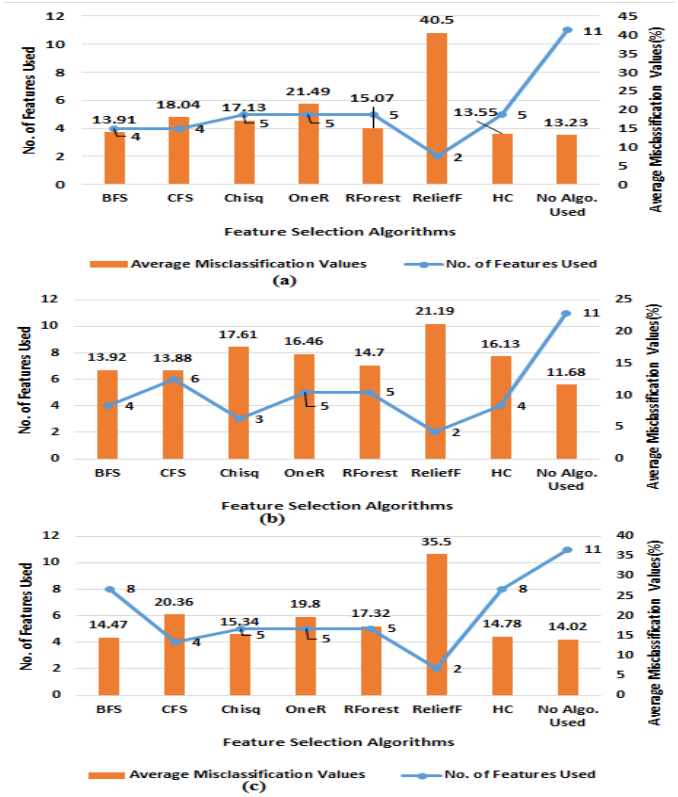

Fig. 6. Average misclassification rate vs. no. of features selected using (a) ITDD (b) ITVD (c) CITDVD respectively

IV. CONCLUSIONS

This paper proposes to utilize the concept of striking features present on both the dorsal and the ventral sides of the leaves and has been modeled around texture and statistical features for dorsal, ventral and dorsal-ventral leaf images. It has been observed that the texture based model, the ITVD model, is giving better average predictive accuracy as compared to other texture based models. This strengthens the proposition of this work that ventral sides of leaf images can be another alternative for extracting and discriminating features. Based on the results of all the statistical feature based models, it is inferred that the ISDD model is the best amongst all the texture and statistical based models used in this work.

The statistical feature set is providing much better predictive accuracy as compared to the models with texture based feature sets owing to the fact that the mutual information (MI) which is based on entropy, provided by the two or more random variables in the dataset is more in the case of statistical feature sets as compared to the texture based feature sets. Based on the extensive analysis, performed in this work, it is proposed that the statistical model (ISDD) which is purely based on calculating the statistical feature values can be applied for studying any object of interest with dorsal side of the image.

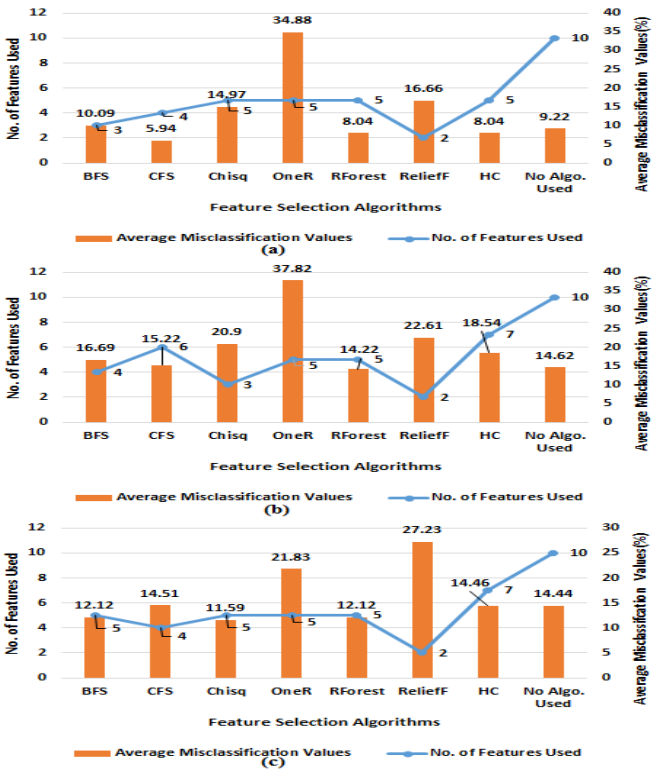

Fig. 7. Average misclassification rate versus no. of features selected using (a) ISDD (b) ISVD (c) CISDVD respectively

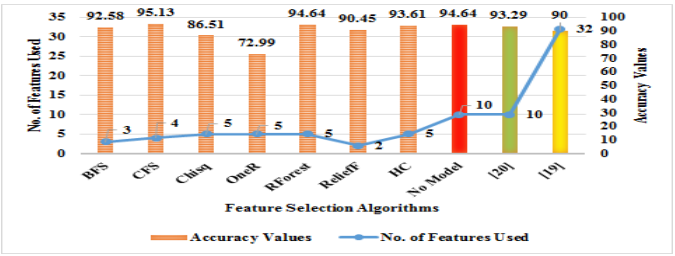

Fig. 8. Comparison of accuracy values versus no. of features used in ISDD with [19, 20]

REFERENCES

[1] Robert M Haralick et al., "Texture features for image classiifcation", IEEE Transactions on Systems, Man and Cybernetics, Vol. 3, No. 6, pp. 610-621, 1973.

[2] Rosenfeld, E. Troy, "Visual texture analysis", Computer Sci. Cent., University of MaryLand, College Park, Technical Report, pp. 70-116, 1970.

[3] G. Castellano et al., "Texture analysis of medical images", Clinical Radiology, Vol. 59, pp. 1061–1069, 2004.

[4] M.S. Oliveira et al., "Texture analysis of computed tomography images of acute ischemic stroke patients", Braz. J. Med. Biol. Research, Vol. 42, No. 11, pp. 1076-1079, 2009.

[5] Igor Pantic et al., " Nuclear entropy, angular second moment, variance and texture correlation of thymus cortical and medullar lymphocytes: Gray level co-occurrence matrix analysis", Annals of the Brazilian Academy of Sciences, Vol. 85, No. 3, pp. 1063-1072, 2012.

[6] Fongaro L. et al., "Assesment of surface aspects of foods using ImageJ plugins", In proceedings of the ImageJ User and Developer Conference, Luxembourg: Centre de Recherche Public Henri Tudor, ISBN: 2-919941-18-6, pp. 245-248, 2012.

[7] Lorenzo Fongaro, Knut Kvaal, "Surface texture characterization of an Italian pasta by means of univariate and multivariate feature extraction from their texture images", Food Research International, Vol. 51, No. 2, pp. 693-705, 2013.

[8] Cem Kalyoncu, Onsen Toygar, "Geometric leaf classification", Computer Vision and Image Understanding, Vol. 133, pp. 102-109, 2015.

[9] Jyotismita chaki et al., "Plant leaf recognition using texture and shape features with neural classifier", Pattern Recognition Letter, Vol. 58, pp. 61-68, 2015.

[10] A.J. Perez et al., "Color and shape analysis techniques for weed detection in cereal fields", Comput. Electron. Agric. Vol. 25, pp. 197-212, 2000.

[11] Rasband, W.S., ImageJ, 1997-2014, U. S. National Institutes of Health, Bethesda, Maryland, USA.

[12] http://imagej.nih.gov/ij/http://www.vision.caltech.edu/Image_Datasets /leaves/leaves.tar (Web page last visited on Aug., 2015)

[13] Database from Oxford, http://www.plant-phenotyping.org/ CVPPP2014-dataset, (Web page last visited on Aug., 2015)

[14] https://archive.ics.uci.edu/ml/machine-learning-databases/00288/leaf.zip ,(Web page last visited on Aug., 2015)

[15] R Development Core Team, R:A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0., 2008.

[16] Piotr Romanski, FSelector: Selecting attributes. R package version 0.19, 2013, http://CRAN.R-project.org/package=FSelector

[17] Anthony J. Viera, Joanne M. Garrett, "Understanding interobserver agreement:The Kappa statistic", Family Medicine, Vol. 37, No. 5, pp. 360-363, 2015.

[18] Julius Sim , Chris C Wright, "The Kappa statistic in reliability studies: use, interpretation, and sample size requirements", Physical Therapy, Vol. 85, No. 3, pp. 257-268, 2005.

[19] Andrik Rampun et al., "Texture segmentation using different orientations of GLCM features", Mirage 6th International Conference on Computer Vision / Computer Graphics Collaboration Techniques and Applications, 2013.

[20] Matthew Shardlow, "An analysis of feature selection techniques", https://studentnet.cs.manchester.ac.uk/pgt/COMP61011/goodProjects/Shardlow.pdf (Web Pages last visited on May 5, 2015).