International Journal of Computing and Digital Systems
ISSN (2210-142X)
Int. J. Com. Dig. Sys. 11, No.1 (January-2022)

https://dx.doi.org/10.12785/ijcds/110111

Geometry Feature Extraction of *Shorea* Leaf Venation Based on Digital Image and Classification U sing R andom Forest

Ishak Ariawan[1,2], Yeni Herdiyeni[2] and Iskandar Zulkarnaen Siregar[3]

[1]*Marine Information System, Universitas Pendidikan Indonesia, Serang, Indonesia*
[2]*Department of Computer Science, IPB University, Bogor, Indonesia*
[3]*Department of Silviculture, IPB University, Bogor, Indonesia*

Received 24 February 2021, Revised 9 July 2021, Accepted 23 July 2021, Published 9 January 2021

Abstract: The characterization of *Shorea spp.* tree species among other forest trees appears relatively complicated. Therefore, certain errors tend to occur during planting stock material collection, particularly at seedling or juvenile stages. This mis-identification could probably be minimized by initial sound identification, although it requires very extensive efforts. As a consequence, precise and rapid identification system is required to differentiate the sample at the seedling phase. The identification process involves usually the use of leaves, in which venation forms a major leaf feature with unique architecture and consistent pattern to segregate *Shorea* species. However, geometric properties also exist and can be extracted, using a geometric mathematical model. The approach determine the position of venation point by applying the linear coordinate values. This study was aimed at identifying *Shorea* species, using using random forest classification techniques. In addition, information on leaf venation's geometric features include the attribute angle, length, distance, scale projection, angle difference, straightness, length ratio, as well as densities of leaf vein, branching and ending points, were necessary. In particular properties, the mean, variance and standard deviation are evaluated. Subsequently, to obtain the most important traits, feature selection was conducted, using Boruta algorithm. The results showed the success of the applied model in classifying *Shorea* species, by leaf venation feature. Also, optimum accuracy was attained at 91.90%, with cut-off training and testing data of 90:10, by analyzing 1000 single trees. Furthermore, extensive sensitivity and precision values were obtained at 89.95 and 90.66%, respectively. These results clearly indicated a superior performance model.

Keywords: Feature Extraction, Random Forest, *Shorea*, Venation Feature Geometry

1. Introduction

Shorea spp. are described as a group of timber-producing tree species in Indonesia with important value of commodity [1]. The genus *Shorea* with 194 species, occurs in the tropics [2], as well as possesses significant economic value, due to the good timber quality. In addition, the timber is commonly processed as a light-to-heavy construction material. *Shorea* also produce non-timber forest products, including resin, tengkawang, nut fruit and tannins [3].

Massive forest exploitation threatens plant sustainability [3]. A total of 156 out of the 194 species were incorporated in red list, while 59.6% and 25% occurred in critically endangered and e endangered categories, respectively [4].

Ex-situ conservation appears to be a promising strategy against species extinction. This concept involves transfer of plants from their natural habitats [5], for example through natural regeneration such as wildlings [6].

During the collection of planting stock materials from natural regeneration, distinguishing correct species is often problematic due to complex morphological traits. Therefore, an error occurrence during collection is also possible and an initial identification method with accuracy and rapid process is required.

In the field, species identification is usually conducted using leaves [7], [8], [9]. These plants morphological traits are easy to locate in the forest floors that exhibit several distinct characteristics [10]. In particular, leaf venation offers a significant feature to differentiate the *Shorea* species, alongside a unique architecture and consistent pattern [11], [12]. Various geometric properties are associated with leaf venation and can be extracted using a linear mathematical model. This approach determines the of the venation point, by applying geometric coordinate values [13], [14], [15].

Several studies have classified plants based on the leaf

E-mail address: : ishak$_a$riawan@upi.edu, yeni.herdiyeni@apps.ipb.ac.id, siregar@apps.ipb.ac.id

venation feature. Properties applied to differentiate fruit trees are growth, include angle (Ang), length (Len) and distance (Dis) [16] while to classify medicinal plants are densities of leaf vein, branching and ending points were employed [15]. However, to distinguish Shorea species, four attributes including, scale projection (Spr), angle difference (Adi), straightness (Str), and length ratio (Lra), were involved [9].

This research generally applied the feature angle, length, distance, scale projection, angle difference, straightness, and length ratio, as well as densities of leaf vein, branching and ending points, in order to identify *Shorea* species. In certain instances, the mean, variance, and standard deviation are calculated. Subsequently, to obtain the major leaf venation properties, feature selection is greatly preferred, using Boruta algorithm. This approach was initially employed to improve the procedure for selecting the most important random forest features (Breiman's importance) [17]. Furthermore, the technique was successfully applied in high-throughput DNA methylation sequencing dataset [18]. The results showed the capacity of Boruta algorithm to enhance classification performance by 6.77%. Another utilization involved the analysis of OMICs dataset (leukemia, lung cancer, poriasis, and peripheral blood mononuclear cells), with accuracies of 98, 99, 99, and 98%, respectively [19]. The findings from feature selection were subsequently adopted to classify individual *Shorea* species.

Random Forest (RF) is a widely applied classification technique. This approach was successfully used to categorize medicinal plants, with 99% accuracy [20], while for grape disease, 86% was achieved [21]. Also, the prediction of activity level of stroke sufferers attained 88.24% accuracy [22], and 99.97% to classify DDoS attacks [23].

2. Method

A. Data Collecting

In this study, *Shorea* leaves were acquired in several nurseries, including the IPB Faculty of Forestry and Environment, Forestry Research and Development Center, as well as Bogor CIFOR Forest. The herbarium process was initiated prior to sample compression, followed by scanning. A total of 212 *Shorea* leaf images was scanned and classified, depending on the species (*S. leprosula, S. selanica, S. ovalis*, and *S. acuminata*). Moreover, individual variety comprises 53 pictures. Figure 1 a sample of leaf images from four *Shorea* species.

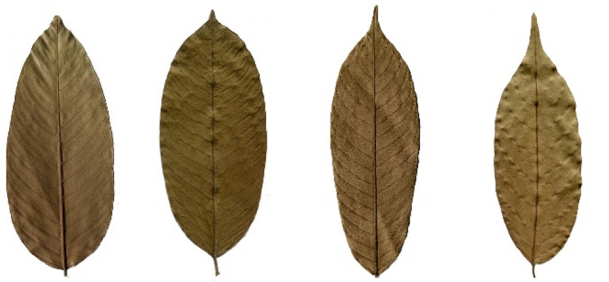

Figure 1. Leaf image samples of four *Shorea* species

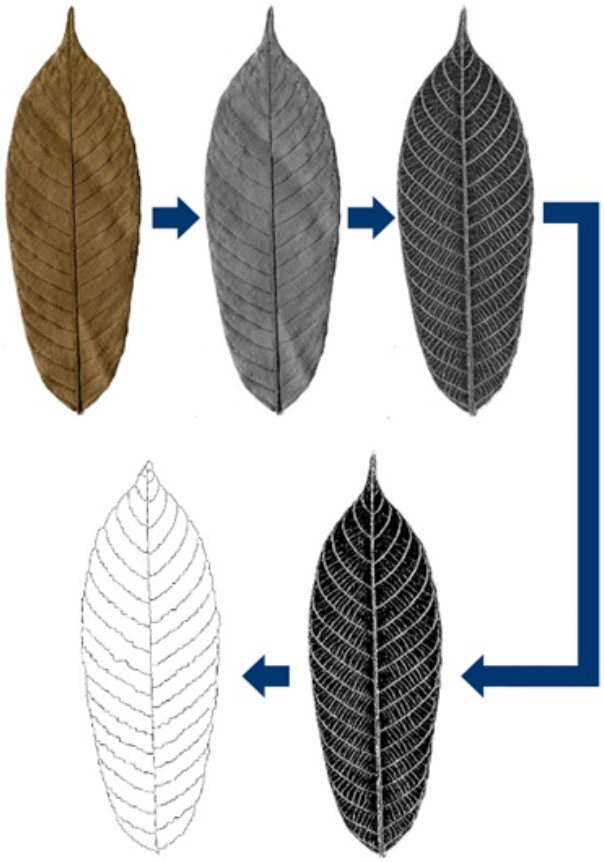

Figure 2. Segmentation process of *Shorea* leaf image data

B. Segmentation

Segmentation is very significant in image processing. The approach separates the image into several homogeneous components, with possible extracting into objects. This procedure was followed by a careful observation to determine the interest region [24].

Figure 2 represents the segmentation of *Shorea* leaf image data. The process was conducted four stages. First, a single surgical technique was used to detect primary, secondary, and tertiary leaves venation. Sobel operation provided a similar technique for edge detection by examining vertical and horizontal edges of the image [25]. Second phase is description as thresholding process that was known to sharpen the the leaf venation colour or alter the image into a binary appearance [26]. The range of threshold values occurred between 1-255, but tends to vary, based on previous segmentations. However, the third phase employed morphological techniques, including dilation and erosion, to eliminate noise and connect broken venations [27], [28]. In addition, the values of both events varied significantly,

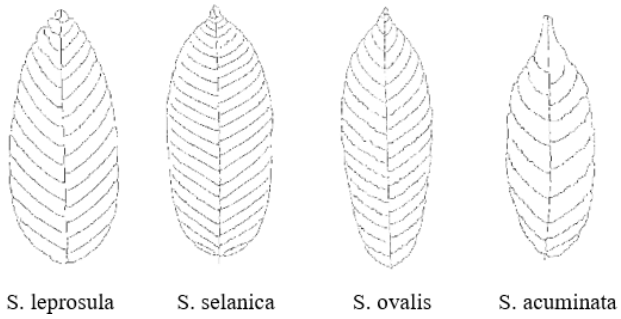

Figure 3. Samples of leaf image segmentation for four *Shorea* species

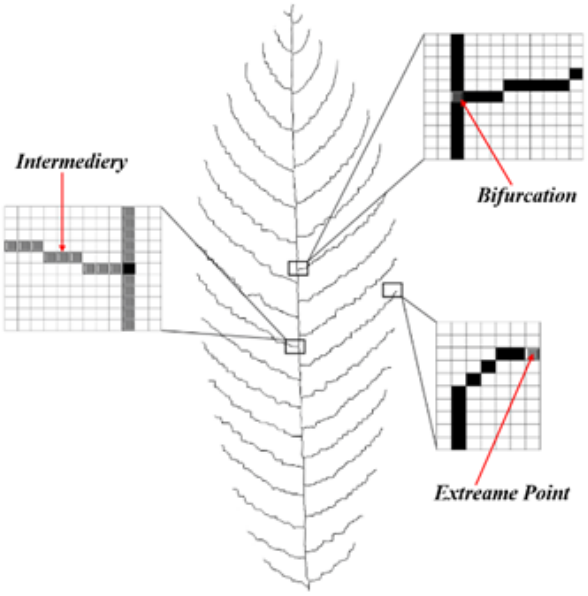

Figure 4. Bifurcation, intermediary and extreme points on the venation of *Shorea ovalis* leaves

depending on thresholding results. Dilation increase object segments' size by extending surrounding layers. Simultaneously, instigated the size reduction by decreasing surrounding layers [24]. The final stage involved the thinning process, used to trim the line [29], or modify the white image (venation) thickness to 1 pixel. Figure 3 shows the leaf image segmentation of four *Shorea* species.

C. Vein Detection

in this stage, the image data segmentation results were determined by bifurcation, intermediary, and extreme point coordinates. Bifurcation is the initial branching and links between primary, secondary, and tertiary venations. In addition, the last pixel is referred as the extreme point. Figure 4 shows the intermediary in the in the form of a link between bifurcation and extreme point.

The basic concept in determining bifurcation and extreme points relate to a pixel with more than two neighbors.

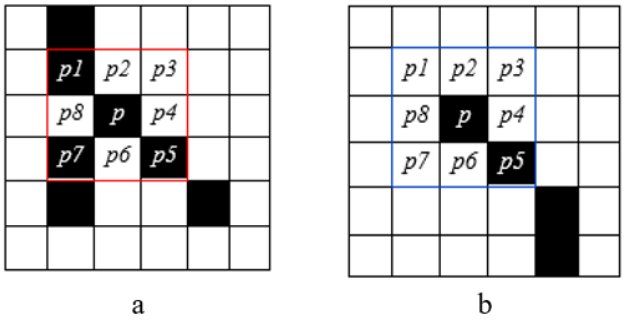

Figure 5. Determination of bifurcation and extreme point pixels in structure: (a) pixel p is a bifurcation, neighbor $p > 2$ pixels (b) pixel p is an extreme point, neighbor $p < 2$.

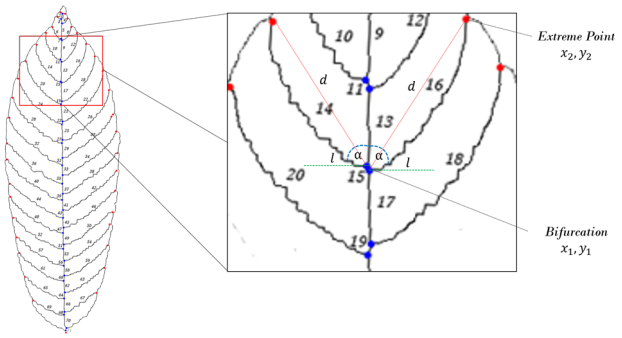

Figure 6. Representation of feature extraction results

Therefore, pixel p is the bifurcation observed in Figure 5a (pixel p has adjacent *p1*, *p5* and *p7*). A pixel p value less than two represents the extreme point as illustrated in Figure 5b (pixel p has a neighbor pixel *p5*). Pixel points apart from bifurcation and extreme points are called intermediary.

D. Feature Extraction

The feature extraction was conducted after detecting the venation. This process calculates the feature angle, distance, length, straightness, angle difference, length ratio, scale projection, Leaf vein density, area, total skeleton length, total branching point and ending point, the densities of branching and ending points. Figure 6 represents the venation feature extraction result, with bifurcation and extreme points. These coordinates are connected to form a segment and also extracted by calculating the values of basic and derived features (Table I and II).

The angle feature relates to the angle formed in each segment, while distance describes the space between bifurcation and extreme points. Also, length represents the extent of bifurcation to the extreme point, while area defines the pixel count of leaf image. In addition, the total skeleton represents the total segment as a complete entity. Furthermore, total branching point refers to the amount of bifurcation exiting the yield. The total ending points are the quantity of generated ending points by the leaf. These

TABLE I. Primary feature of *shorea* leaves venation

Base Feature	Symbols
Angle	α
Length	l
Distance	d
Area	A
Total Skeleton length	T_{SL}
Total Branching Point	T_{BP}
Total Ending Point	T_{EP}

TABLE II. Feature derivated venation of *shorea* leaves

Derived Features	Symbols	Definition
Angle Difference	δ	$\lvert\alpha_i - \alpha_j\rvert$
Straightness	S	l_i/d_i
Length Ratio	R	$l_i/Max(l)$
Scale Projection	P_{ij}	$\vec{i}\cdot\vec{j}/(Max(d_i, d_j))^2$
Vein Density	D_v	T_{ST}/A
Branching Point Density	D_{BP}	T_{BP}/A
Ending Point Density	D_{EP}	T_{EP}/A

seven primary features are used to produce seven derived components the highlighted equation in Table II.

E. Feature Selection

Feature selection is a very significant phase prior to classification, where relevant properties influencing potential outcomes are considered. This aspect reduces data dimensions and irrelevant components [17], as well as improves the effectiveness and efficiency of applied algorithm [30]. However, the major intuitive measure to determine the relevant properties is by attempting the entire feature combinations. For instance, several features are obtained as F, and also the decision to use or not use each feature, results in a blend of 2^F. Under these conditions, an arrangement with the best performance is preferred. However, the above method appears very time-consuming [31].

Conversely, there is also a wrapper technique for feature selection, including the use of Boruta algorithm to eliminate irrelevant attributes, develop and improve data quality, as well as enhance model performance and accuracy [17].

Boruta algorithm functions by creating copies of a particular set of features to expand the available information. The duplicate is referred as a shadow feature. Subsequently, Boruta algorithm resets the shadow elements, in order to eradicate correlations. Therefore, to determine the most important features, mean reduction impurity (MDI) is applied. This approach trains the shadow feature using RF classifier and also evaluates each duplicate's most important property. In addition, the shadow feature with maximum MDI score signifies the best performance. Also, the algorithm develops tests with only the basic features (excluding clone shadow) in determining attribute importance. As a consequence, Z score is considered. Furthermore, the algorithm performs an implicit evaluation by comparing the feature with a higher Z value and maximum shadow feature. A higher estimate is probably saved into a vector known as a hit. This stage is conducted repeatedly to attain a predetermined iteration value, followed by the creation of a hit table. During iteration, the algorithm determines the feature with the best Z score and tags this estimate as important. The last setoff is obtained from the hit vector, while the first to last technique (get hit vector) repeated until the importance level is attained for the entire feature attributes [17], [30]. Figure 7 represents the algorithm.

F. Random Forest

Random forest is a classification method developed by Leo Breiman. This technique comprises several tree c_l models, where each unit display the classification results; and subsequently the most occurring outcome becomes selected [32]. The random forest algorithm functions by performing the best split search using, Gini index calculation [33]. In addition, a higher value shows a sufficient tendency in becoming the next root or splitting node. This approach is also employed to determine the final label. Equation 1 represents the Gini index calculation [34].

$$Gini(t) = 1 - \sum_{i=1}^{N} p(C_i|t)^2 \tag{1}$$

The final result of random forest describes the summation or voting of individual tree classification. Figure 8 depicts this method architecture. Trees $1, 2, \ldots, b$ is the number of trees used for the classification, while k_1, k_2,..., k_b represents class labels. The random forest produces class k labels after voting on each tree's results [35]. The final random forest classification result is the summation of initial learners' outcome or the results of each tree classification. These results are subsequently added on the class basis. Finally, the class with the maximum number is selected as the final classification result. as observed in Equation 2 [36].

$$f(x) = arg^{max}_{y\epsilon Y} \sum_{j=1}^{i} I(y = h_j(x)) \tag{2}$$

Where $f(x)$ is the result of the classification random forest and $h_j(x)$ represents the outcome of each tree grouping. Meanwhile, $I(y = h_j(x))$ is an indicator function, where the value of 1 similar results as class y or otherwise, the value of 0.

3. Result and Analysis

The extraction of base and derivative features from each Shorea leaf image demonstrated various values and the diversity pattern that was believed to distinguish the sample species. Table III, Table IV, and Table V represent the feature extraction results for Figure 6.

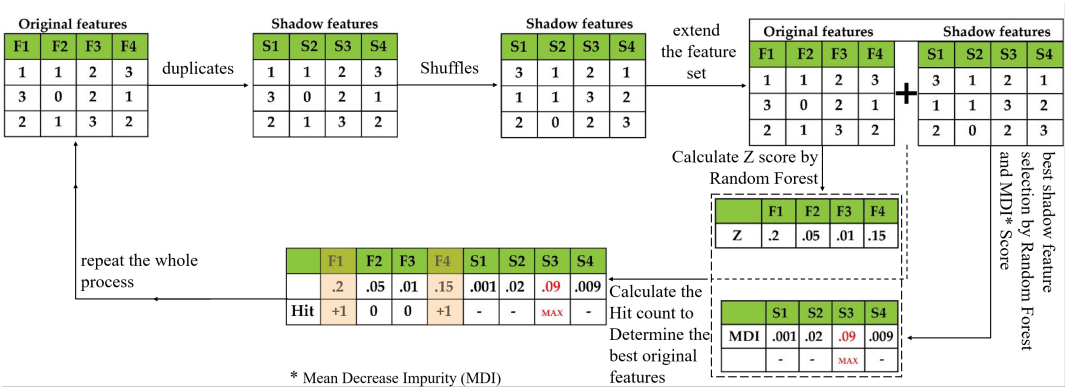

Figure 7. Represents the algorithm

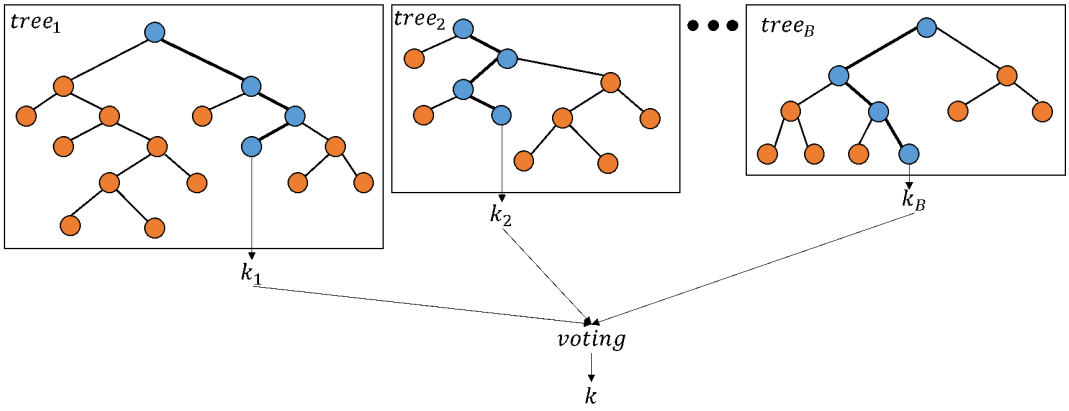

Figure 8. Random forest architecture [35]

TABLE III. Extraction results of angle, length, distance, straightness, and length ratio features

Segment	Ang	Len	Dis	Str	Lra
1	-210.9	6	5.8	1.03	0.02
2	-142.4	26	16.4	1.59	0.10
3	360	9	7	1.29	0.03
4	75.9	13	12.4	1.05	0.05
5	-90	44	43	1.02	0.17
6	347.4	12	9.2	1.30	0.05
7	-230	43	40.5	1.06	0.16
8	63.4	5	4.5	1.12	0.02
9	-91.1	55	54.1	1.02	0.21
10	302.3	46	44.9	1.02	0.18
11	-238.6	96	84.4	1.14	0.37
12	63.4	9	6.7	1.34	0.03
13	88.9	56	55	1.02	0.21
14	330.6	20	18.4	1.09	0.07
15	-233.9	135	122.4	1.10	0.51
16	85.6	53	52.1	1.02	0.20

TABLE IV. Extraction results of scale projection and different angle features

Segment	Adi	Spr
1 and 2	68.532	0.130
2 and 3	149.03	-0.714
3 and 4	142.43	-0.338
4 and 5	102.53	-0.047
4 and 6	166.58	-0.108
5 and 6	148.79	-0.712
6 and 7	175.14	-0.079
7 and 8	118.32	-0.158
8 and 9	177.66	-0.029
9 and 10	113.50	-0.147
10 and 11	142.70	-0.102
11 and 12	132.37	-0.004
12 and 13	158.53	-0.165
13 and 14	63.47	0.351
14 and 15	137.99	-0.104
14 and 16	135	0.2

TABLE V. Extraction results of total skeleton features, total branching point, total ending point, area, branching point density and vein density

Features	TS	TBP	TEP	Area	Branching Point Density	Ending Point Density	Vein Density
Values	43763	37	38	76	0.487	0.5	575.83

Notes: TS = Total Skeleton, TBP = Total Branching Point, TEP = Total Ending Point.

Table III and Table IV represent the value of the feature angle, distance, length, straightness, scale projection, angle difference, and length ratio. However, in order to determine the feature value, a segment (bifurcation and connected extreme points) was acquired. Consequently, calculating the scale projection and angle difference required the segments to overlap. In contrast to the angle, distance, length, straightness, and length ratio features, overlapping was not considered.

Table III and Table IV feature on each leaf data demonstrates more than one value in contrast to the venation attributed listed in Table V, which barely one value. Meanwhile, in Table III and Table IV, the values employed were mean, standard deviation, and variant measurements [15], [16].

Overall, the total features applied in the next process were 28, and were also outlined in Table VI.

The min-max normalization technique was initiated after the value of each feature of *Shorea* leaf venation was obtained.

A. *Feature Selection*

The results acquired using the Boruta algorithm showed 2 features were rejected, while 26 were accepted out of a total of 28. This chosen estimate indicated the components with important information, while two rejected factions well less considered.

Boruta randomly performed 113 forest runs in 24 seconds. The 26 accepted features included Moa, Soa, Voa, Mol, Sol, Vol, Mod, Sod, Vod, Mos, Vos, Msp, Mda, Sda, Vda, Mlr, Slr, Vlr, Ts, Tbp, Tep, Ar, Bpd, and Epd, while the 2 rejected or not important aspects were Ssp and Vsp. Figure 9 shows the feature selection results.

The blue boxplot represents the minimum, average, and maximum z score of the shadow feature, while red and green form the z scores of rejected and accepted features, respectively. These 26 characteristics are adopted as classification datasets, compared to a dataset with complete features.

B. *Application of Random Forest*

In this stage, the classification using the random forest method was conducted with a related package available in R, although several steps are involved.

The first approach is to define the random forest package to request for the library used in developing the classifi-

TABLE VI. Feature value of leaf venation produced from Figure 6

Venation Features	Values
Mean of angle (Moa)	39.957
Stand. dev. of angle (Soa)	221.129
Varian of angle (Voa)	48898.4
Mean of length (Mol)	62.4211
Stand. dev. of length (Sol)	64.3837
Varian of length (Vol)	4145.2603
Mean of distance(Mod)	57.4680
Stand. dev. of distance (Sod)	61.3103
Varian of distance (Vod)	3758.9
Mean of straightness (Mos)	1.1744
Stand. dev. of straightness (Sos)	0.3337
Varian of straightness (Vos)	0.1114
Mean of scale projection (Msp)	0.0046
Stand. dev. of scale projection (Ssp)	0.3009
Varian of scale projection (Vsp)	0.0905
Mean of different angle (Mda)	116.2987
Stand. dev. of different angle (Sda)	44.3795
Varian of different angle (Vda)	1969.546
Mean of length ratio (Mlr)	0.2361
Stand. dev. of length ratio (Slr)	0.2466
Varian of length ratio (Vlr)	0.0608
Total skeleton (Ts)	43763
Total branching point (Tbp)	37
Total ending point (Tep)	38
Area (Ar)	76
Branching point density (Bpd)	0.487
Ending point density (Epd)	0.5
Vein density (Vd)	575.83

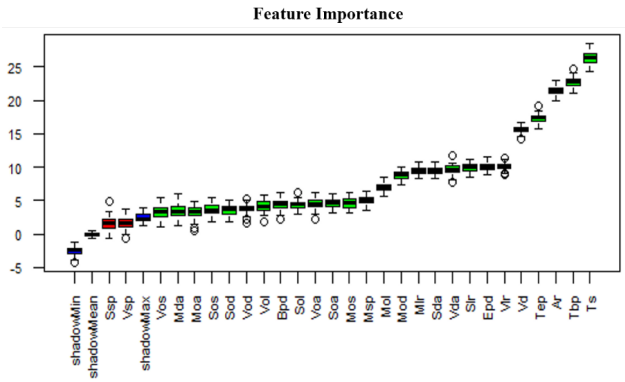

Figure 9. The feature selections results

TABLE VII. Details of accuracy (in%) and running time (in minutes) for each model established from cut-off (90:10, 80:20 and 70:30) and number of single tree 500

Fold	Feature Selection Dataset						All-Feature Dataset					
	90:10		80:20		70:30		90:10		80:20		70:30	
	A	B	A	B	A	B	A	B	A	B	A	B
1	89.29	1.039	97.92	1.045	87.72	1.049	86.67	1.039	88.24	1.040	90.91	1.038
2	90.91	1.046	89.58	1.041	88.41	1.041	90.00	1.039	85.29	1.039	95.38	1.038
3	82.61	1.042	84.78	1.041	90.91	1.040	100.0	1.044	93.18	1.047	90.48	1.040
4	96.00	1.044	95.56	1.047	88.41	1.040	76.92	1.042	90.38	1.039	85.45	1.041
5	90.00	1.047	90.00	1.046	87.30	1.046	85.71	1.041	90.00	1.041	86.21	1.039
Average	89.76	1.043	91.57	1.044	88.55	1.043	87.86	1.041	89.42	1.041	89.69	1.039

Note : A =Accuracy and B = Running time.

TABLE VIII. Details of accuracy (in%) and running time (in minutes) for each model established from cut-off (90:10, 80:20 and 70:30) and number of single tree 700

Fold	Feature Selection Dataset						All-Feature Dataset					
	90:10		80:20		70:30		90:10		80:20		70:30	
	A	B	A	B	A	B	A	B	A	B	A	B
1	80.77	1.051	87.80	1.048	89.23	1.045	96.43	1.051	86.84	1.048	90.16	1.044
2	88.89	1.046	90.00	1.049	83.61	1.045	91.30	1.048	86.36	1.049	94.64	1.043
3	91.67	1.052	91.89	1.051	85.71	1.044	84.62	1.050	91.84	1.044	81.25	1.046
4	88.89	1.055	91.84	1.051	92.19	1.049	94.44	1.044	83.78	1.046	88.71	1.047
5	84.62	1.044	84.21	1.050	89.47	1.046	83.33	1.044	87.10	1.052	85.25	1.045
Average	86.97	1.049	89.15	1.050	88.04	1.046	90.03	1.047	87.18	1.048	88.00	1.045

cation model. Second stage involves parameter setting, including the quantity of single trees built in excess 500, 700, and 1000 with the number of nodes as well as the default predictor variables. Subsequently, the model is developed on the label basin the form of classification reference and training data. Finally, the prediction results are analyzed and further evaluated.

The determination of several cut-offs (90:10, 80:20, and 70:30) was conducted to analyze the sensitive of random forest performance. This circumstance is very important in classifying *Shorea* species, in order to obtain an optimal evaluation value. Table VII shows the calculation results of the classification accuracy obtained from a model known to apply selected and unselected features and the number of single trees 500 as well as the evaluation value of $k-fold$ cross-validation model, where $k = 5$.

The evaluation of the random forest model in Table VII shows the maximum accuracy average value obtained during classification model, using the selection feature dataset with cut-off training and testing data 80:20, with 91.57% accuracy and running time 1.044 minutes. However, the process was not significantly different from similar approach using overall feature dataset, Where the cut-off training and test data 70:30, including the average classification accuracy value of 89.69% with running time 1.039 minutes were applied.

Table VIII shows the classification accuracy calculation results obtained from a model believed to employ selected and unselected features as well as 700 single trees. The evaluation value of $k-fold$ cross-validation model was evaluated, where $k = 5$.

The random forest model evaluation results in Table VIII reported the maximum accuracy average value was obtained using overall feature dataset with a cut-off training and testing data 90:10, and accuracy of 90.03% with running time 1.047 minutes. However, no major variation was observed in comparison to similar approach using the selection feature dataset with cut-off training and testing data 80:20, including the average classification accuracy value of 89.15% with running time 1.050 minutes.

Table IX shows the classification accuracy calculation results obtained from a model known to employ selected and unselected features with 1000 single trees. The evaluation value of $k-fold$ cross-validation model was evaluated, where $k = 5$.

The random forest model evaluation results in Table IX showed the maximum accuracy average value is obtained, using the selection feature dataset with cut-off training and testing data 90:10, with an accuracy of 91.90% with running time 1.054 minutes. Table VIII also revealed the average accuracies of the three cut-off features dataset was more preferred, in comparison to the overall feature dataset.

Based on the evaluation results, Table VII, Table VIII,

TABLE IX. Details of accuracy (in%) and running time (in minutes) for each model established from cut-off (90:10, 80:20 and 70:30) and number of single tree 1000

Fold	Feature Selection Dataset						All-Feature Dataset					
	90:10		80:20		70:30		90:10		80:20		70:30	
	A	B	A	B	A	B	A	B	A	B	A	B
1	77.78	1.054	84.85	1.051	86.89	1.054	90.48	1.068	85.96	1.052	94.03	1.043
2	90.91	1.054	89.19	1.045	89.06	1.050	80.00	1.054	88.10	1.059	85.96	1.047
3	100.0	1.054	100.0	1.054	94.37	1.051	96.00	1.059	91.11	1.051	87.50	1.048
4	95.83	1.054	89.58	1.055	86.49	1.050	78.95	1.055	88.37	1.052	93.33	1.048
5	95.00	1.052	90.00	1.045	93.15	1.050	90.48	1.057	79.07	1.057	88.57	1.049
Average	91.90	1.054	90.72	1.050	89.99	1.051	87.18	1.059	86.52	1.054	89.88	1.047

TABLE X. Comparison of average value of precision and sensitivity (in%) of each dataset using 90:10 cut-Off and number of single trees 1000.

Fold	S.acuminata		S. leprosula		S. ovalis		S. Selanica		Average	
	Preci.	Sensit.	Preci.	Sensit.	Preci.	Sensit.	Preci.	Sensit.	Preci.	Sensit.
1	100.0	100.0	87.50	77.78	81.82	90.00	100.0	100.0	92.33	91.945
2	100.0	100.0	85.71	66.67	75.00	85.71	100.0	100.0	90.18	88.10
3	100.0	100.0	90.91	71.43	60.00	75.00	100.0	100.0	87.73	86.61
4	100.0	100.0	85.71	92.31	87.50	77.78	100.0	100.0	93.30	92.52
5	92.30	100.0	100.0	71.43	66.67	100.0	100.0	90.91	89.74	90.59
Average	98.46	100.0	89.97	75.92	74.20	85.70	100.0	98.18	90.66	89.95

Note: Preci. = Precision and Sensit. = Sensitivity.

and Table IX represent the random forest performance, using a dataset of selection features and cut-off training and testing data 90:10, with 1000 trees as the best model. This conclusion was evidenced by the optimum average accuracy of 91.90% with running time 1.054 minutes.

Apart from using accuracy to measure a model performance, precision and sensitivity also offer effective alternatives. The precision calculation is ascertaining the classification significance. In contrast, sensitivity determine the extent of balance in classifying the actual class correctly. Sensitivity and precision testing applies a selection feature dataset with cut-off training and testing data 90:10,in addition to 1000 single trees. Table X presents the precision and sensitivity values for each class.

Table X shows that *S. leprosula* and *S. ovalis* obtained the minimum average sensitivity and precision value of 75.92 and 74.20%, respectively. Meanwhile, *S. acuminata* and *S. Selanica* both acquired the maximum sensitivity and precision values of 100%. This showed the individual class actual data was correctly classified, in order to attain a high category, containing the proportional estimate. The conclusions were evidenced in the average precision and sensitivity of 90.66 and 89.96%, respectively. Based on the above estimates, the model demonstrated an excellent performance, as the average precision was greater, compared to the sensitivity, although the difference was not very significant.

4. Conclusions

Information regarding leaf morphology is very important to study. Through leaves, botanists can identify the characteristics of plants. Leaf venation is a part of the leaf that can be used as a characteristic of plant species because it has different patterns. This study identifying *Shorea* species, based on the features found in leaf venation. The features in leaf venation were successfully extracted, namely: features angle, distance, length, straightness, scale projection, different angle, length ratio, total skeleton, total branching point, total ending point, area, branching point density, ending point density and vein density. This feature is used as a dataset to build a classification model using the random forest classification technique.

The results showed the success of the applied model in classifying *Shorea* species, by leaf venation feature. Also, optimum accuracy was attained at 91.90%, with a cut-off training and testing data of 90:10, by analyzing 1000 single trees. Furthermore, extensive sensitivity and precision values were obtained at 89.95 and 90.66%, respectively. These results clearly indicated a superior performance model.

Acknowledgment

We greatly appreciate the research support from both Department of Computer Science, Faculty of Mathematics and Natural Sciences and Department of Silviculture, Faculty of Forestry and Environment, IPB University, Bogor, Indonesia.

References

[1] M. F. Newman, P. F. Burgess, T. C. Whitmore *et al.*, *Manuals of dipterocarps for foresters: Java to New Guinea.* Bogor, Indonesia:

Prosea Indonesia, 1998.

[2] E. Kintamani *et al.*, "The diversity of shorea spp.(meranti) at some habitats in indonesia," in *IOP Conference Series: Earth and Environmental Science*, vol. 197, no. 1. IOP Publishing, 2018, p. 012034.

[3] A. Eni, I. Dewantara, and L. Sisillia, "Identification of tengkawang (shorea spp) species as natural days of tenun ikat kapuas hulu regency west borneo," *J Hutan Lestari*, vol. 6, no. 1, pp. 7–15, 2018.

[4] "IUCN Red List: Daftar Merah Spesies Shorea." [Online]. Available: https://www.iucnredlist.org

[5] D. Dodo and H. Wawangningrum, "Caring wildlings methods for ex situ conservation: red balau (shorea guiso (blanco) blume)," in *Prosiding Seminar Nasional Masyarakat Biodiversitas Indonesia*, vol. 4, no. 2, 2018, pp. 139–143.

[6] Bastoni, *Ecological and silvicultural studies of ramin in South Sumatra and Jambi*. Bogor, Indonesia: ITTO and PPD, 2005.

[7] T. Jin, X. Hou, P. Li, and F. Zhou, "A novel method of automatic plant species identification using sparse representation of leaf tooth features," *PloS one*, vol. 10, no. 10, p. e0139482, 2015.

[8] S. Singh and M. S. Bhamrah, "Leaf identification using feature extraction and neural network," *IOSR Journal of Electronics and Communication Engineering*, vol. 5, pp. 134–140, 2015.

[9] I. Ariawan, Y. Herdiyeni, and I. Z. Siregar, "Geometric morphometric analysis of leaf venation in four shorea species for identification using digital image processing," *Biodiversitas Journal of Biological Diversity*, vol. 21, no. 7, 2020.

[10] T.-L. Le, D.-T. Tran, and V.-N. Hoang, "Fully automatic leaf-based plant identification, application for vietnamese medicinal plant search," in *Proceedings of the fifth symposium on information and communication technology*, 2014, pp. 146–154.

[11] L. J. Hickey, "Classification of the architecture of dicotyledonous leaves," *American journal of botany*, vol. 60, no. 1, pp. 17–33, 1973.

[12] A. Roth-Nebelsick, D. Uhl, V. Mosbrugger, and H. Kerp, "Evolution and function of leaf venation architecture: a review," *Annals of Botany*, vol. 87, no. 5, pp. 553–566, 2001.

[13] A. Kadir, L. E. Nugroho, A. Susanto, and P. I. Santosa, "A comparative experiment of several shape methods in recognizing plants," *arXiv preprint arXiv:1110.1509*, 2011.

[14] M. F. Ab Jabal, S. Hamid, S. Shuib, and I. Ahmad, "Leaf features extraction and recognition approaches to classify plant," *Journal of Computer Science*, vol. 9, no. 10, p. 1295, 2013.

[15] A. Ambarwari, Y. Herdiyeni, and I. Hermadi, "Biometric analysis of leaf venation density based on digital image," *Telkomnika*, vol. 16, no. 4, pp. 1735–1744, 2018.

[16] R. De Oliveira Plotze and O. M. Bruno, "Automatic leaf structure biometry: computer vision techniques and their applications in plant taxonomy," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 23, no. 02, pp. 247–262, 2009.

[17] M. B. Kursa, W. R. Rudnicki *et al.*, "Feature selection with the boruta package," *J Stat Softw*, vol. 36, no. 11, pp. 1–13, 2010.

[18] Z. Yang, M. Jin, Z. Zhang, J. Lu, and K. Hao, "Classification based on feature extraction for hepatocellular carcinoma diagnosis using high-throughput dna methylation sequencing data," *Procedia Computer Science*, vol. 107, pp. 412–417, 2017.

[19] V. Fortino, P. Kinaret, N. Fyhrquist, H. Alenius, and D. Greco, "A robust and accurate method for feature selection and prioritization from multi-class omics data," *PloS one*, vol. 9, no. 9, p. e107801, 2014.

[20] K. Ponmalar and K. Krishnaveni, "Random forest classification of medicinal plant leaves using shape and texture features," *J. Appl. Sci. Comput*, vol. 5, no. 8, pp. 561–569, 2018.

[21] B. Sandika, S. Avil, S. Sanat, and P. Srinivasu, "Random forest based classification of diseases in grapes from images captured in uncontrolled environments," in *2016 IEEE 13th International Conference on Signal Processing (ICSP)*. IEEE, 2016, pp. 1775–1780.

[22] N. Apao, L. Feliscuzo, C. Romana, and J. Tagaro, "Multiclass classification using random forest algorithm to prognosticate the level of activity of patients with stroke," *Int. J. Sci. Technol. Res*, vol. 9, no. 4, pp. 1233–1240, 2020.

[23] S. Agrawal and R. S. Rajput, "Denial of services attack detection using random forest classifier with information gain," *International Journal of Engineering Development and Research*, vol. 5, no. 3, pp. 929–938, 2017.

[24] R. C. Gonzalez and R. E. Woods, *Digital image processing and computer vision*, 3rd ed. New Jersey: Pearson Prentice Hall, 2008.

[25] A. McAndrew, "An introduction to digital image processing with matlab notes for scm2511 image processing," *School of Computer Science and Mathematics, Victoria University of Technology*, vol. 264, no. 1, pp. 1–264, 2004.

[26] K. Bhargavi and S. Jyothi, "A survey on threshold based segmentation technique in image processing," *International Journal of Innovative Research and Development*, vol. 3, no. 12, pp. 234–239, 2014.

[27] M. H. Siddiqi, I. Ahmad, and S. B. Sulaiman, "Weed recognition based on erosion and dilation segmentation algorithm," in *2009 International Conference on Education Technology and Computer*. IEEE, 2009, pp. 224–228.

[28] K. Sreedhar and B. Panlal, "Enhancement of images using morphological transformation," *arXiv preprint arXiv:1203.2514*, 2012.

[29] L. Abhishek, "Thinning approach in digital image processing," *Special Issue-SACAIM*, pp. 326–330, 2017.

[30] R. Nilsson, J. M. Pena, J. Björkegren, and J. Tegnér, "Consistent feature selection for pattern recognition in polynomial time," *The Journal of Machine Learning Research*, vol. 8, pp. 589–612, 2007.

[31] J. Wira, "Pengenalan pembelajaran mesin dan deep learning jan wira gotama putra pengenalan konsep pembelajaran mesin dan deep learning," *no. March*, vol. 2019, 2018.

[32] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[33] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and regression trees*. Boca raton: Routledge, 2017.

[34] K. Fawagreh, M. M. Gaber, and E. Elyan, "Random forests: from early developments to recent advancements," *Systems Science & Control Engineering: An Open Access Journal*, vol. 2, no. 1, pp. 602–609, 2014.

[35] A. Verikas, A. Gelzinis, and M. Bacauskiene, "Mining data with random forests: A survey and results of new tests," *Pattern recognition*, vol. 44, no. 2, pp. 330–349, 2011.

[36] A. Cutler, D. R. Cutler, and J. R. Stevens, "Random forests," in *Ensemble machine learning*. Springer, 2012, pp. 157–175.

Ishak Ariawan Ishak Ariawan completed his Master of Computer Science from IPB University in 2019 with a thesis on morphometrics analysis of leaf. He obtained his first-degree informatics and computer engineering education from Universitas Negeri Makassar, Indonesia (2015). He is currently a Lecturer in Marine Information Systems, Universitas Pendidikan Indonesia, Serang, Banten, Indonesia. His research areas are Computer Vision, Machine Learning, and Ecological Informatics.

Yeni Herdiyeni Yeni Herdiyeni is an expert in digital image processing, computer vision and computational intelligence. She earned a PhD in Computer Science with the Dissertation on Semantic Image Similarity using Tree from University of Indonesia (2010). Then she conducted Post Doctoral research at Department of Information Science, Graduate School of Science and Engineering, Saga University, Japan, for 5 months (September - January 2012) . In 2014 to 2016, she also conducted research in Paris Diderot for morphometrics analysis of leaf plant. She had a Master of Computer Science from University of Indonesia with the thesis on 3D Face Recognition (2005).

Iskandar Zulkarnaen Siregar Iskandar Zulkarnaen Siregar is a Professor at the Department of Silviculture, Faculty of Forestry, IPB University, Indonesia. His research interests are Population Genetics of Forest Trees, Molecular Genetics of Forest Plants, Conservation and Sustainable Management of Tropical Forest Genetic Resources, Forest Tree Improvement, Forest Adaptation and Genetics in Silviculture. Current research projects are Genetic Variation of Shorea spp, Aquillaria spp., Melia spp. and Toona spp, Genetic Marker based Chain of Custody (CoC) in Timber (teak, agarwood and dipterocarps), Genetic Risks of Species Extincion (Endangered Shorea spp.), Conservation of Tree Genetic Resources outside Forests. Currently, he serves also as Director for International Program.