

**AUTOMATED CLASSIFICATION OF TROPICAL SHRUB
SPECIES: A HYBRID OF LEAF SHAPE AND MACHINE
LEARNING APPROACH**

MIRAEMILIANA BINTI MURAT

**FACULTY OF SCIENCE
UNIVERSITY OF MALAYA
KUALA LUMPUR**

2018

**AUTOMATED CLASSIFICATION OF TROPICAL SHRUB
SPECIES: A HYBRID OF LEAF SHAPE AND MACHINE
LEARNING APPROACH**

MIRAEMILIANA BINTI MURAT

**DISSERTATION SUBMITTED IN FULFILMENT OF
THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE**

**INSTITUTE OF BIOLOGICAL SCIENCES
FACULTY OF SCIENCE
UNIVERSITY OF MALAYA
KUALA LUMPUR**

2018

UNIVERSITY OF MALAYA
ORIGINAL LITERARY WORK DECLARATION

Name of Candidate: **Miraemiliana Binti Murat)** [REDACTED]

Matric No: **SGR160015**

Name of Degree: **Master of Science**

Title of Project Paper/Research Report/Dissertation/Thesis ("this Work"):

**AUTOMATED CLASSIFICATION OF TROPICAL SHRUB SPECIES:
A HYBRID OF LEAF SHAPE AND MACHINE LEARNING APPROACH**

Field of Study: **Bioinformatics**

I do solemnly and sincerely declare that:

- (1) I am the sole author/writer of this Work;
- (2) This Work is original;
- (3) Any use of any work in which copyright exists was done by way of fair dealing and for permitted purposes and any excerpt or extract from, or reference to or reproduction of any copyright work has been disclosed expressly and sufficiently and the title of the Work and its authorship have been acknowledged in this Work;
- (4) I do not have any actual knowledge nor do I ought reasonably to know that the making of this work constitutes an infringement of any copyright work;
- (5) I hereby assign all and every rights in the copyright to this Work to the University of Malaya ("UM"), who henceforth shall be owner of the copyright in this Work and that any reproduction or use in any form or by any means whatsoever is prohibited without the written consent of UM having been first had and obtained;
- (6) I am fully aware that if in the course of making this Work I have infringed any copyright whether intentionally or otherwise, I may be subject to legal action or any other action as may be determined by UM.

Candidate's Signature

Date:

Subscribed and solemnly declared before,

Witness's Signature

Date:

Name:

Designation:

AUTOMATED CLASSIFICATION OF TROPICAL SHRUB SPECIES: A HYBRID OF LEAF SHAPE AND MACHINE LEARNING APPROACH

ABSTRACT

Plants play an important role in foodstuff, medicine, industry, and environmental protection. The plant recognition is very crucial in some applications, including conservation of endangered species and rehabilitation of lands after mining activities. But, it is a challenging task to identify plant species because it requires specialised knowledge. Therefore, developing an automated classification system for plant species is necessary and valuable since it can help specialists as well as the public in identifying plant species easily. In this study, shape descriptors are applied on the myDAUN dataset that contains 45 tropical shrub species, which are collected from the University of Malaya (UM), Malaysia. Four types of shape descriptors are used in this study namely morphological shape descriptors (MSD), histogram of oriented gradients (HOG), Hu invariant moments (Hu) and Zernike moments (ZM). Single descriptor, as well as the combination of hybrid descriptors are tested and compared. The tropical shrub species are classified using six different classifiers, which are artificial neural network (ANN), random forest (RF), support vector machine (SVM), k-nearest neighbour (k-NN), linear discriminant analysis (LDA) and directed acyclic graph multiclass least squares twin support vector machine (DAG MLSTSVM). In addition, three types of feature selection methods are tested in the myDAUN dataset, namely Relief, Correlation-based feature selection (CFS) and Pearson's coefficient correlation (PCC). The well-known Flavia dataset and Swedish Leaf dataset are used as the validation dataset on the proposed methods. The results showed that the hybrid of all descriptors of ANN outperformed the other classifiers with an average classification accuracy of 98.23% for myDAUN dataset, 95.25% for Flavia dataset and 99.89% for Swedish Leaf dataset. In addition, the Relief feature selection method achieved the highest classification accuracy of 98.13% after 80

(or 60%) of the original features are reduced, from 133 to 53 descriptors in myDAUN dataset with the reduction in computational time. Subsequently, the hybridisation of four descriptors gave the best results compared to others. It is proven that the combination of MSD and HOG are good enough for tropical shrubs species classification. Hu and ZM descriptors also improved the accuracy in tropical shrubs species classification in terms of invariant to translation, rotation and scale. ANN outperformed the others for tropical shrub species classification in this study. Feature selection methods can be used in the classification of tropical shrub species, as the comparable results could be obtained with the reduced descriptors while reducing in computational time and cost.

Keywords: shape descriptors, feature extraction, species identification, machine learning, artificial neural network

KLASIFIKASI AUTOMATIK UNTUK SPESIES TUMBUHAN RENEK TROPIKA: PENDEKATAN HIBRID DENGAN KAEDAH BENTUK DAUN DAN PEMBELAJARAN MESIN

ABSTRAK

Tumbuhan memainkan peranan penting dalam bahan makanan, perubatan, industri, dan perlindungan alam sekitar. Mengenalpasti tumbuhan sangat penting dalam beberapa aplikasi, termasuk pemuliharaan spesies terancam dan pemulihan alam sekitar selepas aktiviti perlombongan. Tetapi, ia adalah tugas yang mencabar untuk mengenal pasti spesies tumbuhan kerana ia memerlukan pengetahuan khusus. Membangunkan sistem klasifikasi automatik untuk spesies tumbuhan adalah penting dan berharga kerana ia dapat membantu pakar serta orang ramai dalam mengenal pasti spesies tumbuhan dengan mudah. Pemerihal bentuk telah digunakan pada dataset myDAUN yang mengandungi 45 spesies tumbuhan renek tropika yang dikumpul dari Universiti Malaya (UM), Malaysia. Empat jenis pemerihal bentuk digunakan dalam kajian ini iaitu *Morphological Shape Descriptor (MSD)*, *Histogram of Oriented Gradients (HOG)*, *Hu invariant moments (Hu)* dan *Zernike moments (ZM)*. Pemerihal tunggal, serta kombinasi pemerihal hibrid telah diuji dan dibandingkan. Spesies tumbuhan renek tropika diklasifikasikan menggunakan enam klasifikasi yang berbeza iaitu *artificial neural network (ANN)*, *random forest (RF)*, *support vector machine (SVM)*, *k-nearest neighbour (k-NN)*, *linear discriminant analysis (LDA)* dan *directed acyclic graph multiclass least squares twin support vector machine (DAG MLSTSVM)*. Di samping itu, tiga jenis kaedah pemilihan ciri telah diuji di dalam dataset myDAUN iaitu *Relief*, *Correlation-based feature selection (CFS)* dan *Pearson coefficient correlation (PCC)*. Dataset Flavia dan Swedish Leaf telah digunakan sebagai dataset pengesahan untuk kaedah yang dicadangkan. Keputusan menunjukkan bahawa hibrid dari semua pemerihal dengan klasifikasi *ANN* mengatasi pengelasan lain dengan ketepatan klasifikasi purata sebanyak 98.23% untuk dataset myDAUN, 95.25% untuk

dataset Flavia dan 99.89% untuk dataset Swedish Leaf. Di samping itu, kaedah pemilihan ciri *Relief* mencapai ketepatan pengelasan tertinggi sebanyak 98.13% selepas 80 (atau 60%) ciri asal dikurangkan, iaitu dari 133 hingga 53 pemerihal dalam dataset myDAUN dengan pengurangan masa pengiraan. Selepas itu, kombinasi empat pemerihal memberikan hasil yang terbaik berbanding yang lain. Ini membuktikan bahawa gabungan *MSD* dan *HOG* cukup baik untuk klasifikasi spesies tumbuhan renek tropika. *Hu* dan *ZM* pemerihal juga meningkatkan ketepatan dalam klasifikasi spesies tumbuhan renek tropika dari segi invarian kepada terjemahan, putaran dan skala. *ANN* mengatasi yang lain untuk klasifikasi spesies tumbuhan renek tropika dalam kajian ini. Kaedah pemilihan ciri boleh digunakan dalam klasifikasi spesies tumbuhan renek, kerana keputusan yang seimbang boleh diperolehi tetapi dengan berkurangnya dari segi pemerihal, masa dan kos.

Kata kunci: pemerihal bentuk, pengekstrakan ciri, identifikasi spesies, pembelajaran mesin, *artificial neural network*

ACKNOWLEDGEMENTS

I would like to thank the Almighty God for the blessings bestowed on me with everything; in whatever I do throughout my life and for giving me strength and guidance for the successful completion of this thesis. A great deal of personal effort was required to successfully complete this thesis. Nonetheless this would not have been possible without the kind support and guidance of many individuals and organizations. I would like to extend my sincere thanks to all of them.

I am highly indebted to my invaluable main supervisor, Dr Chang Siow Wee for her patience, guidance and supervision as well as for providing the necessary information and her support in completing this thesis. A million thanks for her encouragement, helpfulness and quick response whenever I needed advice, but mostly for believing in me in all these years.

I would like to express my appreciation and gratitude towards my co-supervisor Dr Arpah Abu for her encouragement and guidance throughout the process of completing this thesis. Her immediate response to inquiries and uncertainties as well as her constructive criticism have helped me to improve my scientific skill set tremendously.

To my beloved parents, Murat bin Katijan and Rohana binti Murad, without their constant support, guidance and love, I am just nobody. To my siblings, thanks for their understanding and overwhelming support and motivation. From the bottom of my heart, I am greatly indebted to them. To my friends especially to Afrina, Jing Wei and Mei Sze, I owe thanks for their motivation, encouragement, helps and ideas.

Thank you for all your encouragement.

TABLE OF CONTENTS

ABSTRACT	iii
ABSTRAK	v
ACKNOWLEDGEMENTS.....	vii
TABLE OF CONTENTS.....	viii
LIST OF FIGURES	xii
LIST OF TABLES	xv
LIST OF SYMBOLS AND ABBREVIATIONS	xvii
LIST OF APPENDICES	xix
CHAPTER 1: INTRODUCTION.....	1
1.1 Background.....	1
1.2 Problem Statement.....	6
1.3 Research Objective	7
1.4 Scope of the Study	8
1.5 Research Significance.....	8
1.6 Chapter Organization.....	9
CHAPTER 2: LITERATURE REVIEW.....	11
2.1 Introduction.....	11
2.2 Plant Leaf Structure	11
2.2.1 Plant Leaf Image Databases	13
2.3 Feature Extraction.....	19
2.3.1 Shape	20
2.3.2 Morphological Shape Descriptors (MSD).....	20
2.3.3 Region-based Shape Descriptor	21
2.3.4 Boundary-based Shape Descriptors.....	24
2.4 Feature Selection Methodologies	29
2.4.1 Relief	29
2.4.2 Correlation-based Feature Selection (CFS).....	30
2.4.3 Pearson's Correlation Coefficient (PCC)	31
2.4.4 Genetic Algorithm (GA)	31
2.5 Image Classification Methodologies	32
2.5.1 Artificial Neural Network (ANN)	33

2.5.2	Random Forest (RF)	34
2.5.3	Support Vector Machine (SVM)	35
2.5.4	k- Nearest Neighbour (k-NN)	36
2.5.5	Linear Discriminant Analysis (LDA)	37
2.5.6	Naïve Bayes	38
2.6	Previous Studies in Plant Species Classification	39
2.6.1	Boundary-based Shape Descriptor	40
2.6.2	Region-based Shape Descriptor	41
2.6.3	Combination of Shape Descriptors	43
2.7	Automated Plant Species Identification System	49
2.7.1	Web-based Plant Identification System	49
2.7.2	Mobile Apps	54
2.8	Summary	59
CHAPTER 3: RESEARCH METHODOLOGY		60
3.1	Introduction	60
3.2	Proposed Framework	60
3.3	Field Sampling	62
3.4	Image Acquisition	66
3.5	Software and Hardware Requirement	68
3.6	Image Pre-processing	70
3.7	Image Segmentation	71
3.8	Feature Extraction	74
3.8.1	Morphological Shape Descriptor (MSD)	74
3.8.2	Histogram of Oriented Gradients (HOG)	79
3.8.3	Hu Invariant Moments (Hu)	80
3.8.4	Zernike Moments (ZM)	82
3.9	Feature Selection	83
3.9.1	Relief-F	83
3.9.2	Correlation-based Feature Selection (CFS)	84
3.9.3	Pearson's Correlation Coefficient (PCC)	85
3.10	Classification	86
3.10.1	Artificial Neural Network (ANN)	87
3.10.2	Random Forest (RF)	87
3.10.3	k – Nearest Neighbour (k-NN)	88

3.10.4	Support Vector Machine (SVM).....	89
3.10.5	Linear Discriminant Analysis (LDA)	90
3.10.6	Directed Acyclic Graph Multiclass Least Squares Twin Support Vector Machine (DAG MLSTSVM)	91
3.11	Cross-validation (CV).....	92
3.12	Summary.....	93
CHAPTER 4: RESULTS.....		94
4.1	Introduction.....	94
4.2	Results of Data Collection	94
4.3	Feature Extraction Methods.....	95
4.4	Classifiers	96
4.4.1	Artificial Neural Network (ANN)	96
4.4.2	Random Forest (RF).....	101
4.4.3	Support Vector Machine (SVM)	105
4.4.4	k-Nearest Neighbour (k-NN).....	109
4.4.5	Linear Discriminant Analysis (LDA).....	113
4.4.6	Directed Acyclic Graph Multiclass Least Squares Twin Support Vector Machine (DAG MLSTSVM)	117
4.5	Results of Feature Selection	121
4.6	Cross-validation (CV).....	123
4.7	Validation using Flavia and Swedish Leaf dataset.....	124
4.7.1	Flavia Dataset.....	124
4.7.2	Swedish Leaf Dataset	126
4.8	Summary.....	128
CHAPTER 5: DISCUSSIONS		129
5.1	Leaves.....	129
5.2	Leaf Shape	130
5.3	Image Processing.....	133
5.4	Single Descriptor	134
5.5	Hybrid of Descriptor.....	136
5.6	Feature Selection	139
5.7	Cross-validation (CV).....	140
5.8	Validation	140

5.9	Comparison Studies	143
-----	--------------------------	-----

CHAPTER 6: CONCLUSION AND FUTURE WORK145

6.1	Introduction.....	145
6.2	Research Summary	145
6.3	Research Constraints	148
6.4	Research Contributions.....	149
6.5	Future Work.....	150
6.5.1	Larger Amount of Data Collection.....	150
6.5.2	Utilising Other Descriptors	150
6.5.3	Utilising Other Classifiers	151
6.5.4	Mobile Apps	151

REFERENCES.....152

LIST OF PUBLICATIONS AND PAPERS PRESENTED163

APPENDICES165

LIST OF FIGURES

Figure 1.1 : General classification of plants in the world	3
Figure 2.1 : Leaf structure	12
Figure 2.2 : Leaf types	13
Figure 2.3 : Leaf samples in Flavia dataset.....	14
Figure 2.4 : Leaf samples in Swedish Leaf dataset.....	15
Figure 2.5 : Leaf samples in Leafsnap dataset	15
Figure 2.6 : Leaf samples in ICL dataset	16
Figure 2.7 : Leaf samples in ImageCLEF dataset.....	17
Figure 2.8 : Categorization and overview of the most prominent shape feature descriptors in plant species identification	19
Figure 2.9 : Model of artificial neural network (ANN)	34
Figure 2.10 : Model of a two dimensional hyperplane	36
Figure 2.11 : Model of k-nearest neighbour classification.	37
Figure 2.12 : Model of linear discriminant analysis (LDA) classification.	38
Figure 2.13 : Model of Naïve Bayes classification.....	39
Figure 2.14 : iNaturalist web interface.....	50
Figure 2.15 : Pl@ntNet web interface.....	51
Figure 2.16 : Leaf Recognition interface	52
Figure 2.17 : Leafsnap interface.....	55
Figure 2.18 : Pl@ntNet interface	56
Figure 2.19 : FOLIA interface	57

Figure 3.1 : Flowchart for the proposed methodology.....	61
Figure 3.2 : Location of sampling area in the University of Malaya (UM), Kuala Lumpur, Malaysia.	63
Figure 3.3 : Arrangement of leaves before compression	67
Figure 3.4 : Light box and experimental setup	67
Figure 3.5 : Samples of the leaf images in the myDAUN dataset	69
Figure 3.6 : A <i>Lantana camara</i> sample before image enhancement	70
Figure 3.7 : A <i>Lantana camara</i> sample after cleaned using Photoshop CC.....	70
Figure 3.8 : Conversion of RGB image into grey-scale image.....	71
Figure 3.9 : Conversion of grey-scale image into detected edge image.	72
Figure 3.10 : Conversion of detection edge image into binary image.	72
Figure 3.11 : Conversion of binary image to filled binary image.....	73
Figure 3.12 : Conversion of filled binary image to region of interest (ROI) image.	74
Figure 3.13 : The flowchart of the Relief-F feature selection.....	84
Figure 3.14 : The flowchart of the correlation-based feature selection	85
Figure 3.15 : The flowchart of the Pearson's correlation coefficient.....	86
Figure 3.16 : Neural network for tropical shrub species.....	87
Figure 4.1 : Comparison of ANN accuracy with various sets of descriptors	100
Figure 4.2 : Comparison of RF accuracy with various sets of descriptors	104
Figure 4.3 : Comparison of SVM accuracy with various sets of descriptors.....	108
Figure 4.4 : Comparison of k-NN accuracy with various sets of descriptors	112
Figure 4.5 : Comparison of LDA accuracy with various sets of descriptors	116

Figure 4.6 : Comparison of DAG MLSTSVM accuracy with various sets of descriptors	120
Figure 5.1 : A variety of leaf samples of <i>Acalypha wilkesiana</i> species	131
Figure 5.2 : A variety of leaf samples of <i>Loropetalum chinensis</i> species.....	131
Figure 5.3 : A <i>Manihot esculenta</i> leaf sample	132
Figure 5.4 : The overall steps in image pre-processing . (a) original image, (b) grayscale image, (c) detected edge, (d) binary image, (e) filled binary image, (f) ROI image.....	134
Figure 5.5 : A leaf sample of <i>Mussaenda sp.</i>	138
Figure 5.6 : A leaf sample of <i>Ilex macrocarpa</i> and <i>Chimonanthus praecox</i> in Flavia dataset.....	141

LIST OF TABLES

Table 2.1	: A summary of the features of existing leaf dataset.....	18
Table 2.2	: A comparison of classification accuracies on the leaf identification and classification studies	46
Table 2.3	: A summary of the features and requirements of web-based plant identification systems.....	53
Table 2.4	: A summary of the features and requirements of plant automated identification systems of mobile apps	58
Table 3.1	: List of tropical shrub species in myDAUN dataset	64
Table 3.2	: Basic geometrical and morphological shape descriptors	75
Table 4.1	: Data collection of myDAUN dataset	94
Table 4.2	: Combinations of descriptors for feature extraction	95
Table 4.3	: Testing accuracy of three set data division with various set of neurons...	97
Table 4.4	: Classification accuracy for ANN classifier using single descriptor	97
Table 4.5	: Classification for ANN classifier using hybrid of descriptors.....	99
Table 4.6	: Classification accuracy for RF classifier using single descriptor	101
Table 4.7	: Classification for RF classifier using hybrid of descriptors.....	103
Table 4.8	: Classification accuracy for SVM classifier using single descriptor	105
Table 4.9	: Classification for SVM classifier using hybrid of descriptors.....	107
Table 4.10	: Accuracy of number of nearest neighbour.....	109
Table 4.11	: Classification accuracy for k-NN classifier using single descriptor.....	110
Table 4.12	: Classification for k-NN classifier using hybrid of descriptors	111
Table 4.13	: Classification accuracy for LDA classifier using single descriptor.....	113

Table 4.14	: Classification for LDA classifier using hybrid of descriptors	115
Table 4.15	: Classification accuracy for DAG MLSTSVM classifier using single descriptor.....	117
Table 4.16	: Classification for DAG MLSTSVM classifier using hybrid of descriptors	119
Table 4.17	: Categories of descriptors reduction	121
Table 4.18	: Classification accuracy for the selected feature selection methods	122
Table 4.19	: Running time for features extraction	123
Table 4.20	: Validation result with cross-validation	123
Table 4.21	: Classification results of Flavia dataset.....	125
Table 4.22	: Classification results of Swedish Leaf dataset.....	127
Table 5.1	: Classification accuracy for single descriptor	135
Table 5.2	: Classification accuracy of hybrid descriptors	137
Table 5.3	: Classification results of Flavia, Swedish Leaf and myDAUN dataset ...	142
Table 5.4	: The performance of our proposed method compared to other leaf classification studies	144

LIST OF SYMBOLS AND ABBREVIATIONS

ANN	Artificial Neural Network
CCD	Centroid Contour Distance
CCG	Centroid Contour Gradient
CFS	Correlation-Based Feature Selection
CSS	Curvature Scale Space
DAG	Directed Acyclic Graph
MLSTSVM	Multiclass Least Squares Twin Support Vector Machine
GLCM	Gray-Level Co-Occurrence Matrices
HOG	Histogram Of Oriented Gradient
HU	Hu Invariant Moments
IDSC	Inner-distance Shape Context
k-NN	K- Nearest Neighbour
LBP	Local Binary Pattern
LDA	Linear Discriminant Analysis
LDC	Linear Discriminant Classifier
MARCH	Multi- Scale Arch Height Descriptor
MDM	Multi-Scale Distance Matrix
MDP	Modified Dynamic Programming
MLP	Multi Layer Perceptron Using Back Propagation
MMC	Move Median Centre
MSD	Morphological Shape Descriptor
NFC	Scaled Conjugate Gradient Algorithm
PCC	Pearson's Correlation Coefficient
PNN	Probabilistic Neural Network
PRT	Polar Fourier Transform
RF	Random Forest
ROI	Region Of Interest
SIFT	Scale Invariant Feature Transform

SVM	Support Vector Machines
TAR	Triangle Area Representation
TOA	Triangle Oriented Angles
TSL	Triangle Side Length Representation
TSLA	Triangle Side Lengths And Angle Representation
ZM	Zernike moments

University of Malaya

LIST OF APPENDICES

Appendix A	Images in myDAUN dataset	165
Appendix B	Parameters in ANN	180

University of Malaya

CHAPTER 1: INTRODUCTION

1.1 Background

There are more than seven million species of plants and animals live on Earth globally (Chapman, 2009). Plants play an important and vital role in human life in which plants are essential resources for human being and exist everywhere (Tandon et al., 2007). Most of the plants bring significant information for the development of human society and are considered as crucial resource for human being. Plants form a fundamental part of life as they form the base for food chain and a lot of medicines are derived from plants. Plants also play a significant role in environmental protection (Tilman et al., 2002).

The increasing anthropogenic pressure on the natural environment has driven a steadily decline of biodiversity towards the verge of extinction (Hore et al., 1997; Mata-montero & Carranza-Rojas, 2016; Pimm et al., 2014). The resulting ecological crisis has brought many serious environmental effect including flash floods, climate changes, desertification and so on (Geertsema et al., 2009; Wiens, 2016; Wilby & Keenan, 2012). RBG Kew (2016) reveals that there are currently 391,000 vascular plants species known to science and the study found that 2034 new vascular plant species were discovered in 2015 and 1730 new vascular plant species were discovered in 2016. However, this report stated that 20% of the plants are at risk of extinction with threats, including climate change, habitat loss, disease and invasive species, which is one in five plants is estimated to be threatened with extinction.

Nowadays, people have better understanding about the importance and urgency to conserve and protect plant resources (Kazerouni et al., 2015). According to the Willis (2017) there are 28,187 plant species currently recorded as being of medicinal use. Increasing demand of herbal medicines threatens the wild populations of many of these

plants. Thus, in order to conserve and protect the plant species, it is crucial for the general public to be able to identify and recognise the many of plant species (Corlett, 2016).

Due to numerous types of plants, a classification system has been developed to guide the botanists and researchers as well as the general public on the classification of plants (Whittaker, 1969) (see Figure 1.1). In general, plants are divided into two types botanically, which are non-seed, or spore bearing plants and the seed bearing plants. Furthermore, the larger groups are the seed plant, which is subdivided into angiosperms and gymnosperms. Angiosperm is the largest and the most common type of plants that are generally seen by the public and most abundant plants in the environment.

Traditionally, the large groups of flowering plants are divided into two groups, which are monocot and dicot. 25% of angiosperms are monocots and 75% of angiosperms are dicots. The most common examples of monocot are grasses, palms, ginger and banana whereas the common examples of dicot are trees, shrubs and herbs. Tropical rainforests are recognised as one of the most productive type of forests in the world. There are three areas in the world where tropical rainforests are found; South America, Central Africa and Southeast Asia (The Malaysian Rainforest, n.d). There are huge numbers of plants species in Malaysia, thus it is crucial for the public to know the importance and function of the plants.

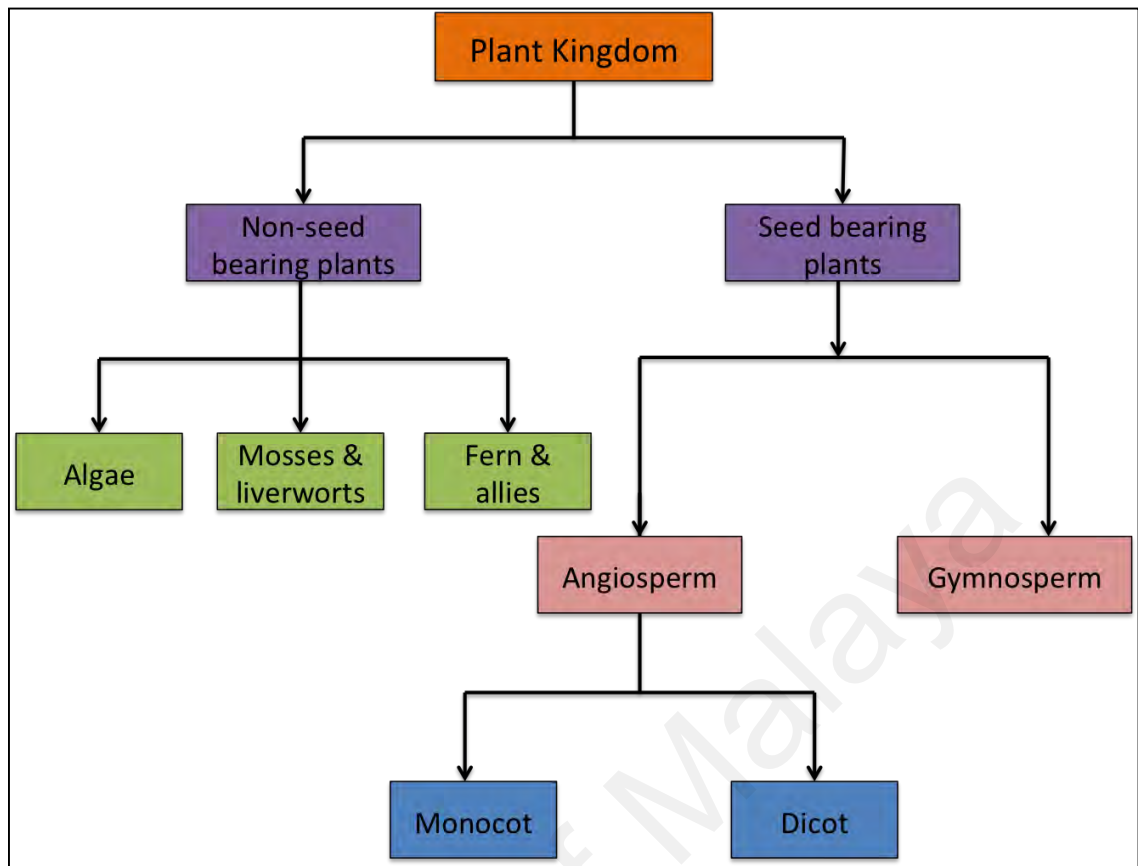


Figure 1.1: General classification of plants in the world

In recent years, image recognition and classification applications have become extremely widespread in several life sectors such as science, engineering and medicine. Computerised system through the concept of image processing and machine learning presents the ability to acquire information about the problem under study in a way that is difficult for a human to acquire.

In a manual recognition process, botanists use different plant characteristics as important parts of identification, which includes examining evenly and adaptively to identify plant species. Significantly, a botanist used identification keys for instance shapes, colours, flowers, number of petals and existence of hairs or thorns in order to answer a series of questions about one or more attributes of an unknown plant species and continuously focusing on the most significant characteristics and narrow down the set of possible

species. This list of possible species eventually leads to the desired species. However, the determination of plant species from field observation needs a substantial botanical skills and expertise.

In spite of the fact that botanist and layman can identify plant species based on botanical and biological methods, both methods are less efficient because plant species identification requires vast knowledge and in-depth training in botany and plant systematics (Wäldchen & Mäder, 2017). The recognition and classification of plant species by using traditional methods are almost impossible for layman. It is a challenging task for professional botanists as well. Professional botanists are required to take a plenty of time in the field in order to master plant species identification (Radermaker, 2000).

Besides that, an automated plant species recognition and classification is a current and popular research trend. Computer vision methods for botanical study have numerous applications, including mobile field guides using computer vision to automate or speed up the identification process, image processing for biological databases, automatic detection, automatic in agricultural field and registration and mapping of plants from publicly available data. However, there are a tremendous amount of challenges when applying the computer vision and learning algorithms for plant species recognition and classification (Wilf et al., 2016). The application of the image processing and analysis with machine learning are important for the development of intelligent systems where these approaches could be beneficial to the public.

Recently, taxonomists and botanists started searching for more efficient and effective techniques to meet species identification requirements, for instance developing digital image processing and pattern recognition methods (Agarwal et al., 2006). The rapid

development and prevalence of appropriate information technologies, for example digital cameras and portable devices, have been driving these ideas closer to reality.

Up until now, various computer vision applications have been developed and implemented. They seem to be the most popular option for the purposes of plant identification, leaf identification and classification (Wäldchen & Mäder, 2017). Plant can be recognised by looking into four aspects; leaf, flower, fruit and bark. Leaf and flower are usually the two characteristics in which plants are classified, whereby leaves are virtually two-dimensional, and flowers are three-dimensional (Viscosi & Cardini, 2011). Because of its less complex two-dimensional structure, leaves are often the preferred characteristic over flowers when determining the classification of a plant (Kellogg, 2016). Furthermore, leaves can be easily found and collected anywhere during any seasons, while flowers and fruits are only available during their respective blooming and fruit seasons (Chaki et al., 2015b). Other than flowers and leaves, bark texture can also be used in determining the plant classification. Despite being easily influenced by its surrounding environment, bark texture is more various than leaves (Lamit et al., 2015).

In this research, an image dataset for tropical shrub species, which was named as “myDAUN” dataset was developed. The samples in myDAUN dataset were sampled and collected locally from the campus of University of Malaya (UM), Kuala Lumpur, Malaysia. This dataset was used to classify tropical shrub species based on leaf shape descriptors. The classification of tropical shrub species was conducted using single and hybrid of two, three or four descriptors, which are morphological shape descriptors (MSD), histogram of oriented gradients (HOG), Hu invariant moments and Zernike moments. Three features selection method were tested on the proposed tropical shrub dataset, which were Relief, Correlation-based feature selection (CFS) and Pearson’s

correlation coefficient (PCC). Then, the selected descriptors were classified using artificial neural network (ANN), random forest (RF), support vector machine (SVM), k-nearest neighbour (k-NN), linear discriminant analysis (LDA) and directed acyclic graph multiclass least squares twin support vector machine (DAG MLSTSVM).

1.2 Problem Statement

Plants play important roles in providing us oxygen, food, medicine and fuel. Plants form a fundamental part in environmental protection (Tilman et al., 2002). However, the increasing of anthropogenic pressure on the natural environments has led many of the native plant species towards the verge of extinction (Hore et al., 1997; Mata-Montero & Carranza-Rojas, 2016) and resulting ecological crisis (Geertsema et al., 2009; Wiens, 2016). Nowadays, people have better understanding about the importance to conserve plant resources. Therefore, it is important for the general public to be able to recognise and identify plant species in order for them to contribute towards the protection of important local plant species (Corlett, 2016).

There are about 391,000 vascular plant species that are present in the world (RBG Kew, 2016) and it is difficult for any botanist or researcher to know more than a tiny fraction of the total number of known species (Fu & Chi, 2006). Plant species identification actually requires vast knowledge and in-depth training in botany and plant systematics. Even botanists take plenty of time to master plant species identification (Rademaker, 2010; Wu et al., 2007). Therefore, developing a plant species identification mechanism or an automated system that could assist the recognition process is needed (Kumar et al., 2012; Wang et al., 2008).

There is lesser emphasis and interest on botanical studies in school or university due to no interesting methods available, which makes this subject boring and less captivating. Therefore, the automated plant species classification by using machine-learning approach could be used in school or university for student's excursion. This is the advance step for inculcating interest and awareness among students in identifying plants. The ideal situation is when every student has handheld devices such as computer or smartphone, they can obtain answers directly if the system is online. Other than that, they can also save all the images and analyses them when they return back to their school or campus. This automated system is beneficial for school students, university students, as well as the general public who are interested in botanical study since it is easier to use opposed to looking through the same information from academical books.

1.3 Research Objective

The aim of this research is to classify tropical shrub species based on leaf shape descriptors and to compare different feature selection methods with various classification tools. The following objectives have been formulated in order to attain the aim of this research.

1. To extract leaf shape features from the images of selected tropical shrub species.
2. To classify tropical shrub species based on various leaf shape descriptors.
3. To identify effective machine learning algorithms for the classification of tropical shrub species.
4. To compare different feature extraction methods on the effectiveness of the tropical shrub classification.

1.4 Scope of the Study

This research focuses on the classification of tropical shrub species by using a hybrid of leaf shape and machine learning approach. In this study, the leaf images of the tropical shrub were collected and stored in myDAUN dataset. Due to time and cost limitations, only 45 species of common tropical shrubs were selected and 30 leaf samples were collected for each species in myDAUN dataset.

This research considered four shape representation techniques, namely morphological shape descriptors (MSD), histogram of oriented gradients (HOG), Hu invariant moments (Hu) and Zernike moments (ZM) for feature extraction. Besides that, three feature selection methods were applied which were Relief, Pearson's correlation coefficient (PCC) and correlation-based feature selection (CFS). Furthermore, six types of classifiers were utilised for tropical shrub species classification, namely artificial neural network (ANN), random forest (RF), support vector machine (SVM), k- nearest neighbour (k-NN), linear discriminant analysis (LDA) and direct acyclic graph multiclass least squares twin support vector machine (DAG MLSTSVM).

1.5 Research Significance

The purpose of this study is to provide an alternative approach in order to assist layman and botanist to identify plant species. Identifying and classifying plants using traditional technique is a very time-consuming task and has usually been carried out only by the experts or trained botanists. However, there are several other limitations in identifying and classifying plants using these features such as the unavailability of important and desired morphological information and use of botanical terms that only expert and botanists understand.

This study also offers expert and non-expert with valuable tools at low cost since the automated system of plant classification can be accessed without any special equipment other than a standard camera and computer processing technology. This automated system aims to significantly speed up the process of plant species identification. It only requires a photograph a leaf sample taken by user, returning the images of top matching species within seconds.

Currently, image processing and computer imaging has grown at a rapid pace, and computer architecture have become sufficiently powerful enough to solve complicated tasks in processing image data. Computer-based image processing approaches are widely implemented in solving many problems in the biological field. Apart from that, computer-aided plant classification system helps user in the process of identification, in which the user, either layman or botanists in the field can quickly search the desired plant species. The process of identification, which previously took hours, can now be completed within seconds. At the same time, this technology can increase the interest of user in studying plant. Based on the literature review, there was no similar work done on tropical shrub species dataset for plant species classification. Thus, this is the first study in development of tropical shrub species image dataset and classification using a hybrid of leaf shape and machine learning approach.

1.6 Chapter Organization

This section describes the structure and content of the thesis. The chapters of this thesis are organised as described below.

Chapter 1 provides the introduction of the proposed study including the problem statements, research objectives, and significance of the study and the scope of the research.

Chapter 2 provides the literature review of this study. It is a summary of a thorough analysis and comparison of previous studies on computer vision approaches for plant species identification.

Chapter 3 describes and explores the methodology used in this study. It gives an overview of the research and presents the experimental setup including system specification, datasets, and experimental setup. Furthermore, this chapter provides the feature extraction, feature selection and classification methods that had been implemented in this research.

Chapter 4 presents the results evaluates the performance of the various sets of the descriptors using proposed dataset and compares the results of each set of the descriptors and validates its performance with benchmark datasets.

Chapter 5 discusses the results, discussions, comparisons and validation of the proposed method.

Chapter 6 provides a summary of the research, its contribution and limitation of the proposed method and proposes some future works.

CHAPTER 2: LITERATURE REVIEW

2.1 Introduction

This chapter presents previous and current studies, which are relevant and related to the field of this study. It introduces a brief overview of the plant data sources. In addition, this chapter highlights the key techniques and algorithms used in the previous research. The fundamental concepts of plant image processing and shape approaches used in automated plant species classification are also discussed.

2.2 Plant Leaf Structure

Recognition and classification of plant species refers to one or more characteristics of a plant and linking it with a name. Commonly, human use one or more of the following characteristics, which are whole plant, bark, flower, fruits and leaves (Prasad et al., 2011). Most of the previous studies utilised leaves as one of the aspect in order to classify the plant species (Wäldchen & Mäder, 2017). Leaves are the most obvious and universal choice for tree species recognition, as they present some fundamental features and a wide pattern variation.

In botany, plant leaves are defined as a usually green, lateral structure attached to a stem, flattened and functioning as a principal organ of photosynthesis and transpiration in most plants (Gupta, 2007; Soni, 2010). Leaves contain cellular organelles chloroplasts, which contains the pigment chlorophyll that helps in making their own food. The stomata of the leaves assist in gaseous exchange, which aids in entry of atmospheric carbon dioxide during the photosynthesis process, as well as the removal of excess water in the form of water vapour during transpiration process (Raven et al., 2013). The leaves also take part

in vegetative propagation. Figure 2.1 shows the main characteristics of a leaf with its comparable botanical terms.

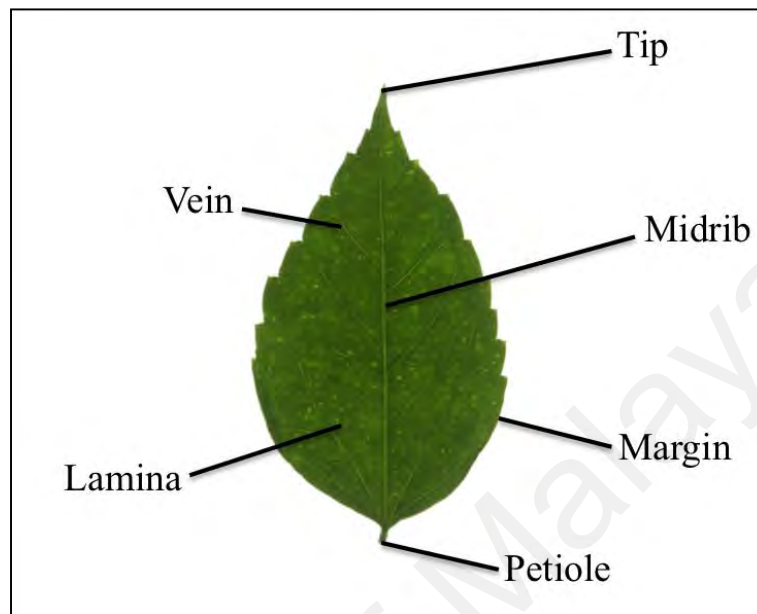


Figure 2.1: Leaf structure

Generally, a leaf consists of a lamina, which is the flat part of leaf and also supported upon a petiole, which is the transition between the stem and the leaf lamina. The structures of the leaves are different from species to species depending on their adaptation to availability of light and climate (Xu et al., 2009). It also depends on factors such as availability of nutrients, ecological competition and predating organisms (Kuzuyakov & Xu, 2013).

Based on the divisions of the blade, two basic forms of leaves can be classified, which are simple and compound leaf (Efroni et al., 2010). A simple leaf has undivided blade and the shape of the leaf are formed of lobes, but the lobes do not reach the main vein or the midrib. Whereas, the compound leaf has the leaf blade that fully subdivided and each leaflet of the blade is separated along a main or secondary vein. The middle vein of a

compound leaf is called a rachis. Figure 2.2 shows the leaf types of simple and compound leaf.

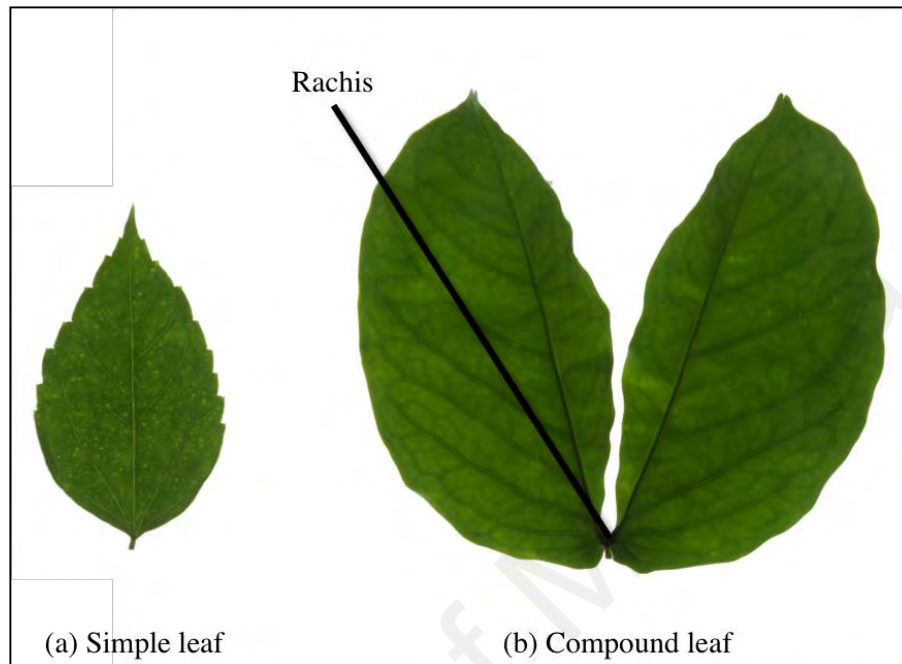


Figure 2.2: Leaf types

2.2.1 Plant Leaf Image Databases

Utilised images in the studies are divided into three categories, which are scans, pseudo-scans and photos. The majority of utilised images in the previous studies are scans and pseudo-scans in order to avoid occlusions and overlapping. The most popular and publicly available leaf image datasets are:

a) Flavia Dataset (<http://flavia.sourceforge.net/>)

The Flavia dataset was sampled in the campus of Nanjing University and Sun Yat-Sen arboretum, Nanking, China. Most of those leaves are common plants in Yangtze Delta, China (Wu et al., 2007). The leaf images were acquired by scanners or digital cameras on white background. All of the leaf samples composed of blades only without petioles. The

dataset of Flavia contains 1907 leaf images of 32 different species with 50 to 77 sample images per species (refer to Figure 2.3).

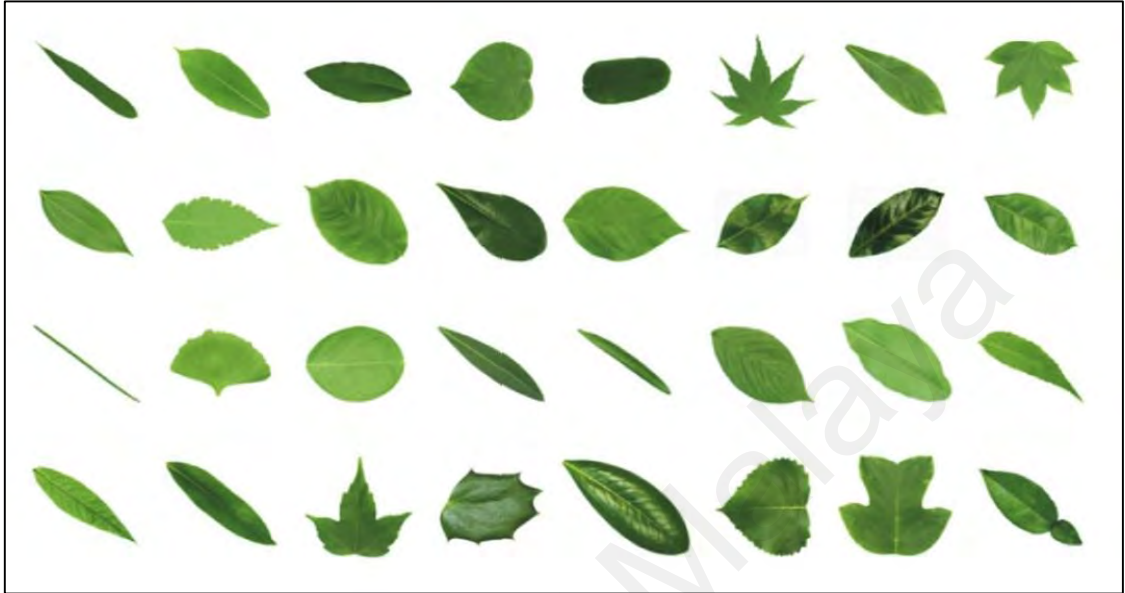


Figure 2.3: Leaf samples in Flavia dataset

b) Swedish Leaf Dataset (<http://www.cvl.isy.liu.se/en/research/datasets/swedish-leaf/>)

The Swedish Leaf dataset is a part of collaboration project between Swedish Museum of Natural History and Linköping University (Söderkvist, 2001). The dataset consists of 1125 images in total of 15 different plant species with 75 samples per species (refer to Figure 2.4). This dataset is appraised quite challenging by the reason of its high inter-species similarity. The original Swedish Leaf images contain petioles, in which the length and orientation of those petioles may heavily depends on the collection process.



Figure 2.4: Leaf samples in Swedish Leaf dataset

c) Leafsnap Dataset (<http://leafsnap.com/dataset/>)

The Leafsnap dataset currently covers 185 tree species from the Northeastern United States. This dataset consists of images of leaves that has been taken from two different sources and are accompanied by automatically generated segmentations. The first sources are high quality lab images taken of pressed leaves from Smithsonian collection with 23147 total images. Whereas, the second sources are field images taken by mobile devices in outdoor environments. These images were different in sharpness, shadows, illumination patterns and noise (refer to Figure 2.5).



Figure 2.5: Leaf samples in Leafsnap dataset

d) ICL Dataset (<http://www.intelegine.cn/English/dataset>)

The ICL dataset was sampled at the Botanical Garden of Hefei, Anhui Province of China by members of Intelligent Computing Laboratory (ICL) in Institute of Intelligent Machines, Chinese Academy of Sciences. All the leafstalks of those leaves have been cut off before the leaves were scanned and photographed on a plain background. The dataset contains 17032 plant leaf images from 220 different plant species with 26 to 1078 sample images per species (refer to Figure 2.6). All of the petioles have been cut off before the leaves were scanned or photographed on a uniform background.



Figure 2.6: Leaf samples in ICL dataset

e) ImageCLEF11 and ImageCLEF12 Dataset (<http://www.imageclef.org/>)

This dataset has been captured as part of a joined leaf identification project between Tela Botanica social network and with researchers specialised in computational botany and this dataset covering of common woody species in the Metropolitan French territory. The dataset contains 71 tree species in 2011 and further increased to 126 species in 2012 (refer to Figure 2.7). ImageCLEF11 dataset consists 5436 images subdivided into three different groups of pictures: scans (56%), scan-like photos (17%) and natural photos (27%). ImagesCLEF12 dataset contains 11572 images with scans (57%), scan-like photo (24%) and natural photos (19%).



Figure 2.7: Leaf samples in ImageCLEF dataset

Table 2.1 is a summary of the features of current existing plant datasets that were discussed in section 2.2.1. Commonly, the development of plant image database helped in developing many systems and tools to assist and support expert and non-expert in performing plant identification tasks.

Table 2.1: A summary of the features of existing leaf dataset

Features	Public leaf dataset				
	Flavia	Swedish Leaf	Leafsnap	ICL	ImageCLEF11 & ImageCLEF12
Developer	Stephen Gang Wu, Forrest Sheng Bao, Eric You Xu, Yu-Xuan Wang, Yi-Fan Chang and Qiao-Liang Xiang	Collaboration of Linkoping University and the Swedish Museum of Natural History	Researchers from Columbia University, the University of Maryland, and the Smithsonian Institution	Researchers from Intelligent Computing Laboratory (ICL) at the Institute of Intelligent Machines, China	Citizen sciences initiative conducted by Telabotanica, a French social network of amateur and expert botanists
Sampling area	Yangtze Delta, China	Linkoping University, Sweden	Northeastern United States	Hefei Botanical Garden	French Mediterranean
Type of plant species	Common plant of Yangtze Delta	Swedish tree species	Common tree of Northeastern United States	Common tree species of the Chinese Anhui	Common tree species of French Mediterranean area
Total species	32	15	185	220	126
Total image samples	1907	1125	30866	17032	11572
Samples per species	50 to 77	75	56 to 448	26 to 1078	-

2.3 Feature Extraction

Feature extraction is a fundamental part of the content-based image classification and usually follows after the image segmentation steps (Thepade et al., 2014). A digital image is simply a collection of pixels represented as large matrices of integers that relates to the intensities of colours at different positions in the image (Gonzalez & Woods, 2010). The aim of the feature extraction is to reduce the dimensionality of this information by extracting characteristics features. These features can be found in shape, colour, texture and specific organ of the leaf. On the other hand, most of the previous studies highlight that shape plays an important role, even though the shape of leaves presents a wide pattern variation. In the following sections, an overview of the main features and the descriptors proposed for automated plant species classification (see Figure 2.8) are discussed.

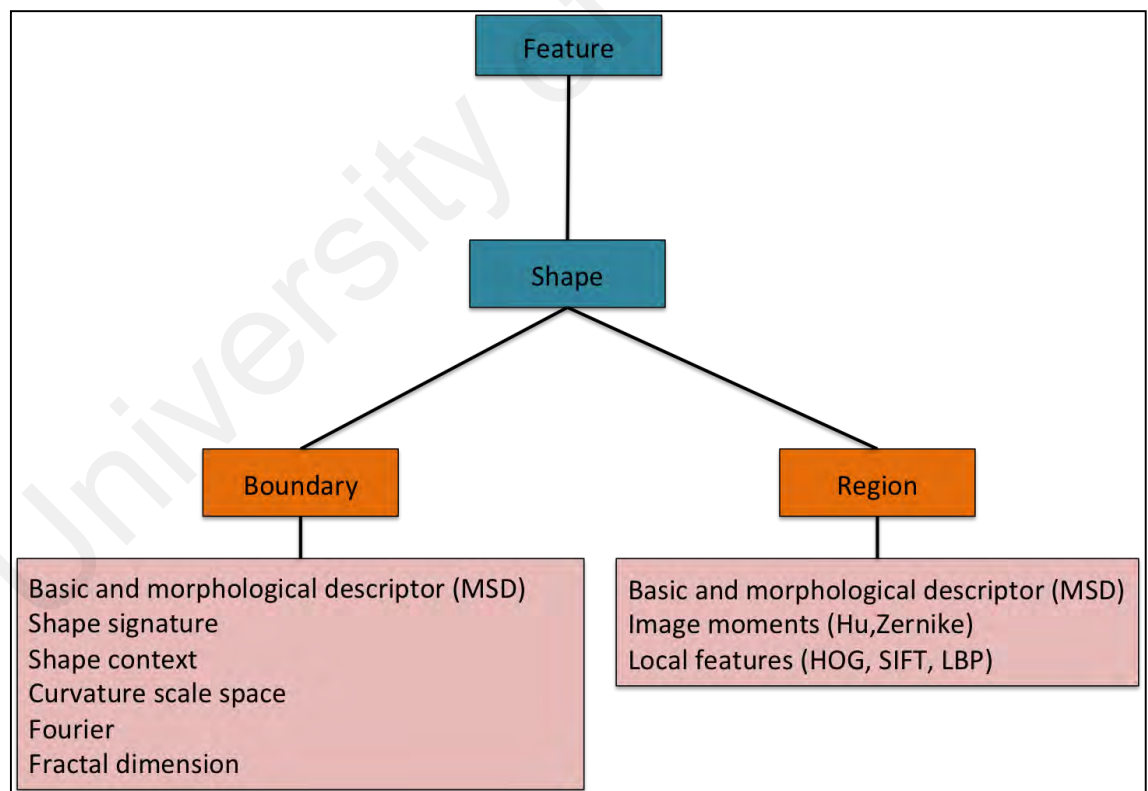


Figure 2.8: Categorization and overview of the most prominent shape feature descriptors in plant species identification

2.3.1 Shape

Shape is known as fundamental aspect for human when identifying objects. A shape measure is typically a quantity, which corresponding to a particular shape characteristic of an object. The relevant shape descriptor should be invariant to geometrical transformations; rotation, reflection, scaling and translation. Shape descriptors are classified into two categories; boundary based and region based. Region based shape descriptors obtain the shape features from the whole region of the shape (Kadir et al., 2011a; Zhang & Lu, 2004). On the other hand, boundary based shape descriptors extract shape features merely from the contour of the shape. Besides, there are also some methods, which cannot be classified as either region-based or contour-based. Since the majority of the primary and previous studies had focused on plant identification using leaves, the discussion in this study will focus on the feature extraction using leaf shape.

2.3.2 Morphological Shape Descriptors (MSD)

Across the studies, there are five basic shape descriptors generally used for leaf analysis. These are specific to basic geometric properties of the leaf's shape, which are diameter, major axis length, minor axis length, area and perimeter (Wu et al., 2007). Based on that, studies computed and applied morphological descriptors based on these basic descriptors such as aspect ratio, form factor, rectangularity and perimeter to area ratio. Ratios are uncomplicated which is simple to compute and naturally invariant to translation, rotation and scaling. These make them more robust against different representations of the same leaf. Moreover, some studies proposed more leaf specific descriptors.

Leaf Width Factor (LWF) is computed when the leaf is sliced perpendicular to the major axis, into a number of vertical strips (Hossain & Amin, 2010). Then the LWF per strip is calculated as a ratio of the width of the strip to the length of the entire leaf (major axis length). Yanikoglu et al. (2014) introduced an area width factor (AWF), which computed

the area of each strip normalised by the global area. In addition, Prasad et al. (2013) proposed a porosity feature in order to highlight the cracks in the leaf image.

Previous studies have showed that morphological shape and simple descriptors (MSD) are too simplified to discriminate the leaves apart from those with large differences. Therefore, most of the studied combined MSD with other descriptors with more complex shape analysis. The uncertainty of MSD is that any attempt to describe the shape of a leaf using only five to twelve descriptors may reduce matters to the extent that essential analysis becomes impossible, even if it look sufficient to classify small set of images. In addition, most of single value descriptors are highly correlated with each other, which makes the task of choosing enough independent features to differentiate categories of interest especially though (Cope et al., 2012).

2.3.3 Region-based Shape Descriptor

Region-based method is a technique that takes all pixels within a shape region to obtain the shape representation and the pixels of same type identified and grouped together into same type of region. The main purpose of region-based method is to partition an image into different or same types of regions. Generally, region-based descriptors for plant species identification is divided into two categories; image moments and local feature techniques.

a) Image Moment

Image moments are generally applied category of descriptors in object classification. Roughly, image moments are statistical descriptors that are invariant to translation, rotation and scale. Hu invariant moments (Hu) proposed seven invariant moments computed from central moments through order up to three and two-dimensional. Hu are

a widely applied in computer vision research (Wäldchen & Mäder, 2017). Hu known as geometric moments are computationally simple but exceedingly sensitive to noise.

Based on the previous studies, the combination of hybrid descriptors of Hu with the MSD for leaf classification analysis was used (Chaki et al., 2015a; Du et al., 2007; Pauwels et al., 2009; Wang et al., 2005; Zhang et al., 2008). Apriyanti et al. (2013) used Maximal Similarity based on Region Merging (MSRM) method for segmenting and extracting the shape feature such as centroid point, aspect ratio, roundness, Hu invariant moments, fractal dimension and colour feature. Whereas, Du et al. (2007) combined geometric features with Hu, Zernike moments (ZM) and Polar Fourier Transform (PFT) in order to identify the plants. Whereas, Kalyoncu and Toygar (2015) proposed a set of features to describe a leaf and the feature extraction by applied Hu, convexity, perimeter ratio, multi scale distance matrix, average margin distance and margin statistics.

In addition, Kadir et al. (2011), Wang et al. (2008) and Zulkifli et al. (2011) studied leaf analysis and evolved Zernike moments and Legendre moment. Both moments are also invariant to arbitrary rotation of the object but they are not sensitive to image noise. In spite of that their computational complexity is very high. Kadir et al. (2011) found that classification using Zernike moments did not produce better accuracy than Hu invariant moments. Besides, three moment invariants methods were compared which are Zernike moment, Legendre moment and Tchebichef moment invariant in order to determine the most effective technique in extracting features from leaf images (Zulkifli et al., 2011). As a result, Tchebichef moment invariant is the most effective descriptors among others moment invariants. In Novotny and Suk (2013), they found that Tchebichef moment invariant produced the best results compared with Hu invariant moments and Zernike moments. However, Tchebichef moment is a time consuming process and the

computational complexity increased by increasing the moment order (Wang & Wang, 2006).

b) Local Features

Generally, local features are defined as the selection of scale-invariant interest points in an image and their extraction into local descriptors per point. The obtained interest point can be compared with another image. For example, local feature approach is the histogram of oriented gradient (HOG) descriptor (Du & Wang, 2001; Lowe, 2004; Pham et al., 2013; Xiao et al., 2010; Zhang & Feng, 2010). Lowe (2004) introduced HOG descriptor and used in image processing for object detection and it is the local statistic of the orientations of the image around key points. HOG descriptor method determines occurrences of gradient orientation in localised portions of an image or ROI.

HOG descriptor is similar to Scale-invariant feature transform (SIFT) descriptors but it uses overlapping local contrast normalization across neighbouring cells grouped into a block. As HOG computes histograms of all image cells, it contains much redundant information that reduces the dimensionality necessarily for further extraction of discriminant features. Pham et al. (2013) compared Hu invariant moments with HOG features and the achieved result showed that HOG gave better result than Hu invariant moments for species identification.

Lowe (2004) introduced the SIFT approach and combine a feature extractor and detector. SIFT algorithm are invariant to image scale and rotation in term of feature detected and extracted. This algorithm is suitable for object recognition rather than object comparison because of the invariance and robustness of the features extracted. There are also many research groups working on an automated identification for identification of plant species

using HOG descriptors (Hsiao et al., 2014; Hussain et al., 2013; Lavania & Matey, 2014; Priyankara & Withanage, 2015). Priyankara and Withanage (2015) described a leaf image based plant identification system using SIFT features combining with the Bag of Words (BoW) model. The BoW model reduced the high dimensionality of the space data.

Hussain et al. (2013) presented a method of shape feature extraction that is Scale Invariant Feature Transform (SIFT) and colour feature extraction Grid Based Colour Moment (GBCM) to identify plant. Hsiao et al. (2014) applied SIFT with sparse representation and correlated their results with BoW model. In order to improve the leaf image classification, there are a few studies such as Wang et al. (2011) and Kebapci et al. (2011) combined both local and global shape with SIFT descriptors. Larese et al. (2014) presented that the accuracy by using SIFT method is significantly lower when compared with combination of SIFT and global shape features. One of the common issues in leaf analysis using SIFT is often a lack of characteristic key points since the leaves are not in uniform texture.

2.3.4 Boundary-based Shape Descriptors

Boundary based shape descriptors merely consider the contour of the shape and disregard the information contained in the shape interior. A boundary-based descriptor is a sequence values calculated at points from the object's outline.

a) Shape Signatures

Shape signatures typically use boundary-based shape descriptors, which performed a shape by a one-dimensional function derived from shape boundary point. There are varieties of shape signatures that had been studied for leaf analysis, for example centroid contour distance (CCD) (Beghin et al., 2010; Chen et al., 2011; Fotopoulou et al., 2013;

Teng et al., 2009), triangle area representation (TAR), the triangle side length representation (TSL), triangle oriented angles (TOA) and triangle side lengths and angle representation (TSLA) (Mouine et al., 2013a).

The CCD descriptor contains a sequence of distances between point of the contour of a shape and centre of the shape. On the other hand, centroid-angle (AC) and the tangential angle (AT) are the example of descriptors that consists of a sequence of angles to represent the shape. Fotopoulou et al. (2013) compared CCD and AC sequences and the result showed that CCD sequences are more informative than AC sequences. The reason is CCD included both location information of contour details and global information of the leaf area and shape. Hence, combining CCD and AC is expected to increase the classification accuracy.

As stated in Mouine et al. (2013a), two multi scale triangular approaches for leaf shape description which are triangle area representation (TAR) and the triangle side length representation (TSL) has been proposed. TAR descriptor is computed based on the area of triangles formed by points on the shape contour, whereas TSL descriptor is computed based on the side lengths. Mouine et al. (2013a) introduced triangle oriented angles (TOA) and triangle side lengths and angle representation (TSLA). TOA usually uses angle values to represent triangle and TSLA is multi-scale triangular contour descriptor that represent the triangles by their lengths and angle. The limitation of the shape signatures for leaf analysis is the high matching cost and sensitive to noise and changes in the contour. Therefore, it is unenviable to describe a shape using a shape signature directly.

b) Shape Context

Belongie et al. (2002) proposed shape context (SC) descriptor that represents log polar histograms of contour distribution. Each contour point is described by a histogram in the context of entire shape. Hu et al. (2012) presented a contour based shape descriptor, called multi-scale distance matrix (MDM) in order to capture the geometric structure of shape and in the same time invariant to translation, rotation, scaling, and bilateral symmetry. MDM provides the most effective technique because it avoids the use of dynamic programming for building the point-wise. By comparing SC with MDM, it showed that MDM achieved comparable result of recognition and more computationally efficient (Hu et al., 2012).

In spite of MDM being effective in describing the broad shape of leaf, it fails in capturing information for example leaf margin. Thus, a combination MDM with average margin distance (AMD), margin statistics (MS), MSD and Hu had been proposed by Kalyonchu and Toygar (2015). The result of classification achieved higher accuracy by using combination of MDM, MSD and Hu.

c) Scale Space Analysis

Florindo et al. (2010) proposed an approach to classification of leaf shape using curvature scale space (CSS). CSS piles up curvature measures at each point of the contour (Zhang & Lu, 2004). A curve describing the complexity of the shape can be used as descriptor. In addition, studies found that CSS is an effective descriptor but too informative because the implementation and matching of CSS is very complex. Besides, curvature provides a compact description of curvature optima and is able to detect point of interest. Caballero and Aranda (2010) and Cerruti et al. (2013) found that the stand out points based on graph curvature values of the contour as descriptor. Whereby, Chen et al. (2011) used a

simplified curvature of the leaf contour, named velocity and the result showed that the velocity algorithms were more justifiable and faster at finding contour shape characteristics than CSS.

Lavania and Matey (2014) presented the contour as a chain code, which extract high curvature points on the contour and enumerated direction codes. Kumar et al. (2012) as well suggested a method, called histogram of curvature over scale (HoCS). The HoCS method is implemented from CSS and it apparently creates histograms of curvature values with different values of scales. However, the drawback of the HoCS method is not articulation invariant. This is because the blade and petiole of the simple leaf or the leaflet of the compound leaves can cause significant changes in calculation of the HoCS descriptor. Therefore, the authors suggested to detect and removed the petiole before classification.

d) Fourier Descriptor

Fourier descriptor is the simple method for shape identification and a general method to encode various shape signatures. A leaf can be analysed in the frequency domain instead of spatial domain. A set of Fourier descriptors are calculated for the outline of the object and the global shape features in the low and high frequency terms will be captured. As stated in Cope et al. (2012), the dominance of this method is easy to implement and it is based on the well-known theory of Fourier analysis. Moreover, Fourier descriptors can be easily normalised to represent the shape independently so it is easy to compare between shapes. Yet, the limitation of the Fourier descriptors is that they do not present the information of the local shape because this information is distributed across all coefficients after the transformation (Zhang & Lu, 2004).

There is one study that uses Fourier descriptors to compute the distances of the contour point from the centroid and this method works well for smaller datasets. Furthermore, Kadir et al. (2012) proposed a method known as Polar Fourier transform (PFT) that extracts the shape of the leaves and compared it with other methods which are MSD, Hu invariant moments and Zernike moments. PFT showed that an eventual result of the classification (Kadir et al., 2012). Most previous studies used FD in combination with MSD and the result of combining all descriptors showed prospective classification result (Aakif & Khan, 2015; Yanikaglo et al., 2014). In order to improve the effectiveness of classification result, there are a few studies such as Florindo et al., (2010), Hu et al. (2012), Wang et al. (2013), Yang and Wang (2012) and Zhao et al. (2015) that proposed novel methods for leaf analysis, benchmarking their descriptor against Fourier descriptor.

e) Fractal Dimension

Fractal dimension is used to perform a shape filling the dimensional space to which it belongs and contributes a useful measure of leaf shape's complexity. There are a few studies that used fractal dimension for leaf analysis (Bruno et al., 2008; Du et al., 2013; Jobin et al., 2012). Bruno et al. (2008) compared fractal dimension with two methods which are box-counting and multi-scale Minkowski. Multi-scale Minkowski method shows better results in terms of characterizing plant species, whereas box-counting method just provides a good result.

Due to the wide variety of leaf shapes, the fractal dimensional descriptor may only be useful in combination with other descriptors because fractal dimension characterised the leaf shape by the single value descriptor of complexity. Du et al. (2013) presented the leaf analysis with fractal dimension and Hu invariant moments and the classification accuracy showed that fractal dimension achieved significant result than Hu invariant moments.

However, the result of combining both Hu invariant moments and fractal dimensional achieved a better result.

2.4 Feature Selection Methodologies

Feature selection is used to select the inputs, which are most significant and meaningful in the modelling process, in order to obtain more accurate outputs. The objective of the feature selection is to reduce the number of inputs in the modelling process, while still retaining the accuracy of the outputs if compared to the full input model. Thus, this can have a better predictive and less computationally intensive model.

Feature selection can be classified into three main groups, which are filter, wrapper and embedded methods. The filter methods rank the variable and select the variables with highest criteria. The examples of the filter selection methods are Pearson's correlation coefficient (PCC), Relief, independent component analysis (ICA) and linear discriminant analysis (LDA) method. Whereas, wrapper methods is used to evaluate the variables in subsets and use the heuristic search methods for the optimal subset. One example of the wrapper approach is genetic algorithm (GA). Next, the embedded method learns the best features that contribute to the accuracy of the model while the model is being created. The embedded method is built into classifier to search for a subset and it is specific to the learning algorithm (Saeys et al., 2007; Song et al., 2005).

2.4.1 Relief

Kira and Rendall (1992) formulated the Relief algorithm. Relief algorithm was inspired by instance-based learning (Aha et al., 1991) and as an individual evaluation filtering feature selection method. Relief calculates a proxy statistic for each feature that can be applied to estimated feature quality or relevance to the target concept, which predicts endpoint value. These feature statistics are referred to as feature weights and also known

as feature scores that can range from worst to best. The original Relief algorithm is rarely applied in practice anymore and has been supplanted by ReliefF as the best known and more utilised Relief-based algorithm to date (Kononenko, 1996). The ReliefF algorithm has been detailed in a several other studies (Kononenko et al., 1996; Robnik-Sikonja & Kononenko, 1997).

The Relief algorithm that has been proposed by Kononenko (1995) relies on a number of neighbours user parameter k that specifies the use of k nearest hits and k nearest misses in the scoring update for each target instance. In order to solve the scoring in multi class, Relief finds k nearest misses for each other class and the averages of the weight update based on the prior probability of each class. Generally, this algorithm is able to estimate the ability of the features to separate all pairs of classes regardless of which two classes are closest to one another. The process is repeated for m times.

2.4.2 Correlation-based Feature Selection (CFS)

The Correlation-based feature selection (CFS) algorithm relies on a heuristic for calculating the worth of a subset of features. This heuristic finds the usefulness of individual features for predicting the class label. The hypotheses on which the heuristic is based can be stated (Hall, 1999):

“Good feature subsets contain features highly correlated with (predictive of) the class, yet uncorrelated with (not predictive of) each other” (Hall, 1999)

Gennari et al. (1989) also stated “Features are relevant if their values vary systematically with category membership.” Which means, a feature is useful if it is correlated with or predictive of the class; otherwise it is irrelevant. Empirical information from the feature

selection literature found that the irrelevant features and the redundant information should be eliminated as well (Kononenko, 1995, 1996).

A feature is considered redundant if one or more of the other features are highly correlated with it. The above definitions for relevance and redundancy show the idea of the best features for classification are those that are highly correlated with one of the classes and have a minimal correlation with the rest of the features in the set. The correlation between a composite consisting of the summed components and the outside variable can be predicted if the correlation between each of the components is known, and the inter-correlation between each pair of components is disposed (Borcherding, 1977; Ghiselli, 1973; Zajonc, 1962).

2.4.3 Pearson's Correlation Coefficient (PCC)

The Pearson's correlation coefficient (PCC) was introduced by Pearson (1920). The concept of the PCC is to measure the linear correlation between two random variables X and Y. The range values of PCC are from -1 to 1 and the value of 0 indicates no linear correlation between X and Y. Besides, the value -1 indicates a total negative correlation, whereas +1 indicates a total positive correlation between X and Y. It can be defined on real values variables. The major drawback of the PCC is that it can only detect linear correlations. PCC is a parametric measure, as it assumes that distribution of each attribute can be described using a Gaussian distribution (Rosner, 2006).

2.4.4 Genetic Algorithm (GA)

Genetic algorithm (GA) is a search algorithm method that classifies a given dataset based on natural selection and genetics in biological systems. It can be used to solve different and diverse types of problems. The algorithm starts with a population of group of individuals called chromosomes. Each of the chromosomes is evaluated using a fitness

function. The process is iterated for multiple times for a number of generations until a termination criterion is reached. The reached termination criterion could be a single individual or a group of individuals obtained by repeating the GA process (Li, 2004).

2.5 Image Classification Methodologies

The digital images are electronic snapshots taken of a scene or scanned from document such as photographs, manuscripts, printed texts and artwork. A digital image is made of picture elements known as pixels. Usually, pixels are organised in an ordered rectangular array and typically the size of an image is determined by the dimensional of this pixel array. Each pixel is assigned a black and white, grey shade or RGB colour. Classification is one of the most significantly used techniques in machine learning, including medical diagnosis, spam detection, risk assessment and image classification. The basis goal of classification is to predict a category or class y from some input x . The image classification is an important task in various fields such as biometry, remote sensing and biomedical images (Kamavisdar et al., 2013). Supervised classification can be simplified as first of all, training took place through known group of pixels. Then, the trained classifier is used to classify other images. Whereas, the unsupervised classification uses the properties of the pixels to group them, and these groups are called cluster, and the process that took place is known as clustering.

An algorithm that implements classification, especially in implementation, is known as a classifier. Different image classification methods have their advantages and some disadvantages. There are some methods that used the combination of another classifier in image classification. A classifier is considered efficient if they can predict precisely and correctly. Hence, classifier is important to extract the pattern or feature from the available

input dataset. The following algorithms have been used to classify species by several authors, based on leaf images.

2.5.1 Artificial Neural Network (ANN)

In order to simplify the tasks of prediction of classification, neural networks are being introduced. Neural networks are simplified and straightforward models of biological neuron system. It consists of great many parallel-processing units, hence it is able to execute many different parts of a program at the same time. There are various existent learning mechanisms that enable the neural network to acquire knowledge. The neural network architectures have been classified into many types based on their learning mechanisms and other features.

Neural network are simplified to represent the central nervous system (Rajasekaran & Pai, 2003), and thus, have been inspired by the kind of computing performed by the human brain. The word network in ANN represents the interconnection between neurons present in various layers of a system. Basically, every system has three-layered systems, which are input layer, hidden layer and output layer (see Figure 2.9). The input layer has input neurons that transfer data through synapses to the hidden layer and the hidden layer transfers the data to the output layer through more synapses. The synapse stores values known as weights that help to manipulate the input and the output to various layers. An ANN can be performed based on the following characteristics:

- The number of layers and the number of the nodes in each of the layers.
- The applied learning mechanism is used for updating the weights of the connections.
- The activation functions applied in various layers

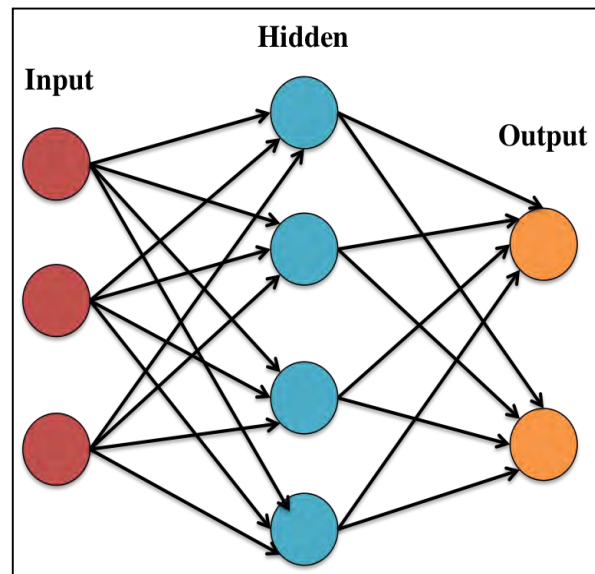


Figure 2.9: Model of artificial neural network (ANN)

2.5.2 Random Forest (RF)

A random forest (RF) is a successful ensemble of a prediction technique that capitalised on many decision trees and prudent randomization to generate accurate predictive models. RF is introduced by Breiman (2001), which showed that replacing a single tree by an ensemble of decorrelated trees provides very good generalisation. Breiman (2001) defined that a RF is a classifier consisting of a collection of tree-structured classifiers $\{(h(x), k = 1, \dots)\}$ where $\{x\}$ are independent identically distributed random vectors and each tree casts a unit vote for the most popular class at input x . In other words, a RF is built to comprise the task of generating random vector to grow an ensemble trees and letting those vote for the most popular class.

RF are trained through the bagging method, where the bagging or bootstrap aggregating method consists of randomly sampling subsets of the training data, fit a model with smaller dataset and aggregating the predictions. The tree bagging includes of sampling subsets of the training set, fitting each decision tree and aggregating the result. Instead of searching greedily for the best base to create branches, the elements of the base space are

randomly sampled. This process is named as feature bagging and it is the powerful method that leads to a more robust model.

RF is the most popular ensemble algorithms that uses decision tree as base classifier. The construction of a RF consists of three main following phases:

(i) Gaining ensemble diversity

RF algorithm attains ensemble diversity by manipulating training sets. A list of learning sets is produced using the bootstrap sampling method.

(ii) Construct base classifiers

RF applied random tree on different training sets generated in the previous step to create base classifiers. Each node, which a small group of input attributes, is selected randomly. The group size is decided by users, but commonly it is chosen as the greatest integer than the number of input attributes. Then, the best attributes or the split point would be selected to split on.

(iii) Combining base classifiers

The greater voting method is employed in the RF algorithm.

2.5.3 Support Vector Machine (SVM)

Support vector machine (SVM) are supervised learning method used for regression problems or classification of samples into two or more classes or group. SVM is a technique of classification with an output resembling the neural network. They have been used to encounter the classification problems in multispectral images (Mitra et al., 2004), gene selection in cancerous tissues (Guyon et al., 2002), and handwritten digits (Cortes & Vapnik, 1995). SVM was introduced by Boser et al. (1992), but the high level mathematics to support its operation and this theory can be referred back to literature concerning hyperplane decision boundaries (Vapnik & Chervobekis, 1968). SVM

performs the classification by constructing an N-dimensional hyperplane that optimally separates the data into two classes. An example of a two dimensional hyperplane is shown below in Figure 2.10.

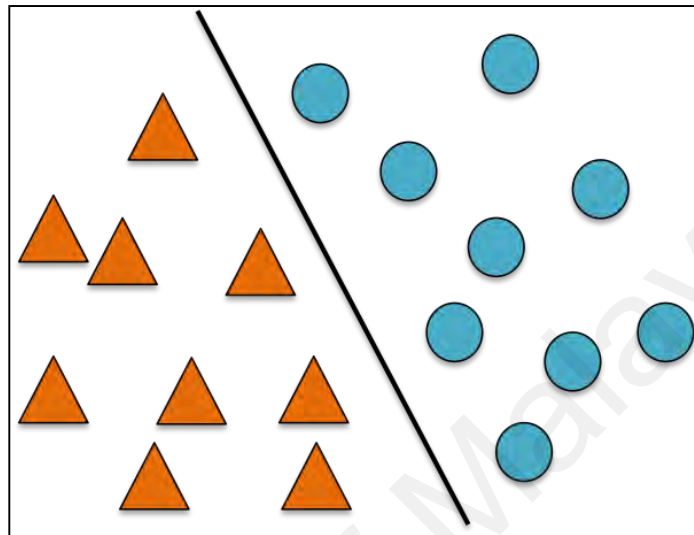


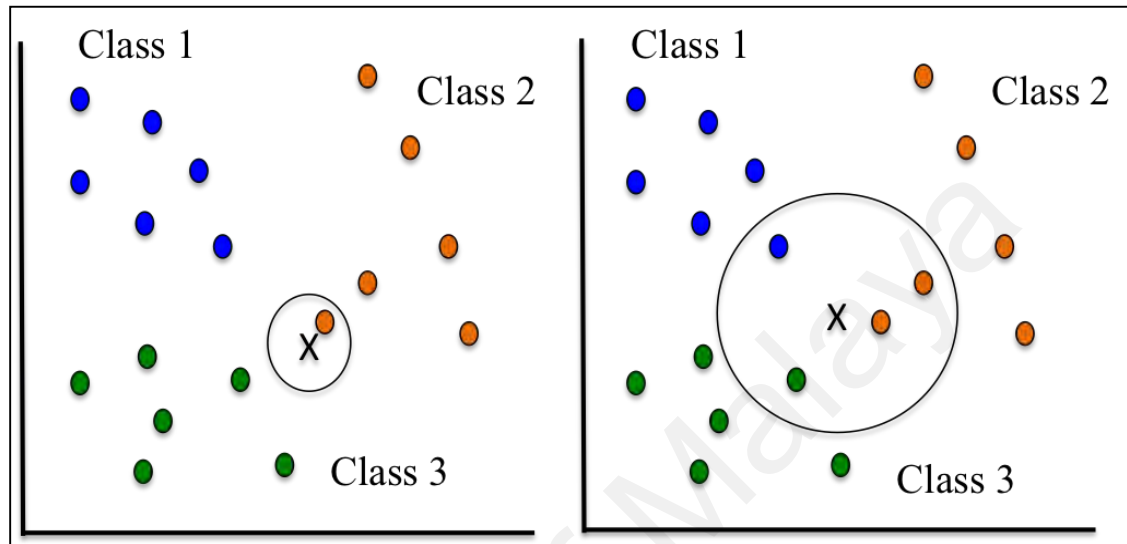
Figure 2.10: Model of a two dimensional hyperplane

2.5.4 k- Nearest Neighbour (k-NN)

The k nearest neighbour classifier is comparable to the minimum classifier. It is a supervised classification-learning algorithm used to classify sample. The purpose of k-NN is to classify a new sample based on its features and labelled training samples. The k-NN algorithm is memory-based and does not require a model to be fit. It considers the k-nearest points to an unknown tuple and assigns the tuple to the majority of its neighbours instead of considering the means of each class.

The accuracy of a classifier is influenced greatly by choosing the correct value for k greatly. If the value of k is too large, it will encompass all of the training data and assign the tuple to the class with more training examples. On the other hand, if the value of k is too small, it will create a problem of classifying the tuple as part of the wrong class. Thus, k-NN classifier takes only k-NN classes, which majority vote is then taken to predict the

best-fit class point (Silva, 2013). For example, consider Figure 2.11 (a) where $k=1$ and Figure 2.6 (b) where $k=4$. The point X presents the best-fit class according to majority votes of the nearest point.



(a) 1-NN classifier

(b) 4-NN classifier

Figure 2.11: Model of k-nearest neighbor (k-NN) classification.

2.5.5 Linear Discriminant Analysis (LDA)

Linear discriminant analysis (LDA) is a supervised statistical method to classify samples into two classes or group based on the features that describe the samples. LDA builds a linear classifier based on the features of the samples in the dataset. LDA technique considers that the classes or group having a common covariance matrix. The function of LDA is to measure each sample in each class is cross-validated with the corresponding class and the accuracy of the classification is obtained.

LDA is a feature mapping method that used both of dimensional reduction and classification. LDA was implemented as a feature mapping method in order to transfer the original data into a new space where different classes can be divided linearly by

finding a decision region between that given classes in the new map spaced (Mohammadi et al., 2011).

LDA faces challenges in cases of high dimensional data, where the LDA matrices are almost always singular (Yu & Jie, 2001). Figure 2.12 shows a two dimensional dataset before and after implementing LDA. The features are mapped into a new feature space by using LDA, which is more linearly discriminant compared to the original feature.

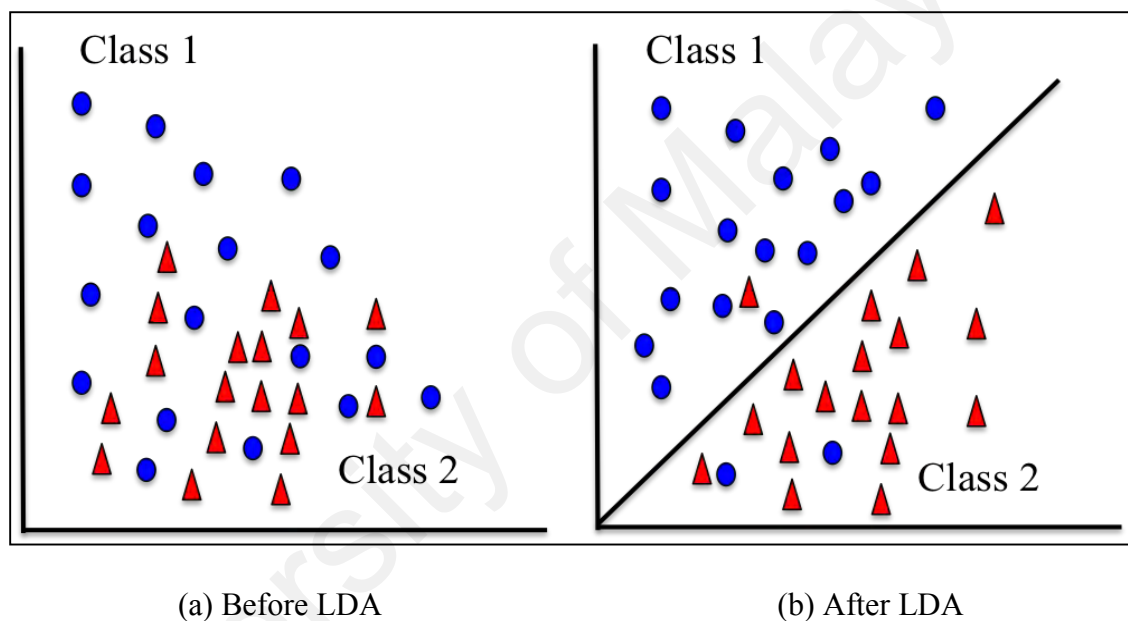


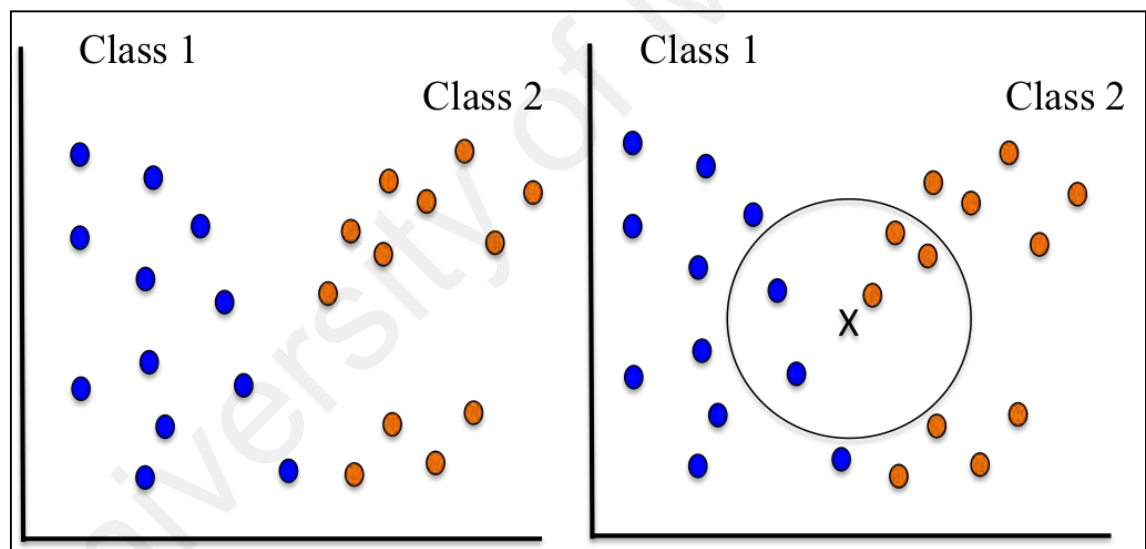
Figure 2.12: Model of linear discriminant analysis (LDA) classification.

2.5.6 Naïve Bayes

A simple Bayesian classifier known as Naïve Bayes classifier is based on Bayes's theorem with the independence assumptions between predictors. The Naïve Bayes is particularly suited when the dimensionality of the input is high and it is comparable to decision trees in terms of performance (Han & Kamber, 2006). This classifier required

two parameters which are a priori probability and class conditional density to do the classification.

The Bayes's classifier tries to estimate the class based on the control conditions (Bandyopadhyay & Pal, 2007). The probability of the class labels can be estimated from the training data, but the distribution of attribute values of the given class are more complex to be estimated. Figure 2.13 shows the Naïve Bayes classification concept. As indicated, the objects can be classified into two groups. From the illustration below, it is clear that the likelihood of X given class 2 is smaller than the likelihood of X given orange, since the circle encompasses one blue object and 3 orange objects.



(a) Before Naïve Bayes

(b) After Naïve Bayes

Figure 2.13: Model of Naïve Bayes classification.

2.6 Previous Studies in Plant Species Classification

There are different experimental methods in the previous and current studies in terms of dataset, features and classifier, which make it very challenging to compare results and the proposed approaches themselves. In this section, primary studies that used boundary

based, region based or the combination of these two shape descriptors techniques were selected and a comparison of their results was performed.

2.6.1 Boundary-based Shape Descriptor

There are six major studies, which used boundary based shape descriptors (Bong et al., 2013; Fu et al., 2004; Mouine et al., 2013b; Ren et al., 2012; Wang et al., 2015b; Xiao et al. 2010). Fu et al. (2004) and Bong et al. (2013) performed centroid contour gradient (CCG) that calculates the gradient between pairs of boundary points corresponding to interval angle. After utilising feed forward back propagation as classifier the accuracy of CCG and CCD are 96.6% and 74.4% respectively in Bong et al. (2013). Different features were also tested for the classification system, which are Fourier coefficient, moment invariant and chain code. The accuracy of CCD, Fourier coefficient, moment invariant and chain code by using feed forward neural network are 94.26%, 85.71%, 69.52% and 42.86% respectively in Fu et al. (2004).

Furthermore, Mouine et al. (2013b) determined that TLSA are outperformed in classification accuracy result compared to TAR, TOA and TSL. Next, inner-distance shape context (IDSC) descriptor stands out than HOG in Xiao et al. (2010). Multi scale overlapped block local binary pattern (LBP) with SVM classifier are used by Ren et al. (2012) and obtained accuracy of 93.73%. Wang et al. (2015b) introduced multi- scale arch height descriptor (MARCH), which follows desirable properties of invariance, compactness, low computational complexity and coarse to fine representation structure. The performance of the proposed method has been evaluated on four leaf datasets, which were the Swedish Leaf dataset, the Flavia dataset, the ICL dataset and the ImageCLEF dataset.

2.6.2 Region-based Shape Descriptor

Across the studies, the region-based shape descriptors were commonly used for leaf analysis. Kulkarni et al. (2013) presented a system for recognizing and identifying plants using shape, vein, colour, textures features which are combined with Zernike moments. Radial basis probabilistic Neural Network (RBPNN) has been used as a classifier. The accuracy for this combined method has achieved the highest accuracy of 93.82%. The techniques include segmentation, a combination of general feature extraction and classification methods in order to classify plant leaves was proposed by Kalyoncu and Toygar (2015). Linear Discriminant Classifier (LDC) is used as classifier; hence using the features that are noisy for some leaf types does not affect the performance of the system. The experimental results give accuracy more than 90% when using Flavia dataset and 71% when using Leafsnap dataset.

The SIFT and contour based edge detection approach for plant recognition were presented in (Lavania & Matey, 2014). The system was able to classify plant species in the Flavia dataset, gives a correct recognition accuracy of 87.5%. Whereas, Bhardwaj et al. (2013) compared the classification result of SVM with RBF kernel and k-NN based on shape and vein features and the SVM result outperformed the k-NN, which obtained 94.5% in SVM and 78% in k-NN. In addition, Caglayan et al. (2013) applied four types of classifier, which are k-NN, SVM, Naïve Bayes and Random forest based on shape and colour features. The result shows that random forest obtained the outperformed result than others with result accuracy of 93.95%.

Hossain and Amin (2010) and Wu et al. (2007) proposed Probabilistic neural network (PNN) for classification of leaf shape features and achieved an accuracy of 90.31% and 91.40% respectively. Multi-layer perceptron using back propagation (MLP) and neuro-

fuzzy classifier using a scaled conjugate gradient algorithm (NFC) were applied in (Chaki et al., 2015a; 2015b). The accuracy achieved by merely using texture-based descriptors are 81.6% with NFC and 87.1% using MLP. Whereas, the accuracy achieved by using shape-based descriptors are 50.16% with NFC and 41.6% using MLP. Chaki et al. (2015a) found that the combination of texture and shape obtained the best results, which is 97.6% with NFC and 85.6% with MLP.

Du et al. (2007) approached a leaf database and each species includes 20 sample images. The digital morphology feature extraction and moment feature (MF) are implemented. The move median centre (MMC) hypersphere was chosen as classifier and the performance of the MMC classifier were also compared with the nearest neighbour (1-NN) and kNN classifier. The accuracy of MMC, 1-NN and kNN are 91%, 93% and 92% respectively. (Bhardwaj et al., 2013) presented the automated system for plant identification using shape features with four parameters those are area convexity, volume fraction, moment invariants and inverse different moment. The database contains various shape, colours and size. There are 320 leaves of different 14 plants taken, which are totally different in their shape and colour. Within 320 leaves, 293 were classified and 27 misclassified then a recognition accuracy of 91.5% was achieved in k-NN.

Arora et al. (2012) studied in identifying the plant of identification system using shape and morphological features on segmented leaflets. The dataset used in this system is ImageCLEAF Pl@ntLeaves dataset and the feature vector can be divided into three categories, which are 50 complex network, 28 tooth and 12 morphological features. Random Forest was chosen as the classifier. The average accuracy is 88%.

Kadir et al. (2012) proposed a method by using Zernike moments, which combined with three types of features that are geometric features, colour moments and gray-level co-occurrence matrices (GLCM). By using the proposed system Euclidean distance, City block distance and PNN the accuracy that obtained are 94.69%, 93.75% and 93.44% respectively. Besides, Wang et al. (2005) proposed MSD descriptor and the moving centre hydrosphere (MCH) classifier were applied to extract the shape features from preprocessing images. The experiment resulted in 20 classes of plant leaves being successfully identified and obtained classification result of 92.2%.

Lin & Peng (2008) combining the shape features and the texture features of the leaves of the broad-leaved tree, and then composing a synthetic feature vector of broad leaves. By using Probabilistic Neural Network (PNN), thirty kinds of broad-leaved trees give the accuracy around 98.3%. The Zernike moment and HOG approach have lent in the automatic recognition system based on the leaf shapes descriptors in Salve et al. (2016). By using Zernike moments, the recognition rate achieved is 84.66%, while HOG is 92.67% and Euclidean minimum distance classifier.

2.6.3 Combination of Shape Descriptors

Furthermore, some other previous studies have implemented both shape descriptors methods. Aakif and Khan (2015) proposed an algorithm in order to identify a plant in three levels, which are: pre-processing, feature extraction, and classification. The morphological features were extracted which included aspect ratio, eccentricity, roundness and convex hull. There are two additional features that were applied in this experiment which were Shape Defining Feature (SDF) and Fourier Descriptor. The classifier that had been used in this study is ANN with back propagation. The algorithm

has been applied to three types of leaf datasets, which are Flavia dataset, ICL dataset and their own dataset. The accuracy of greater than 96% was achieved.

Whereas, Ahmad et al. (2016) proposed an approach of the feature set is based on twelve geometrical features, five vein features and Fourier descriptors were performed. The multiclass of SVM is used for classification after dimensionality reduction using principal component analysis. The accuracy is 87.4%. Kadir et al. (2013a) proposed to integrate shape, vein, colour, and texture features in plant classification and uses PNN as a classifier. The Fourier descriptors, slimness ratio, roundness ratio and dispersion are used to represent shape features. Colour moments that contain mean, standard deviation and skewness are used to represent colour. Twelve textures features were extracted from lacunarity. The accuracy gives 93.75%, which is good enough for its performance.

Polar Fourier Transform (PFT) and three kinds of geometric features were used to represent shape features, then four kinds of colour moments were applied which are mean, standard deviation, skewness and kurtosis were proposed by Kadir et al. (2013b). Texture features were extracted from the gray-level co-occurrence matrices (GLCM) and vein features were applied and used PNN as classifier. The experimental results give an accuracy of 94.69% when using Flavia dataset and 93.08% when using Foliage dataset.

Besides, Prasad et al. (2013) presented shape and colour information of leaves using MSD and Fourier descriptor in order to represent the shape. The initial classification is calculated merely based on these shape descriptors using k-NN classifier and two classes with the highest result were selected. Then, the colour features were analysed and Prasad et al. (2013) found that colour features of the leaves increased the accuracy from 84.45%,

which used shape features only to 91.34%, which used combination of the shape and colour features.

Pham et al. (2012) proposed computer-aided plant species identification (CAPSI), which is based on plant leaf image by using shape-matching technique. Six method were tested in this experiment which are Fourier descriptors, Hu invariant moment, contour moment, curvature scale space, geometrical features and Modified dynamic programming (MDP). The experimental result gives the accuracy up to 92% and k-NN as classifier. Based on Gwo and Wei (2013) proposed a feature extraction technique for leaf contours and the outperform Zernike moments and curvature scale space are proposed. The experimental result show the accuracy is 92.7%.

Table 2.2 shows a comparison of the previous studies.

Table 2.2: A comparison of classification accuracies on the leaf identification and classification studies

Shape feature descriptor	Studies	Dataset	Descriptor	Classifier	Accuracy
Boundary-based	(Fu et al., 2004)	Own dataset	CCD	1-NN	94.26%
			FD	1-NN	85.71%
			Hu	1-NN	69.52%
			Chain code	1-NN	42.86%
	(Xiao et al., 2010)	Swedish	HOG	1-NN	93.17%
			IDSC	1-NN	93.73%
		ICL	HOG	1-NN	98.92%
			IDSC	1-NN	98.00%
	(Ren et al., 2012)	Swedish Leaf	IDSC	SVM	93.73%
			HOG	SVM	93.17%
			LDP	SVM	96.67%
		ICL	IDSC	SVM	95.79%
			HOG	SVM	96.63%
			LDP	SVM	97.70%
	(Bong et al., 2013)	Own dataset	CCG	1-NN	96.60%
			CCD	1-NN	74.40%
	(Mouine et al., 2013b)	Swedish Leaf	TAR	k-NN	90.40%
			TSL	k-NN	95.73%
			TSLA	k-NN	96.53%
	(Wang et al., 2015)	Flavia	MARCH	1-NN	73.00%
		Swedish	MARCH	1-NN	97.33%
		ICL	MARCH	1-NN	86.03%
Region-based	(Wang et al., 2005)	Own dataset	MSD, Hu	1-NN	92.60%
				4-NN	92.30%
				BPNN	92.40%
				H-S	92.20%

Table 2.2, continued.

Shape feature descriptor	Studies	Dataset	Descriptor	Classifier	Accuracy
Region-based	(Wu et al., 2007)	Flavia	MSD	PNN	70.09%
			MSD, Vein	PNN	90.31%
	(Lin & Peng, 2008)	Own dataset	MSD, Texture	PNN	98.30%
	(Du et al., 2007)	Own dataset	MSD, Hu	1-NN	93.00%
				k-NN	92.00%
				MMC	91.00%
	(Hossain & Amin, 2010)	Flavia	MSD	PNN	91.40%
	(Arora et al., 2012)	ImageCLEF	MSD	RF	88.00%
	(Kadir et al., 2012)	Own dataset	MSD, GLCM, Vein, Colour, ZM	PNN	93.44%
				Euclidean distance	94.69%
				City Block	93.75%
	(Bhardwaj et al., 2013)	Own dataset	MSD, moment features	PNN	91.50%
	(Caglayan et al., 2013)	Flavia	MSD	RF	87.61%
				k-NN	82.34%
				Naiye Baiyes	80.26%
				SVM	72.89%
			MSD, Colour moment	RF	93.95%
				k-NN	92.46%
				Naiye Baiyes	88.77%
				SVM	86.50%
			MSD, Colour moment & histogram	RF	96.30%
				k-NN	94.21%
				Naiye Baiyes	89.25%
				SVM	92.89%
	(Kulkarni et al., 2013)	Flavia	MSD, ZM, Vein, Colour, Texture	RBPNN	93.83%
	(Lavania & Matey, 2014)	Flavia	SIFT	SVM	87.50%

Table 2.2, continued.

Shape feature descriptor	Studies	Dataset	Descriptor	Classifier	Accuracy
Region-based	(Chaki et al., 2015a)		MSD	NFC	50.16%
				MLP	41.60%
			Texture	NFC	81.60%
				MLP	87.10%
			MSD, Texture	NFC	97.60%
				MLP	85.60%
	(Chaki et al., 2015b)	Flavia	MSD	NFC	97.50%
	(Kalyoncu & Toygar, 2015b)	Flavia	MSD, Hu	LDC	90.00%
		Leafsnap	MSD, Hu	LDC	71.00%
	(Salve et al., 2016)	Own dataset	Zernike, HOG	Euclidean distance	92.67%
Combination features	(Gwo & Wei, 2013)	Own dataset	Zernike, curvature scale	Statistical model	92.70%
	(Kadir et al., 2013a)	Flavia	MSD, PFD, Vein, Colour, texture	PNN	93.75%
	(Kadir et al., 2013b)	Flavia	MSD, PFD, Colour, Texture	PNN	94.69%
		Foliage	MSD, PFD, Colour, texture	PNN	93.08%
	(Pham et al., 2013)	Own dataset	MSD, FD, Hu, MDP	k-NN	92.00%
	(Prasad et al., 2013)	Flavia	PFT	k-NN	76.69%
			MSD, FD	k-NN	84.45%
			MSD, FD, Colour	k-NN	91.30%
	(Aakif & Khan, 2015)	Flavia	MSD, FD	BPNN	96.00%
		ICL	MSD, FD	1-NN	96.30%
		Own dataset	MSD, FD	1-NN	96.50%
	(Ahmad et al., 2016)	Flavia	MSD, FD, Vein	SVM	87.40%

2.7 Automated Plant Species Identification System

The regularly updated dataset and algorithm make the online service more attractive. In addition of studying classification approaches, several studies provide an implementation of the proposed methods as an app for web-based and mobile device. All of these depend on a reliable Internet connection. Despite in remote areas where plant identification applications are potential to be most valuable and useful, Internet connection may be unreliable or unavailable. The other approach is using efficient algorithms that run directly on the device without the need for Internet connection or the support server. However this is likely to be limitations in their classification performance (Wang et al., 2013).

2.7.1 Web-based Plant Identification System

Typically, with advancement in information technology, many systems and tools have been developed to assist and support botanists and experts in performing their research works. Three main web service of automated plant identification system have developed in the literature, which are iNaturalist, Pl@ntNet and Leaf Recognition. A summary of the features and requirements of web-based plant identification systems is listed in Table 2.3.

a) iNaturalist

iNaturalist.org was founded in 2008 and until now has been merely a crowdsourcing site. iNaturalist is an online social network of people that share the information of biodiversity to help each other to learn about nature (refer to Figure 2.14). It is also a crowdsourced species identification system and an organism occurrence-recording tool. A user can use it to record their observations, get help with identification and collaborate with others or access the data collected by others iNaturalist users.

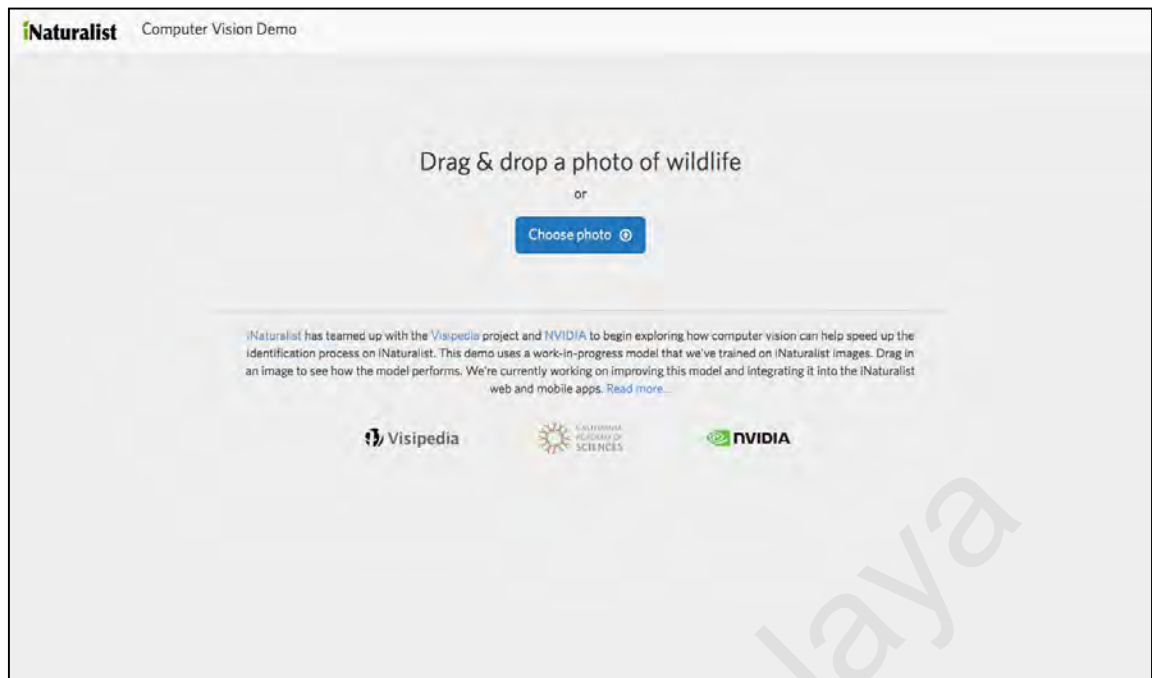


Figure 2.14: iNaturalist web interface

b) Pl@ntNet

Pl@ntNet (Joly et al., 2014) is an application for the identification of plants. Pl@ntNet is composed of three parts, which are an interactive web GUI for client, a content-based visual search engine, and a multi view fusion module on the server side (refer to Figure 2.15). Pl@ntNet is a research and educational initiative on plant biodiversity and was supported by Agropolis Foundation since 2009. Currently, Pl@ntNet consists of 16,675 plant species with a total of 709,411 images.

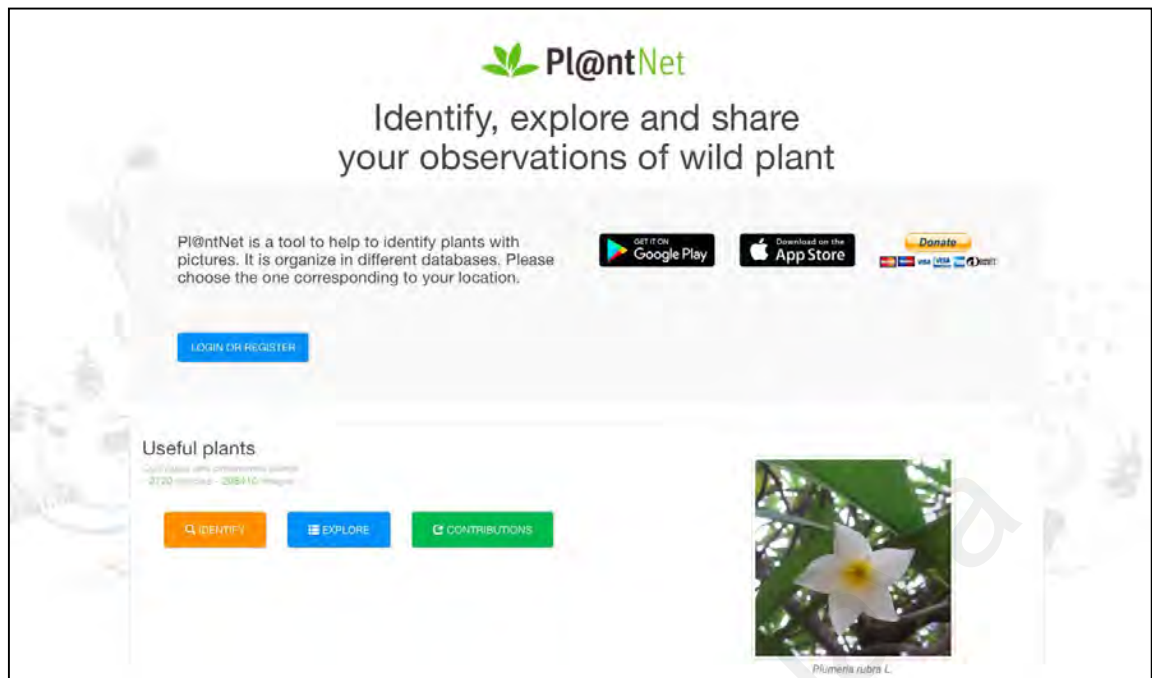


Figure 2.15: Pl@ntNet web interface

c) Leaf Recognition

The Leaf Recognition algorithm is a web-based tool for leaf identification of woody plant from Western Europe and was developed for the trees native in Central Europe and trees that are often planted in that region. The Leaf Recognition is a simple web application that is capable of determining an unknown leaf in the following based on five stages which are single image uploading, thresholding with user correction, user correction of calculated image, top ten results with similarity rate and filtering results by leaf type meta-data. The application code has been written in PHP, image processing uses ImageMagick Studio LLC 2013 and C++, which includes OpenCV library 2013 (refer to Figure 2.16).

Recognition of deciduous trees by leaf shape

Leaves as organs available most of the growing season, are an important element in the process of identifying an unknown plant. On these pages we introduce the automatic recognition of leaves - we publish here information about the project, links to datasets used and more. You can also try your success algorithms collected leaves. Please note recommendations for collecting and digitizing unknown leaves.

The output of the project GAUK no. 2012/524512 - *Digital Recognition trees by leaf* . Thank grant agency for this support.

Upload a photo sheet:

Choose File | no file selected

tell

Recommendations for collection of samples



For best results, recognition, please try to observe some rules. It is very important to note that the application accesses the worksheet definition differently from ordinary botanical terminology. In the composite sheet processed by only single tickets or upload your entire worksheet (eg ash, horse chestnut - horse chestnut, etc.). You get strange results. For palmate leaves folded card use your middle, with pinnate leaves turn leaves from the middle spring. This approach simplifies mathematical operations such as resolution petiole and blades, as well as a general description of shape; is also positive contribution to real usability lay people who in some leaves not distinguish whether they are posted or not. Also play a role proportions sheet, which in some folded sheets fit even up to A3 (eg. *Gymnocladus dioica* - Kentucky Coffeetree)

List of recognized taxa see [information about the dataset](#) .

List can be photographed should be placed on a sheet of paper, but it is preferable nascent color with a white background to 300dpi resolution - and thereby raise their own database entries. Application at the present stage of development does not allow segmentation sheet of ordinary photographs and this action must eventually be done manually before uploading the picture.



Figure 2.16: Leaf Recognition interface

Table 2.3: A summary of the features and requirements of web-based plant identification systems

Features / Requirements	Plant automated identification system		
	iNaturalist	Pl@ntNet	Leaf Recognition
Developer	Alex Shepard	French research organisations (Cirad, INRA, Inria and IRD), and the Tela Botanica network, with the financial support of Agropolis foundation	Petr Novotný, Tomáš Suk
Aim	Provide a crowd-sourced species identification system.	Help in the identification process, and extract the closest matches in the database rather than manually searching through thousands of entries.	Web-based tool for the leaf identification of woody plant from Western Europe
Operating system	Window, Mac, Linux	Window, Mac, Linux	Window, Mac, Linux
Background	Plain	Plain and natural	Plain
Analysis	Online	Online	Online
System requirement			
Query method	Image-based	Image-based	Image-based
Retrieval approach	Image recognition	Image recognition	Image recognition
Organ	Multi organ	Multi organ	Single leaf
URL	https://www.inaturalist.org/computer_vision_demo	https://identify.plantnet-project.org	https://www.inaturalist.org/computer_vision_demo

2.7.2 Mobile Apps

A mobile app is a software application developed specifically for use on small, wireless computing devices such as smartphones and tablets, rather than desktop and computers. A mobile app carries everything required for the implementation of a mobile plant identification system, along with a camera, a processor, a user interface and an Internet connection. These necessities make mobile app highly suitable for field use by professionals and the general public, despite these devices having less available memory, storage capacity, network bandwidth and computational power than desktop or server machines. Due to these constraints, some of the mobile apps offload some of the processing tasks to a high performance server.

Most recently, the plant identification method has shifted to portable devices such as tablets and smartphones. Three main mobile apps of automated plant identification system that are commonly available are Leafsnap, Pl@ntNet and Folia. A summary of the features and requirements of mobile apps for plant identification are listed in Table 2.4.

a) Leafsnap

One of the most established identification systems is Leafsnap (Kumar et al., 2012). Leafsnap (refer to Figure 2.17), so far is the most popular mobile app based on iOS platform for plant species identification. A user can take a photo of a leaf on plain background, and then transfer the image onto the Leafsnap server for analysis process and at last, see the information about the identified species. This application is limited to tree species of the Northeastern United States and it can be used only with an access to the Internet.



Figure 2.17: Leafsnap interface

b) Pl@ntNet

Pl@ntNet is an image sharing and retrieval application for the identification of plants (Joly et al., 2014). This application is developed by scientists from four French research organizations, which are Cirad, INRA, Inria and IRD and with the financial support of Agropolis foundation, Tela Botanical network. The number of species and images used by the application emerge with contributions of end users to the project. Images of tree leaves on plain background provide the most relevant results. Pl@ntNet was the first botanical identification system that is able to consider a combination for leaf, flower, fruit bark, and habit images for classification (refer to Figure 2.18). Currently, Pl@ntNet covers eleven regions, which are Canada, Western Europe, USA, Eastern Mediterranean, Hawaii, North Africa, Mauritius, Amazonia, Tropical Andes, Caribs and R union.

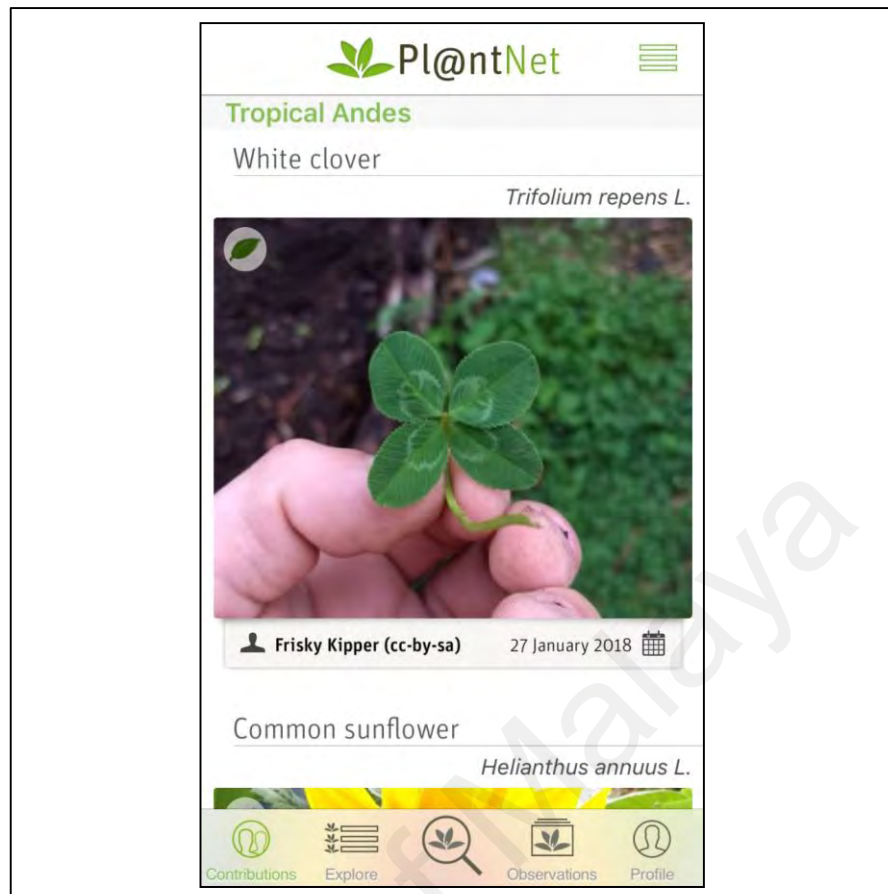


Figure 2.18: Pl@ntNet interface

c) FOLIA

FOLIA (see Figure 2.19) is an identification system that was developed by Cerutti et al., (2013), providing an educational iOS application that helps users in recognizing a plant species in its natural environment. FOLIA was developed by the LIRIS algorithm that analyses the picture. In order to perform this task, the users have to take a photo of an unknown leaf by using the smartphone's camera. The image then will be transferred to the FOLIA server for analysis with extracted high-level morphological features to predict the most exact and corresponding species. FOLIA only works for species of deciduous trees that grow naturally in France.



Figure 2.19: FOLIA interface

Table 2.4: A summary of the features and requirements of plant automated identification systems of mobile apps

Features / Requirements	Plant automated identification system		
	Leafsnap	Pl@ntNet	FOLIA
Developer	Columbia University, the University of Maryland, and the Smithsonian Institution	French research organisations (Cirad, INRA, Inria and IRD), and the Tela Botanica network, with the financial support of Agropolis fondation	Gaillaume Cerutti, Laura Tougne, Julien Mille, Antoine Vacavant, Didier Coquin
Aim	The free mobile apps use visual recognition software to help identify tree species from photographs of their leaves	A tool to help to identify plants with pictures	A tool to recognise trees by shooting leaves.
Application type	Mobile (iOS)	Web and Mobile (Android & iOS)	Mobile (iOS)
Background	Plain	Plain	Natural
Analysis	Online	Online	Online
System requirement			
Query method	Image-based	Image-based	Image-based
Retrieval approach	Image recognition	Image recognition	Image recognition
Organ	Single leaf	Multi organ	Single leaf
URL	http://leafsnap.com	https://identify.plantnet-project.org	http://liris.univ-lyon2.fr/rees/content/en/foliaen.php

2.8 Summary

This chapter provides the findings of previous studies as well as the current status of plant leaf image databases and the automated plant identification systems. A computer aided plant recognition system can offer experts and non-experts to take part in recognition process of unknown plant species. There are many biodiversity databases but most of them covered the plant species from the western regions. Specifically, there are two ways in categorization of the feature descriptors for leaf shape, which are boundary-based and region-based. In boundary-based, shape descriptors extract shape feature entirely from the contour of a shape, while region-based shape descriptors obtain shape features from the whole region of a shape. The advantages and disadvantages of both methods are discussed.

Neural network, SVM and K-NN algorithms are the most commonly implemented classifiers in the previous studies. Currently, three main mobile apps and web services for automated plant identification are: Leafsnap, Pl@ntNet and Folia for mobile apps and whereas: iNaturalist, Pl@ntNet and Leaf Recognition for web service.

In summary, rapidly environment degradation and a limited number of professional botanists represents significant challenges to the future of botanical study and conservation. Moreover, some of the plants are at margin of extinction, thus there is a need to protect and manage them. Therefore, rapid and accurate plant recognition and classification system is crucial for effective study and management of biodiversity.

CHAPTER 3: RESEARCH METHODOLOGY

3.1 Introduction

In the previous chapter, various feature extraction methods, feature selection algorithms and machine learning techniques, which were used in plant identification, were reviewed and the promising results from various plant image datasets are presented. This chapter covers the research methodology, which explains in details how this research was conducted. This chapter covers techniques, tools and research framework used in this research. Manual tasks such as image digitalization of fresh captured leaves from the sampling areas were required as the reference dataset. The proposed research framework involved is required to attain the objectives of this research and the methodology adopted will be discussed in the following sections.

3.2 Proposed Framework

In this section, leaf images are taken for image processing process and the framework for tropical shrub species classification is shown in Figure 3.1. Basically, there are five main steps, which are field sampling, image acquisition, image pre-processing, feature extraction and classification process. Each of these steps will be discussed in more detail in this and following sections. The image, obtained by camera, usually has some restrictions or interference with noises and unrelated object. Before feature extraction, it is necessary and indispensable to carry out image processing including image de-noising, image enhancement and image segmentation.

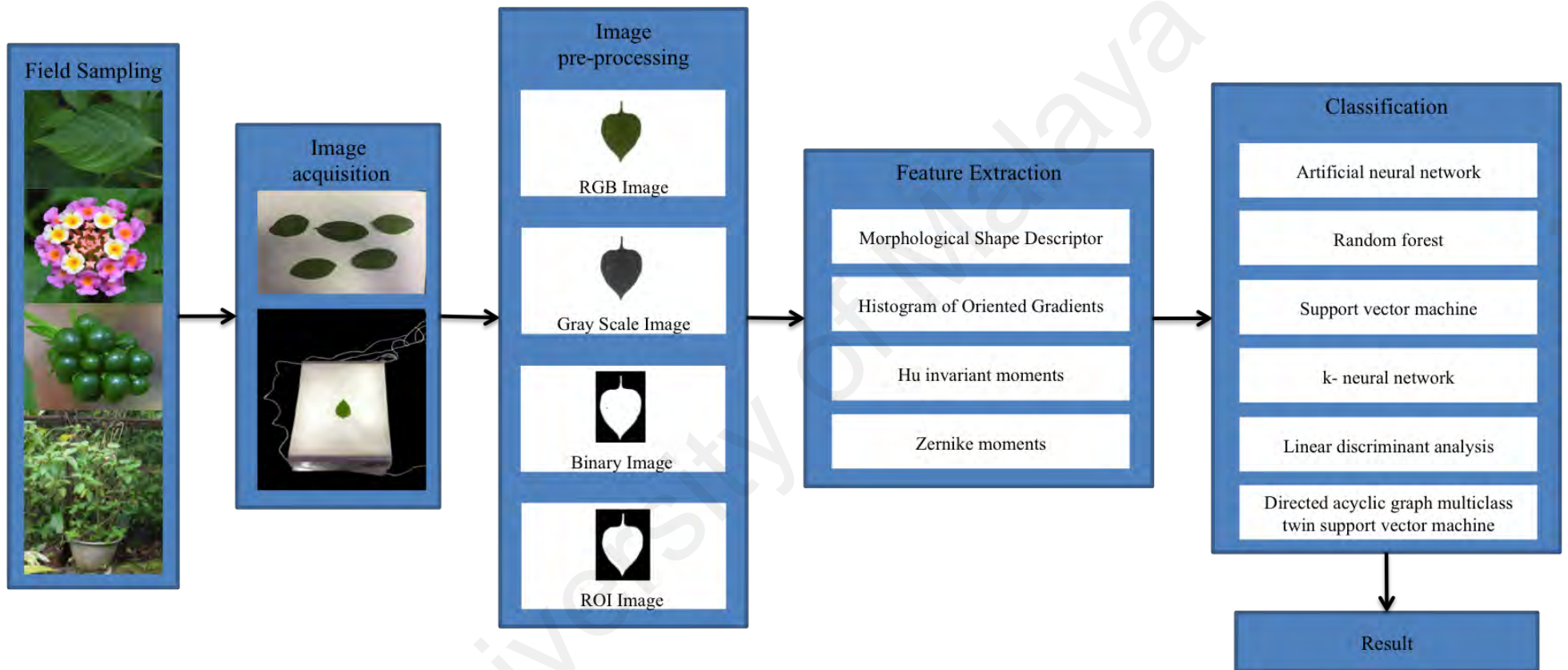


Figure 3.1: Flowchart for the proposed methodology

3.3 Field Sampling

Tropical rainforests are recognised as one of the most productive type of forests in the world. There are three areas in the world where tropical rainforests are found, which are South America, Central Africa and Southeast Asia (The Malaysia Rainforest, 2017). Generally, tropical rainforests in Malaysia are tropical lowland rainforests (Lowland Forest. 2017). The term of tropical lowland forest is used to represent a forest where there is little or no seasonal water storage and the climate is continuously warm and humid. There are about 391,000 vascular plant species of plant in the world, and Malaysia has about 15,000 species of vascular plants (Willis, 2017). There are 12,300 plants that Malaysia shares with other countries and the balance of 2,700 species is endemic to Malaysia.

The scope of this research is focused on the tropical shrub species, which are commonly seen in Malaysia. Firstly, the data collection was conducted by collecting the leaf images of tropical shrub species. This dataset is named as myDAUN, in which ‘my’ represents Malaysia and ‘DAUN’ means leaf in Malay language. The images in myDAUN dataset were sampled from the campus of University of Malaya (UM), Kuala Lumpur, Malaysia. UM is Malaysia’s oldest university and situated on a 922-acre (373.12-hectare) campus in the southwest of Kuala Lumpur (Our History, 2017). There is a botanical garden named as “Rimba Ilmu” that is situated in the UM campus and it consists of over 1600 plant species and is one of the most important biological conservatories in Malaysia (Rimba Ilmu Botanic Garden. 1999). Four main locations in UM with more variety of tropical shrubs were chosen and the sampling took place in these locations. The four main locations of sampling are Faculty of Science, Dewan Tunku Canselor (Tunku Canselor Hall), Varsity Lake and Main Library as shown in Figure 3.2.



Figure 3.2: Location of sampling area in the University of Malaya (UM), Kuala Lumpur, Malaysia.

The species of tropical shrub were identified and selected with the help of botanists. In botany and ecology, a shrub or bush is a small to medium sized woody plant. It is different with the herbs because shrubs have persistent woody stems above the ground and they are unlike trees since they have multiple stems and shorter in height. Since shrubs have a variety of species and cultivar, thus the advice from the professional botanists and the staff from the botanical garden are crucial and valuable.

As the result, 31 tropical shrub species were collected from Faculty of Science, 8 tropical shrub species collected from Tunku Canselor Hall, 3 tropical shrub species were collected from Varsity Lake and the rest of tropical shrub species were collected from the Main Library. Thus, in total, 45 species of common tropical shrubs were selected and 30 leaf samples were collected for each species. Hence there are 1350 images of tropical shrub leaf images (see Appendix A). Table 3.1 shows the selected species in myDAUN database.

Table 3.1: List of tropical shrub species in myDAUN dataset

Sampling location	Label	Scientific Name	Common name or General name
Faculty of Science	1	<i>Acalypha siamensis</i>	Tea Leaves
	2	<i>Acalypha wilkesiana</i>	Copperleaf
	5	<i>Brunfelsia calycina</i>	Yesterday, Today and Tomorrow
	6	<i>Clinacanthus nutans</i>	Sabah Snake Grass
	7	<i>Dillenia suffruticosa</i>	Yellow Simpoh
	8	<i>Dracaena surculosa</i>	Japanese Bamboo
	9	<i>Dracaena reflexa</i>	Song of India
	12	<i>Graptophyllum pictum</i>	Caricature
	15	<i>Lagerstroemia indica</i>	Crepe Myrtle
	16	<i>Lantana camara</i>	Lantana
	17	<i>Lawsonia inermia</i>	Henna
	19	<i>Magnolia figo</i>	Banana Shrub
	20	<i>Malvaviscus arboreus</i>	Sleepy Mallow
	22	<i>Melastoma malabathricum</i>	Sesenduk
	27	<i>Polyscias balfouriana</i>	Dinner-plate Aralia
	28	<i>Sauropus androgynus</i>	Star Gooseberry
	29	<i>Strobilanthes crispus</i>	Bayam Karang
	30	<i>Tabernaemontana divaricata</i>	Ceylon Jasmine
	31	<i>Tibouchina urvilleana</i>	Glory Bush
	32	<i>Citrus microcarpa</i>	China orange
	33	<i>Mentha piperita</i>	Peppermint
	34	<i>Andrographis paniculata</i>	King of bitters
	35	<i>Rhodomyrtus tomentosa</i>	Downy rose myrtle
	36	<i>Orthosiphon aristatus</i>	Cat's whiskers
	37	<i>Centratherum punctatum</i>	Lark daisy
	38	<i>Polygonum minus</i>	Laksa leaf
	40	<i>Justicia gendarussa</i>	Gendarusa
	41	<i>Tetracera scandens</i>	Stone leaf
	42	<i>Piper sarmentosum</i>	Wild pepper

Table 3.1, continued.

Sampling location	Label	Scientific Name	Common name or General name
Faculty of Science	44	<i>Flemingia strobilifera</i>	Wild hops
	45	<i>Cananga odorata</i>	Ylang- ylang
Tunku Canselor Hall	10	<i>Duranta erecta</i>	Golden Dew-Drop
	11	<i>Excoecaria cochinchinensis</i>	Chinese Croton
	14	<i>Ixora javanica</i>	Jungle Geranium
	23	<i>Murraya paniculata</i>	Kemuning
	24	<i>Mussaenda erythrophylla</i>	Red Flag Bush
	25	<i>Mussaenda philippica</i>	White Mussaenda
	26	<i>Phyllanthus myrtifolius</i>	Ceylon Myrtle
	43	<i>Rauvolfia serpentine</i>	Indian snakefoot
Varsity Lake	3	<i>Allamanda cathartica</i>	Golden Trumpet
	4	<i>Bougainvillea spectabilis</i>	Great Bougainvillea
	13	<i>Hibiscus rosa-sinensis</i>	Chinese Hibiscus
Main Library	18	<i>Loropetalum chinense</i>	Chinese Fringe-flower
	21	<i>Manihot esculenta</i>	Manihot
	39	<i>Tabernaemontana coronaria</i>	Crepe jasmine

3.4 Image Acquisition

Firstly, leaf samples that could represent the existing population were identified. Then, the leaf samples from different parts of the shrub and size were plucked and collected. There are standard criteria to follow which is the selected leaf is not ruptured, abnormal or damaged. Leaves that are off-colour, grazed, over mature, diseased, or otherwise not normal, were avoided to be used in this study. Secateurs was used for a clean cut of the stem. Secateurs is a type of scissors for use on plants. They are strong enough to prune hard branches of trees and shrubs, which sometimes is up to two cm thick.

While at the site, each sample was recorded in field notebook by giving it a number and the plant name was recorded (if it is known). Next, 30 samples of leaf were collected for each species. After that, the collected samples were brought back to the laboratory for image acquisition. If the leaves were not flat enough, the leaves were compressed by using a press to flatten them for three to four hours to ensure the leaves become delicate and brittle. Before compressing, the leaves were cleaned up by gently wiping them clean with a tissue paper to remove any dirt or moisture, and leaf stalks were removed. Next, the leaves were placed on a paper and were compressed with books and newspapers to keep the leaves from curling (see Figure 3.3).

In order to obtain a quality leaf photo, the light box was designed and the sizes of light boxes were 37cm x 59cm x 13.5cm. A light box is a contraption with translucent sides that diffuses light coming from multiple sources and this allows for even nearly shadow less lighting against a solid background. The setup of image acquisition step is shown in Figure 3.4.

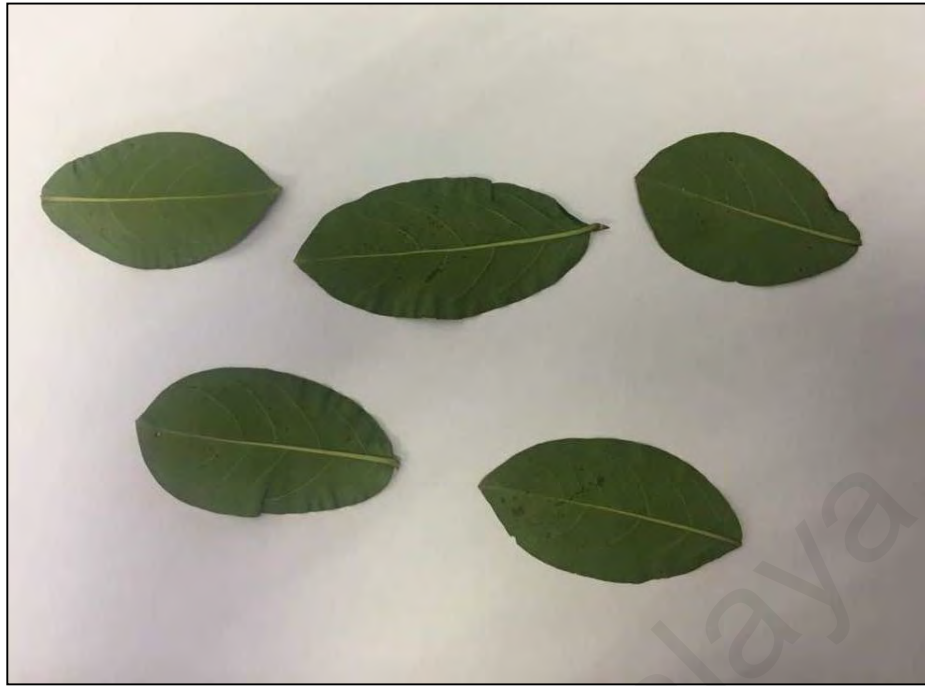


Figure 3.3: Arrangement of leaves before compression



Figure 3.4: Light box and experimental setup

3.5 Software and Hardware Requirement

A total of 1350 fresh leaf images were captured. The main software that were used in this research are the Adobe Photoshop CC and Matlab R2015b. Adobe Photoshop CC was used to enhance the image quality and to eliminate the illumination and contrast problem, which would affect the process of segmentation. Adobe Photoshop CC was used to manually pre-process the leaf images. All the shadow, dust and any undesired object or noise from all images were cleaned up. This process is applied to the dataset in order to facilitate the leaf segmentation. Besides that, Matlab was used for extracting the features and to train the classifiers for image classification.

The images were taken in the same standard with uniform background. The image of the leaf samples was captured on the front side, from a distance of 55cm from the camera.

The camera specifications are as following:-

- Name: Nikon D750 DSLR
- Total pixels: 24.3 Megapixels
- Image sensor type and size: CMOS Sensor Type with 35.99mm/24.0mm
- Highest resolution size: 6016 x 4016 pixels
- Lens: AF-S Nikkor 24-120 mm F4G ED VR lens

The images are stored as 32- bits RGB colour in uncompressed Tagged Image File Format (tiff) format. Tiff is a great choice for archiving images because all details are preserved and no image data is lost. Figure 3.5 shows the samples of all tropical shrub species in myDAUN dataset.

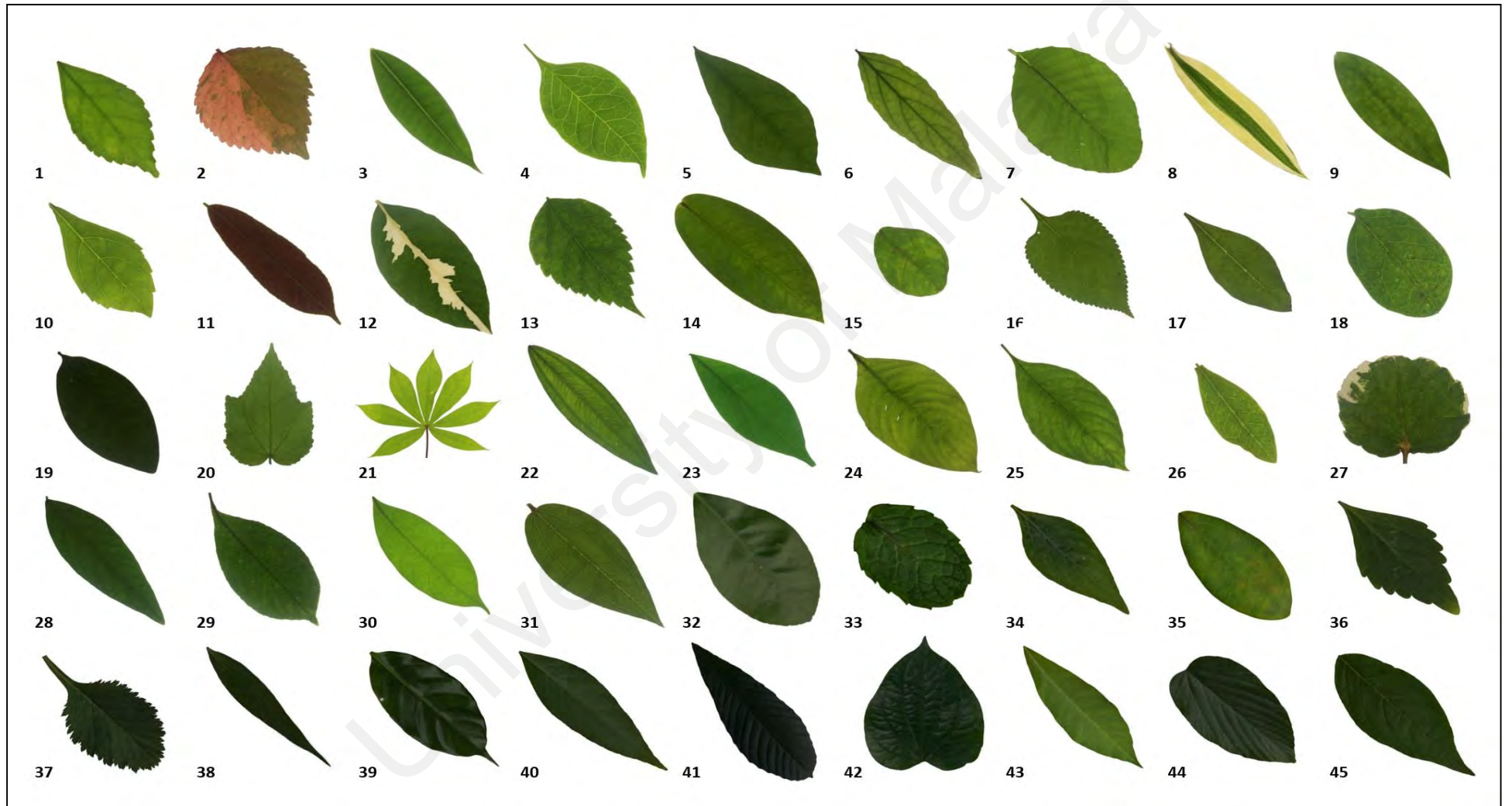


Figure 3.5: Samples of the leaf images in the myDAUN dataset

3.6 Image Pre-processing

The image pre-processing does not change the image information content. It is valuable on a variety of situations where it helps to conceal information that is not relevant to the specific image processing. The main objective of image pre-processing is to identify the main object, which is the leaf shape, and to get rid of all other unrelated and undesired information. The region of interest (ROI) of the leaf must be obtained before the extraction of the morphological descriptors through image segmentation process. Figure 3.6 presented a noisy leaf image sample and Figure 3.7 shows a sample of the tropical shrub leaf image after cleaned up using Photoshop CC.

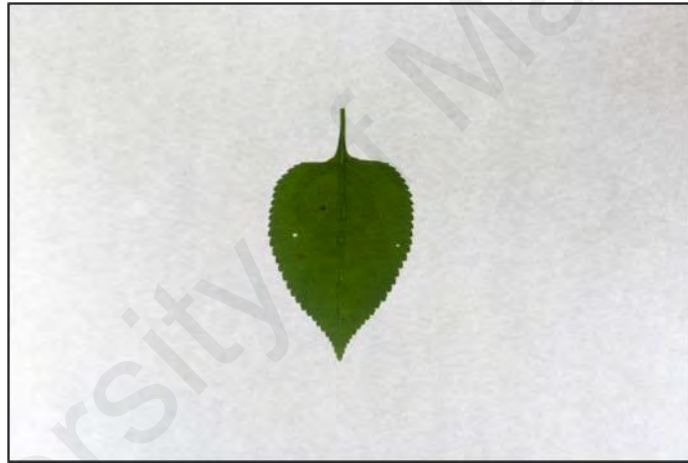


Figure 3.6: A *Lantana camara* sample before image enhancement



Figure 3.7: A *Lantana camara* sample after cleaned using Photoshop CC

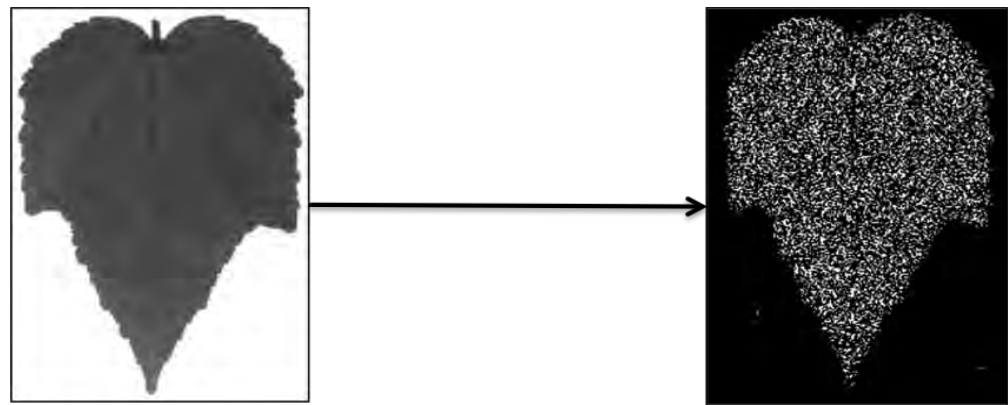
3.7 Image Segmentation

After image enhancement process, the images were successively using MATLAB R2015b through image segmentation process. Firstly, the RGB image or the original image is converted to grey-scale image (see Figure 3.8). This process was to convert the true colour image RGB to the grey-scale intensity image by eliminating the hue and saturation information while retaining the luminance.



Figure 3.8: Conversion of RGB image into grey-scale image

The next operation is edge detection by binary gradient mask or “Canny” edge detector. The binary gradient mask was applied to ensure the maximum contrast of the boundary of the object of interest with the background. “Canny” edge detector is commonly used in edge detection algorithms by computing the image gradient intensity function. The Canny method finds edges by examining for local maxima of the gradient of grey-scale. The edge function calculates the gradient by applying the derivative of the Gaussian filter. This method used two thresholds to detect strong and weak edges and it is less likely than other methods to be tricked by noise, but is more likely to detect true weak edges. Figure 3.9 shows the Canny edge detection on a grey-scale image.

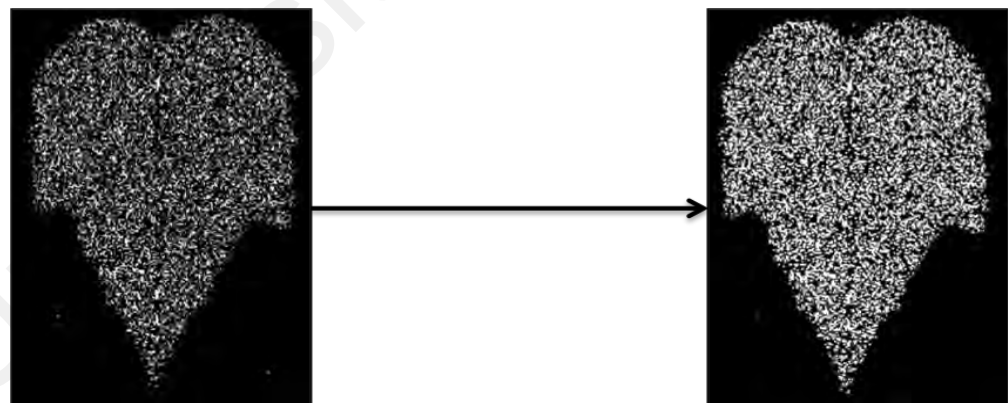


(a) Grey-scale image

(b) Detected edge image

Figure 3.9: Conversion of grey-scale image into detected edge image.

Image after the “Canny” edge detector was performed to form images with edge detected object. The output of binary image replaces all pixels in the input image with luminance greater than level with the value 1 and replaces all other pixels with the value 0. A canny edge detector is considered as the most powerful edge detector for image segmentation (Canny, 1986). Figure 3.10 shows the binary image on edge detector image.



(a) Detected edge image

(b) Binary image

Figure 3.10: Conversion of detection edge image into binary image.

After performing binary image conversion, holes and gaps still exist within the boundary of the region of interest. Hence, the dilation operator was subsequently executed on the current image. The dilation was done in order to enhance and enlarge the lines within the

boundary of the region of interest, reducing holes and gaps. The binary image is then converted to a filled binary image using *imdilate* function. Figure 3.11 shows the filled binary image on the binary image.

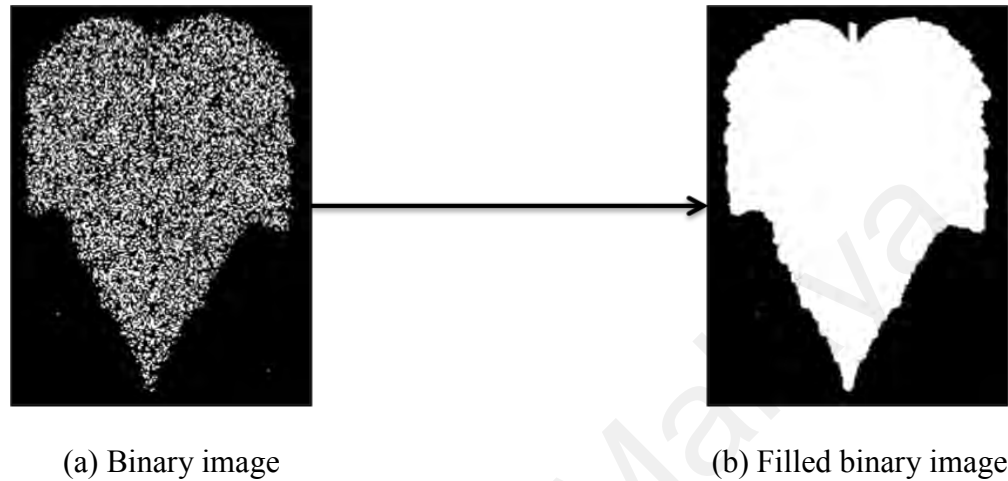


Figure 3.11: Conversion of binary image to filled binary image.

In order to obtain the desired shape of leaf in the image, the small particles in the surrounding were removed and the ROI was obtained through the segmentation process. The images would then perform a flood-fill operation on the background pixels of the input of filled binary image by using “fill” operator. The detected holes will be converted into lighter pixels, which to its original dark pixel value that make the pixel values of the holes equal to the area surrounding them. The filled binary image is then converted to region of interest (ROI) using *imfill* function. Figure 3.12 shows the ROI image on the filled binary image.

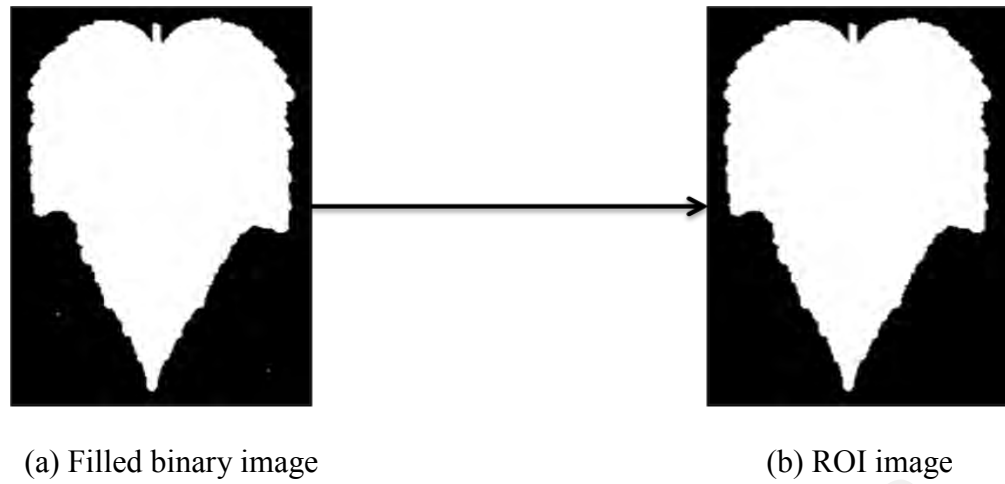


Figure 3.12: Conversion of filled binary image to region of interest (ROI) image.

3.8 Feature Extraction

The next stage in the tropical shrub species classification is the feature extraction phase. The main benefit of this phase is that it rid and removed redundancy from the image and represents the leaf of tropical shrub species by a set of numerical features. All of these features would be used for the classifiers to classify the data. The ROIs obtained from the image pre-processing step are used as input in the feature extraction steps. There are four types of representations that are applied and tested, which are morphological shape descriptors (MSD), histogram of oriented gradients (HOG), Hu invariant moments (Hu) and Zernike moments (ZM).

3.8.1 Morphological Shape Descriptor (MSD)

In this study, five basic shape descriptors were used for leaf analysis, namely diameter, major axis length, minor axis length, area, and perimeter. Based on these basic shape descriptors, 15 morphological descriptors are computed, for example, aspect ratio, rectangularity, circularity, form factor, and etc. Thus, there are 20 descriptors of MSD were implemented in this study as listed in Table 3.2.

Table 3.2: Basic geometrical and morphological shape descriptors






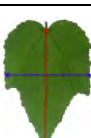
Descriptor	Description	Pictogram	Formula
Diameter	The longest distance between any two points on the margin of the leaf		D
Major axis length	Line segment that connect the base and the tip of the leaf		H
Minor axis length	Maximum width which perpendicular to the major axis		W
Area	Number of pixels in the region of the leaf		A
Perimeter	Total sum of the distance between each adjoining pair of pixels around the border of the leaf		P
Aspect ratio	Ratio of major axis length over minor axis length		$AR = \frac{H}{W}$

Table 3.2, continued.





Descriptor	Description	Pictogram	Formula
Form factor	Illustrates the difference between a leaf and a circle		$FF = \frac{4\pi A}{p^2}$
Rectangularity	Indicates how rectangle a shape is, that represent how much it fills its minimum bounding rectangle		$R = \frac{A}{HW}$
Solidity	Ratio between leaf's area and area of the leaf's convex hull		$S = \frac{A}{CA}$
Eccentricity	Ratio of the distance between foci of the ellipse over its major axis length		$EC = \frac{foci}{major\ length\ foci}$
Narrow factor	Ratio of the diameter over the major axis length		$NF = \frac{D}{H}$

Table 3.2, continued.


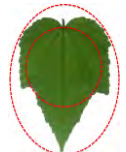

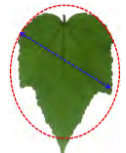





Descriptor	Description	Pictogram	Formula
Convex area	Smallest region that include its convex and leaf's region		CA
Irrectangularity	Ratio of the radius of the inside circle of the bounding box over the radius of the outside circle of the bounding box		$IR = \frac{ri}{ro}$
Entirely	Normalised difference of the convex hull area with leaf's area over area of leaf		$EN = \frac{(convex\ hull - A)}{A}$
Equivalent diameter	Diameter of circle with the same area as the leaf's area		$ED = \sqrt{\frac{4A}{\pi}}$
Perimeter ratio of major and minor axis length	Ratio of the perimeter over the total of sum of the length of major axis and minor axis		$PMM = \frac{P}{(H + W)}$

Table 3.2, continued.

Descriptor	Description	Pictogram	Formula
Perimeter of convexity	Ratio of the perimeter of convex and perimeter of the leaf		$PC = \frac{P_{convex}}{P}$
Perimeter of area	Ratio of the perimeter over the leaf's area		$PA = \frac{P}{\sqrt{A}}$
Perimeter ratio of diameter	Ratio of the perimeter of the leaf over the diameter of the leaf		$PD = \frac{P}{D}$
Perimeter ratio of major axis length	Ratio of the perimeter of the leaf over the major axis length		$PM = \frac{P}{H}$

3.8.2 Histogram of Oriented Gradients (HOG)

The Histogram of Oriented Gradients (HOG) are descriptors used in image processing for object detection and it is the local statistic of the orientations of the image gradients around key points (Dalal & Triggs, 2005; Xiao et al., 2010). HOG descriptors method determines occurrences of gradient orientation in localised portions of an image or ROI. This technique is alike to scale-invariant feature transformation descriptors, edge orientation histograms and shape context. Gradient computation, G and gradient orientation, θ are computed using Equation 1 and Equation 2 respectively.

$$|G| = \sqrt{G_x^2 + G_y^2} ; \quad \text{Equation 1}$$

$$\theta = \arctan \frac{G_x}{G_y} ; \quad \text{Equation 2}$$

where G_x is gradient in X direction and G_y is gradient in Y direction

Each pixel within a cell casts a weighted vote for an orientation-based histogram based on the gradient magnitude, G and gradient orientation, θ . One histogram was counted for each cell based on the number of bins orientation binning. After that the image was split into a number of cells. A cell can be represented as a region like a square with a predefined size in pixels. Each block has 3x3 cells and for each cell, the histogram of gradient by splitting votes into bins for each orientation. The normalization is executing among a group of cells, called as a block and a normalization factor was calculated over the block. All the histograms within this block were normalised and linked together in single feature vector. Normalised vector, V can be performed by

$$V = \frac{V_K}{\|V_K\|_2^2 + \epsilon^2} ; \quad \text{Equation 3}$$

where V_K is the vector for combined histogram and ϵ is a small constant.

The histogram of all the blocks accumulated into a whole HOG descriptor was processed. In this study, the number of bins, K was set to 9, whereas the block size was 3x3 cells. Thus, there were 81-dimensional vector for each of leaf image based on the computation of 3x3x9.

HOG can capture gradient or edge structure that is very characteristic of local shape and it contains better invariance to local geometric and photometric transformations by using gradient and histogram normalization. HOG descriptors that were applied in tropical shrub species classification do not need pre-understanding of leaf structure, since HOG do not extract features from the typical and general botany characteristic of leaf, for example length-width ratio or number of lobes. The HOG not only simplifies the classification procedure but at the same time it removes the influence of botany conception, which may changes all the time.

3.8.3 Hu Invariant Moments (Hu)

Moments and functions of moments have been fully employed as invariant global features of images in pattern recognition. Image moment is a particular weighted average moment of the image pixel intensities or a function of such moments, commonly chosen to have some attractive property of interpretation (Bhardwaj et al., 2013). The idea of using moments in shape recognition was first introduced by (Hu, 1962). Moment functional have attracted due to their mathematical simplicity and various physical interpretations. Let $\{\eta_n\}$ be a real sequence of numbers and defined by Equation 4.

$$\Delta^m \eta_n = \sum_{i=0}^m (-1)^i \binom{m}{i} \eta_{n+i} \quad \text{Equation 4}$$

where $\Delta^m \eta_n$ is the m^{th} order derivative of η_n .

A basic and necessary condition that there exists a function F(x) satisfying the system is given by Equation 5

$$\eta_n = \int_0^1 x^n dF(x), \quad n = 0, 1, 2, \dots \quad \text{Equation 5}$$

Hence, the system of linear inequalities.

$$\Delta^m \eta_n \geq 0 \quad k = 0, 1, 2, \dots \quad \text{Equation 6}$$

where, if $f(x)$ is appositive function that is the case in machine processing, then the set of functional is given by Equation 7.

$$\int_0^1 x^n f(x) dx, \quad n = 0, 1, 2, \dots \quad \text{Equation 7}$$

Hu defined seven invariant moments computed from central moments through order up to three and two-dimensional those are invariant under object translation, scaling and rotation. Hence, a set of seven invariant moments can be derived from the normalised central moments as stated in Equation 8.

$$Hu1 = \eta_{20} + \eta_{02};$$

$$Hu2 = (\eta_{20} - \eta_{02})^2 + 4\eta_{11}^2;$$

$$Hu3 = (\eta_{30} - 3\eta_{12})^2 + (\eta_{03} - 3\eta_{21})^2;$$

$$Hu4 = (\eta_{30} + 3\eta_{12})^2 + (\eta_{03} + \eta_{21})^2; \quad \text{Equation 8}$$

$$Hu5 = (\eta_{30} - 3\eta_{12})(\eta_{30} + \eta_{12})[(\eta_{30} + \eta_{12})^2 - 3(\eta_{03} + \eta_{21})^2] + (3\eta_{21} - \eta_{03})(\eta_{21} + \eta_{03})[3(\eta_{30} + \eta_{12})^2 - (\eta_{03} + \eta_{21})^2];$$

$$Hu6 = (\eta_{20} - \eta_{02})[(\eta_{30} + \eta_{12})^2 - (\eta_{03} + \eta_{21})^2] + 4\eta_{11}(\eta_{30} + \eta_{12})(\eta_{03} + \eta_{21});$$

$$Hu7 = (3\eta_{21} - \eta_{03})(\eta_{30} + \eta_{12})[(\eta_{30} + \eta_{12})^2 - 3(\eta_{03} + \eta_{21})^2] + (3\eta_{21} - \eta_{30})(\eta_{21} + \eta_{03})[3(\eta_{30} + \eta_{12})^2 - (\eta_{03} + \eta_{21})^2];$$

3.8.4 Zernike Moments (ZM)

In order to compute Zernike moments (ZM), three steps are required, namely computation of radial polynomials, Zernike basis functions, and ZM. The approach to obtain ZM from an input image starts with the computation of Zernike radial polynomials (Hwang & Kim, 2006). ZM are based on Zernike polynomials that are orthogonal to the circle $x^2 + y^2 = 1$. The form of these polynomials is formulated in below.

$$K_{ab}(x, y) = R_{ab}(r) \exp(jb\theta) ; \quad \text{Equation 9}$$

where a is a non-negative integer, b is positive or negative integer satisfying constrains $a-|b| = \text{even}$ and $|b| \leq a$. r is the radius of (x, y) to the centroid where $r = \sqrt{x^2 + y^2}$, θ is the angle between r and x-axis where $\theta = \tan^{-1} \frac{y}{x}$, $j = \sqrt{-1}$. R_{ab} is the radial polynomial defined as

$$R_{ab}(r) = \sum_{s=0}^{(a-|b|)/2} (-1)^s \frac{(a-s)!}{s! \left[\frac{a+|b|}{2} - s \right]! \left[\frac{a-|b|}{2} - s \right]!} r^{a-2s} ; \quad \text{Equation 10}$$

The ZM for order a and b repetition of continued function $f(x, y)$, if $f(x, y)$ is a digital image, is defined below:

$$Z_{ab} = \frac{a+1}{\pi} \sum_x \sum_y f(x, y) K_{ab}^*(r, \theta) ; \quad \text{Equation 11}$$

In this case, K^* is the complex conjugate, while K_{ab} is the Zernike basis functions order a with b repetitions, where $K_{ab}(x, y) = K_{ab}(r, \theta) = R_{ab}(r) \exp(jb\theta)$

These descriptors need to be normalised before classification. The normalised ZM can be calculated using

$$Z'_{ab} = \frac{Z_{ab}}{m_{00}} ; \quad \text{Equation 12}$$

where m_{00} is spatial moment order $(0, 0)$ that indicates the area of a leaf.

The ZM with order a counting from 0 to 8 were selected as the descriptors and 25 descriptors of ZM were obtained.

3.9 Feature Selection

Feature selection is a process of identifying and removing the irrelevant and redundant features to describe the target concept. Feature selection reduced the dimensionality of the data and allowed learning algorithms to operate faster and more effectively. It is important to have a feature selection algorithm in the proposed model to avoid over-fitting since there are a vast numbers of variables and the small sample size.

In this study, three feature selection methods are proposed in order to find out the most optimum feature subset for the tropical shrub species. The purpose is to minimise the number of input variables and thus, reducing the time and costs required for tropical shrub species. Three feature selection methods are selected and implemented in this research, which are Relief, Correlation-based feature selection (CFS), and Pearson's correlation coefficient (PCC).

3.9.1 Relief-F

Relief belongs to the filter approach. In this method, each of the feature input is ranked and weighted using the k-nearest neighbours classification, in which the value of k is set to 1. The top features with large positive weights are chosen and selected for each of all descriptors, which are MSD, HOG, Hu and ZM. Figure 3.13 presents the flowchart for the Relief-F method.

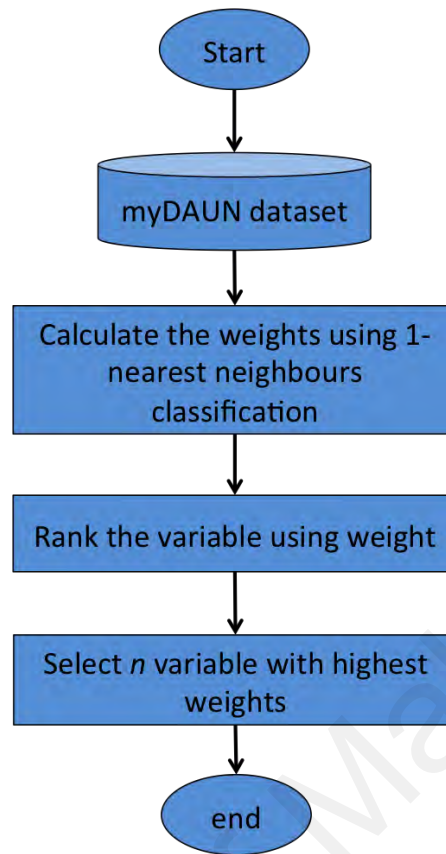


Figure 3.13: The flowchart of the Relief-F feature selection

3.9.2 Correlation-based Feature Selection (CFS)

The correlation-based feature selection is sorted and ranked for each of the feature input according to pairwise correlations. The ranking of the feature input is followed accordingly to minimum correlations. The top features with highest ranked are chosen and selected for each of all descriptors, which are MSD, HOG, Hu and ZM. Figure 3.14 presents the flowchart for the correlation-based feature selection method.

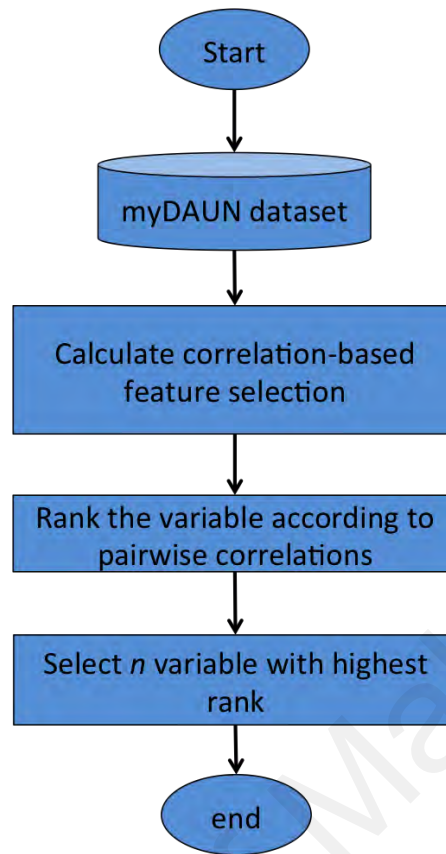


Figure 3.14: The flowchart of the correlation-based feature selection

3.9.3 Pearson's Correlation Coefficient (PCC)

In this research, the Pearson's correlation coefficient, m , is computed and ranked for each of the feature input and the one with the highest m is chosen. For example, there are 20 input variable for MSD descriptors, thus for the 50% of the input model, the top 10 inputs with the highest m value is chosen and selected. This is repeated for the HOG, Hu and ZM descriptors. The flowchart for this method is shown in Figure 3.15.

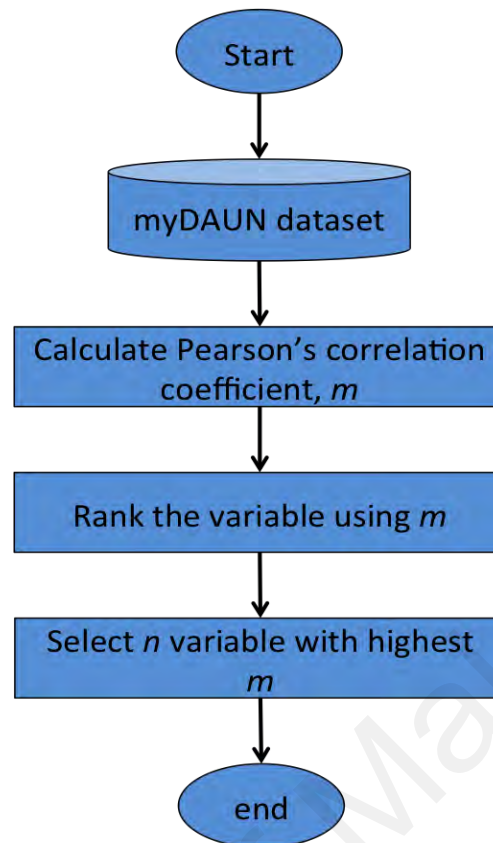


Figure 3.15: The flowchart of the Pearson's correlation coefficient

3.10 Classification

In this study six classifiers were tested and applied in tropical shrub species classification, which are artificial neural network (ANN), random forest (RF), support vector machine (SVM), k-nearest neighbour (k-NN), linear discriminant analysis (LDA) and directed acyclic graph multiclass least squares twin support vector machine (DAG MLSTSVM). All of classification algorithms were tested using random sampling approach. The myDAUN dataset was randomly divided into 80% for training and 20% for testing. This process was repeated 10 times and the average of 10 runs was taken as the final result.

3.10.1 Artificial Neural Network (ANN)

ANN is used as the classification tool and is a biologically inspired program designed to stimulate the system in which the human brain processes information. The general neural network consists of three layers, which are input layer, hidden layer, and output layer. The ANN is composed of a set of neurons that is interconnected with each other.

The total of 133 descriptors that was obtained, which includes 20 descriptors from MSD descriptor, 81 descriptors from HOG descriptors, seven descriptors from Hu and 25 descriptors from ZM. All of the parameters of ANN were setup as in Appendix B. The number of output neurons was present by the number of species classified, which in this case, are 45 classes. The networks were two-layer feed forward with 133 input nodes and 45 output nodes as shown in Figure 3.16.

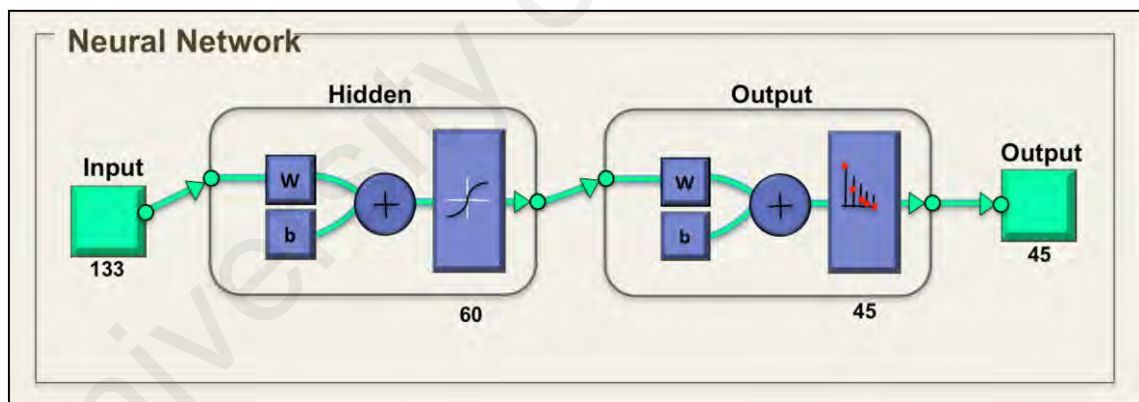


Figure 3.16: Neural network for tropical shrub species

3.10.2 Random Forest (RF)

Additionally, in this study implemented RF classifier. According to Breiman (2001) RF is substantial modifications of bagging in order to build a large collection of de-correlated trees and average them. RF is predictions of multiple classification trees are aggregated for a dataset and each tree in the forest is grown using bootstrap samples. At prediction time, trees in the forest use their votes for target class and classification results are taken

from each tree. The forest selected the class that achieved the most votes among the separate trees. Below are the algorithm used for classification of RF.

For $g=1$ to G :

- (i) Draw a bootstrap sample T^* of size N from the training data.
- (ii) Grow a tree of random forest (R_g) to the bootstrapped data until the minimum node size n_{min} is attained.

Then, output the ensemble of trees $\{T_g\}_1^G$ in order to make a prediction at a new dot of q .

Let $M_g(q)$ be the class prediction of the g th of the random forest tree. Then, the random forest function is calculated by

$$M_{rf}^G(q) = \text{majority vote } \{M_g(q)\}_1^G \quad \text{Equation 13}$$

RF can give an estimate of important input variables in the classification and it runs efficiently on large dataset with high accuracy. However, random forest has some constraints on computing time and memory.

3.10.3 k – Nearest Neighbour (k-NN)

The k-NN implemented majority vote, which indicate

$$Y(x) = \frac{1}{k} \sum_{xi \in N_{k(x)}} y_i \quad \text{Equation 14}$$

The k-NN classifier necessitates three conditions:-

- (i) The set of stored records
- (ii) The distance metric in order to compute distance between records
- (iii) The value of k , which is the number of nearest neighbours to retrieve

There are three steps in order to classify the unknown records: -

- (i) Compute the distance to train other records
- (ii) Identify k nearest neighbours
- (iii) Determine the class label of unknown record by using class labels of nearest neighbours.

In this research, the k value is equal to 1. In order to compute the distance between two points of k -nearest neighbour classifier, the Euclidean distance was applied as followed

$$d(p, q) = \sqrt{\sum_i (p_i - q_i)^2}; \quad \text{Equation 15}$$

The class from nearest neighbour was determined by taking the majority vote of class labels among the k -nearest neighbours and the weight of the vote based on the distance as weight factor, $w = 1/d^2$

3.10.4 Support Vector Machine (SVM)

Furthermore, SVM was also tested in this study. SVM (Cortes & Vapnik, 1995) is a classification method that maps the input data to a higher dimensional feature space through some nonlinear transformation that are separated by optimal hyperplane, which maximises the gap of positive samples and negative samples. The term SVM was originated from points in the training set that are closest to the decision surface, which was named as support vector. Training set with instance label pairs is formulated in Equation 16

$$(x_k, y_k), \quad k = 1, 2, \dots, n, \text{ where } x_k \in \mathbb{R}^r \text{ and } y \in \{1, -1\}^n; \quad \text{Equation 16}$$

The SVM solve the solution of the following optimization step

$$\frac{1}{2} w^B w + C \sum_{k=1}^n \xi_k; \quad \text{Equation 17}$$

$$y_k(w^B \phi((x_k) + b) \geq 1 - \xi_k, \quad \xi_k \geq 0; \quad \text{Equation 18}$$

The vector x_k are higher dimensional space by the function ϕ . Then, SVM indicates a linear separating hyperplane with maximum margin in infinite dimensional space. $K(x_k, x_l) = \phi(x_k) \phi(x_l)$ is known as the kernel function.

Polynomial: $K(x_k, x_l) = (\gamma x_k^B x_l + f)^d, \gamma > 0$ Equation 19

In this research, SVM used the kernel function of polynomial to transform the input data into a higher dimensional space and optimal hyperplane is constructed with maximum margin. The classification involved 45 classes in this study; therefore SVM classifier was trained using the one versus all approach. In this approach, every class was trained with test cases of that class as positive and all other as negative.

3.10.5 Linear Discriminant Analysis (LDA)

Discriminant analysis is a usually used statistical method for classification to reduce the dimensionality of data. LDA is also known as the Fisher discriminant analysis (FDA) (Alpaydin, 2014). LDA was proposed to find a linear combination of features, which characterises or separates different classes. LDA maximises the ratio of between class variance to within class variance, thus achieving maximum discrimination.

LDA contains weights for each feature separately for every class that allows it to ignore features that have no significant meaning for some classes. In the research case of LDA, the covariance matrices are assumed to be equal. Suppose T is an J by K class membership matrix where if observation j is from class k , $T_{j,k} = 1$, otherwise $T_{j,k} = 0$.

The parameters of the Gaussian distribution are indicated the priori probability and formulated below

$$P(G=K) = \pi_k = \frac{J_k}{J} ; \quad \text{Equation 20}$$

Where J_k is the number of sample class k

The estimated of the class mean for the data is indicates in equation below

$$\mu_k = \frac{\sum_{j=1}^J T_{j,k} x_j}{\sum_{j=1}^J T_{j,k}}, \quad \text{Equation 21}$$

The unbiased estimate of the pool in covariance matrix is formulated

$$\Sigma = \frac{\sum_{j=1}^J \sum_{k=1}^K T_{j,k} (x_n - \mu_k)(x_n - \mu_k)^F}{J-K} \quad \text{Equation 22}$$

Then, the linear discriminant function is calculated by

$$d_k^R(x) = \mu_k^R \Sigma^{-1} x - \mu_k^R \Sigma^{-1} x \mu_k^R + \log(\pi_k); \quad \text{Equation 23}$$

3.10.6 Directed Acyclic Graph Multiclass Least Square Twin Support Vector Machine (DAG MLSTSVM)

DAG MTSVM uses direct acyclic graph structure in order to arrange sub classifiers. According to Gu et al. (2014) and Tomar and Agarwal (2015) a directed acyclic graph is a finite directed graph with directed cycles. There are many nodes of a rooted binary directed acyclic graph and each node has zero two arcs leaving it. The combinations of least square twins support vector machine and directed acyclic graph were proposed by Tomar and Agarwal (2015). The proposed approach retains the characteristics of DAG MTSVM, and the least squares loss makes the training process faster.

In this research, the DAG MLSTSVM was based on “one-versus-one” algorithm. The approach forms a binary classifier for each pair of classes in order to deal with the k-class classification problem. The method builds $k(k-1)/2$ binary twin SVM which can classify two classes directly. The sub classifier of “one-versus-one” DAGMLSTSVM for the n -th and m -th class is constructed by solving the equation below.

$$\text{Min } \frac{1}{2} \| A_n w_{nm} + e_{nm}^{(1)} b_{nm} \|^2 + \frac{c_{nm}}{2} e_{nm}^{(2)T} \xi_{nm}$$

$$\text{Where, } (A_m w_{nm} + e_{nm}^{(2)} + \xi_{nm} \geq e_{nm}^{(2)}, \xi_{nm} \geq 0 \text{ and}$$

Equation 24

$$\text{Min } \frac{1}{2} \| A_n w_{nm} + e_{nm}^{(2)} b_{nm} \|^2 + \frac{c_{nm}}{2} e_{nm}^{(1)T} \xi_{nm}$$

$$\text{Where, } (A_m w_{nm} + e_{nm}^{(1)} + \xi_{nm} \geq e_{nm}^{(1)}, \xi_{nm} \geq 0$$

According to Ding et al. (2017) the “one-versus-one” strategy based on DAG MLSTSVM is always better than “one-versus-more” SVM can solve a linear equation problem instead of using quadratic programming in the multiclass classification, which can lead to higher computational costs.

3.11 Cross-validation (CV)

Cross validation (CV) is a popular strategy for model selection, and more generally algorithm selection. According to Artlot and Celisse (2010), the main idea behind CV is to split the data (once or several times) for estimating the risk of each algorithm, which some part of the data (the training sample) is used for training each algorithm, and the remaining part (the validation sample) is used for estimating the risk of the algorithm. Then, CV selects the algorithm with the smallest estimated risk. If compared to the re-substitution error, CV avoids over fitting because the training sample is independent from the validation sample.

Generally, the k-fold and leave-one-out are the popular and benchmarking methods for CV. The main purpose of k-fold is partitioning the dataset into k subsets randomly and leave-one-out approach is used leave out for testing in every iteration and others are used for training. The cross-validation approach was applied in the myDAUN dataset that consist of 30 samples of each tropical shrub species in order to reduce the over fitting. Two approaches of CV were applied which are 5-fold and 10-fold.

3.12 Summary

This chapter summarises the overview of the proposed solution. The materials and methods used for the development of automated classification of tropical shrub species were described. This chapter presents an overview of the current research and explained the experimental setup including field sampling, image acquisition, image pre-processing, feature extraction and classification.

This study introduced the dataset named as myDAUN, which currently stored 45 species of tropical shrubs and all of these data collection were sampled from four main locations, which are Faculty of Science, Tunku Canselor Hall, Varsity Lake and Main Library. Four types of shape descriptors were applied which were MSD, HOG, Hu and ZM. All of the data of extracted features were tested in the classification step. The classifiers used are ANN, SVM, RF, k-NN, RF and DAG MLSTSVM. Based on the literature review, this is the first study in the development of tropical shrub species image dataset and classification using a hybrid of leaf shape and machine learning approach.

CHAPTER 4: RESULTS

4.1 Introduction

This section covers the results achieved in this research. Several experiments were executed, accuracy of various sets of descriptors were measured, which are single and hybrid of two, three and four descriptors using myDAUN dataset. The feature selections are applied in the proposed method into three categories which are 50%, 60% and 70% descriptors. In addition, the proposed method was validated with cross validation and the most popular benchmark datasets, which are Flavia and Swedish Leaf dataset.

4.2 Results of Data Collection

As mentioned in Chapter 3, four main locations with variety of tropical shrub species were chosen and selected. Table 4.1 shows the number of tropical shrub that has been stored in myDAUN dataset. The four sampling locations took place in four main locations which are the Faculty of Science, Dewan Tunku Canselor, Varsity Lake and Main Library. Faculty of Science achieved the highest number of shrubs, followed by Tunku Canselor Hall. These results showed that the Faculty of Science has more number of shrubs compare to other locations

Table 4.1: Data collection of myDAUN dataset

Location	Faculty of Science	Tunku Canselor Hall	Varsity Lake	Main Library
Number of shrub	31	8	3	3

4.3 Feature Extraction Methods

Various combination sets of descriptors extracted from the myDAUN dataset were tested. These descriptors include single descriptor and hybrid of two, three and four descriptors. Single descriptor consists of only one descriptor, hybrid of two consists of two types of descriptors, hybrid of three consists of three types of descriptors and hybrid of four consists of four types of descriptors. A total of 133 features were extracted from myDAUN dataset, which included 20 features from MSD descriptor, 81 features from HOG descriptors, 7 features from Hu and 25 features from ZM. The classification methods were implemented on all of the combinations of the descriptors and Table 4.2 lists the combinations of the descriptors for feature extraction.

Table 4.2: Combinations of descriptors for feature extraction

Methods	Descriptor
Single descriptor	MSD
	HOG
	Hu
	ZM
Hybrid of two descriptors	{MSD + HOG}
	{MSD + Hu}
	{MSD + ZM}
	{HOG + Hu}
	{HOG + ZM}
	{Hu + ZM}
Hybrid of three descriptors	{MSD + HOG + Hu}
	{MSD + HOG + ZM}
	{MSD + Hu + ZM}
	{HOG + Hu + ZM}
Hybrid of all descriptors	{MSD + HOG + Hu + ZM}

4.4 Classifiers

In this section, the features extracted from myDAUN dataset are tested using six types of classifier models, which are artificial neural network (ANN), random forest (RF), support vector machine (SVM), k-nearest neighbour (k-NN), linear discriminant analysis (LDA) and direct acyclic graph multi least square twins support vector machine (DAG MLSTSVM). The myDAUN dataset was set randomly with 80% for training and 20% for testing, in which 1080 of sample images were for training and 270 of sample images were for testing. The classification process was repeated for 10 times and average accuracy of 10 runs was taken as final result.

4.4.1 Artificial Neural Network (ANN)

ANN that was implemented in this research is the feed forward neural network, which is the most common type of ANN. The feed forward neural network was trained using the scaled conjugate gradient algorithm and with random data division. Table 4.3 shows the results generated to determine the best number of neurons for feed forward neural network classification. One hidden layer with 10 to 100 neurons with three different sets of data division for training and testing were tested in order to determine the best neurons. In this research, one hidden layer with 60 neurons (achieved the best result) with 80% of training and 20% of testing achieved the best result. Thus, it was used throughout this research. The average classification of accuracy for 10 runs was taken.

Table 4.3: Testing accuracy of three set data division with various set of neurons

Neurons	Accuracy (%)		
	(60,40)	(70,30)	(80,20)
10	79.46	81.82	83.17
20	81.34	82.11	83.71
30	82.90	84.15	84.57
40	82.44	85.56	85.83
50	82.71	85.73	83.97
60	84.86	85.97	86.21
70	84.54	85.44	84.57
80	83.84	85.30	84.51
90	83.46	84.91	85.27
100	83.92	84.91	85.62

a) Single Descriptor

Each of the descriptors method, which were MSD, HOG, Hu and ZM were tested individually. The average of classification accuracy of feed forward neural network using single descriptors by using ANN classifier is shown in Table 4.4. The ANN accuracy for MSD descriptor achieved the highest result followed by HOG descriptor. Hu obtained the lowest accuracy for single descriptor with an accuracy of 82.27%.

Table 4.4: Classification accuracy for ANN classifier using single descriptor

Descriptor	Average accuracy of ANN** (%)
MSD	96.39
HOG	95.82
Hu	82.27
ZM	91.79
**Average accuracy = average of 10 runs	

b) Hybrid of Descriptors

The average of classification accuracies of feed forward neural network using hybrid of two, three and four descriptors, were shown in Table 4.5. The hybridisation of two descriptors of {MSD + HOG} achieved the highest accuracy of 97.49%, followed by hybridisation of {MSD + Hu}, {MSD + ZM}, {HOG + Hu}, {HOG + ZM} and {Hu + ZM} which obtained 96.67%, 96.60%, 96.24%, 93.70% and 93.6% respectively.

Whereas, the hybridisation of three descriptors for ANN classifier showed that, the accuracy of {MSD + HOG + ZM} achieved the highest accuracy of 97.63%, followed by hybridisation of {MSD + HOG + Hu} which obtained 97.59%. Furthermore, {MSD + Hu + ZM} and {HOG + Hu + ZM} obtained 96.64% and 97.06% respectively. As shown in Table 4.5 the hybridisation of two and three descriptors achieved almost comparable results.

Next, by using ANN classifier, the accuracy of combination of all descriptors of {MSD, HOG, Hu + ZM} obtained the highest accuracy of 98.23%, if compared to the accuracy achieved by using single, hybrid of two and three descriptors. By comparing the classification accuracy in Table 4.4 and Table 4.5, it is clear that the hybridisation of descriptors improved the classification accuracy. Figure 4.1 summarises the graphical comparison of ANN accuracy on myDAUN dataset with various combination sets of descriptors.

Table 4.5: Classification for ANN classifier using hybrid of descriptors

Methods	Descriptor	*Average accuracy of ANN (%)
Hybrid of two descriptors	{MSD + HOG}	97.49
	{MSD + Hu}	96.67
	{MSD + ZM}	96.60
	{HOG + Hu}	96.24
	{HOG + ZM}	93.70
	{Hu + ZM}	93.67
Hybrid of three descriptors	{MSD + HOG + Hu}	97.59
	{MSD + HOG + ZM}	97.63
	{MSD + Hu + ZM}	96.64
	{HOG + Hu + ZM}	97.06
Hybrid of all descriptors	{MSD + HOG + Hu + ZM}	98.23
**Average accuracy = average of 10 runs		

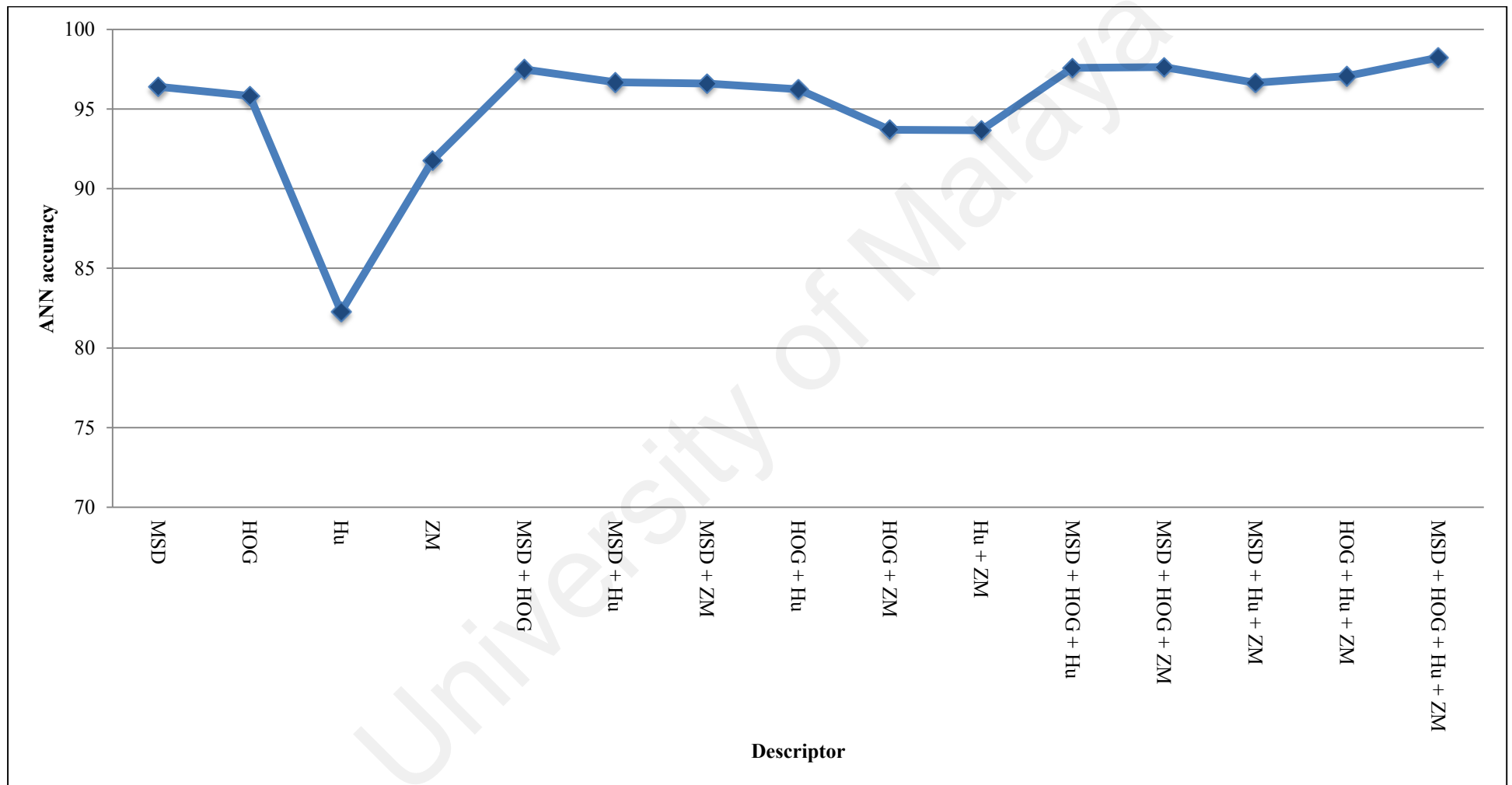


Figure 4.1: Comparison of ANN accuracy with various sets of descriptors

4.4.2 Random Forest (RF)

The RF, or also known as *TreeBagger* in Matlab, was implemented in this research. The “Random” word comes from the term “bootstrap aggregating”, or “bagging”, which means that each tree within the forest only gets to train on some subset of the full training dataset. Then, the elements of the training data for each tree are held “out-of-bag” for estimation of accuracy. The randomness also helped and assisted in order to decide which feature input variables are seen at each node in each decision tree. In this research the decision tree was trained using bagging algorithm and with random data division.

a) Single Descriptor

Each of the descriptors method, which were MSD, HOG, Hu and ZM were tested individually. The average of classification accuracy of decision tree using single descriptors by using RF classifier was shown in Table 4.6. The RF accuracy for MSD descriptor achieved the highest result of 92.58% followed by HOG descriptor which achieved 91.58%. Hu obtained the lowest accuracy for single descriptor with 83.36%.

Table 4.6: Classification accuracy for RF classifier using single descriptor

Descriptor	Average accuracy of RF** (%)
MSD	92.58
HOG	91.58
Hu	83.36
ZM	87.85
**Average accuracy = average of 10 runs	

b) Hybrid of Descriptors

The average of classification accuracy of RF using hybrid of two, three and four descriptors are shown in Table 4.7. For the hybridisation of two descriptors by using RF classifier, the combination of {MSD + HOG} achieved the highest accuracy of 93.45%, followed by hybridisation of {MSD + ZM} with 92.86%.

The hybridisation of three descriptors of {MSD + HOG and Hu} by using RF achieved the highest accuracy of 93.62%, followed by hybridisation of {MSD + HOG + ZM}, {HOG, Hu and ZM} and {MSD + Hu + ZM}, which obtained 93.52%, 93.37% and 93.24% respectively. As shown in Table 4.9 the hybridisation of two and three descriptors achieved almost comparable results for RF classifier.

In addition, the combination of all descriptors of {MSD, HOG, Hu + ZM} obtained the highest result of 93.83% if compared to the accuracy by using of single, hybrid of two and three descriptors. By comparing the classification accuracy in Table 4.6 and Table 4.7, it is clear that the hybridisation of descriptors improved the classification accuracy. Figure 4.2 summarises the graphical comparison of RF accuracy on myDAUN dataset with various combination sets of descriptors.

Table 4.7: Classification for RF classifier using hybrid of descriptors

Methods	Descriptor	Average accuracy of RF ** (%)
Hybrid of two descriptors	{MSD + HOG}	93.45
	{MSD + Hu}	92.84
	{MSD + ZM}	92.86
	{HOG + Hu}	92.39
	{HOG + ZM}	92.58
	{Hu + ZM}	90.07
Hybrid of three descriptors	{MSD + HOG + Hu}	93.62
	{MSD + HOG + ZM}	93.52
	{MSD + Hu + ZM}	93.24
	{HOG + Hu + ZM}	93.37
Hybrid of all descriptors	{MSD + HOG + Hu + ZM}	93.83
**Average accuracy = average of 10 runs		

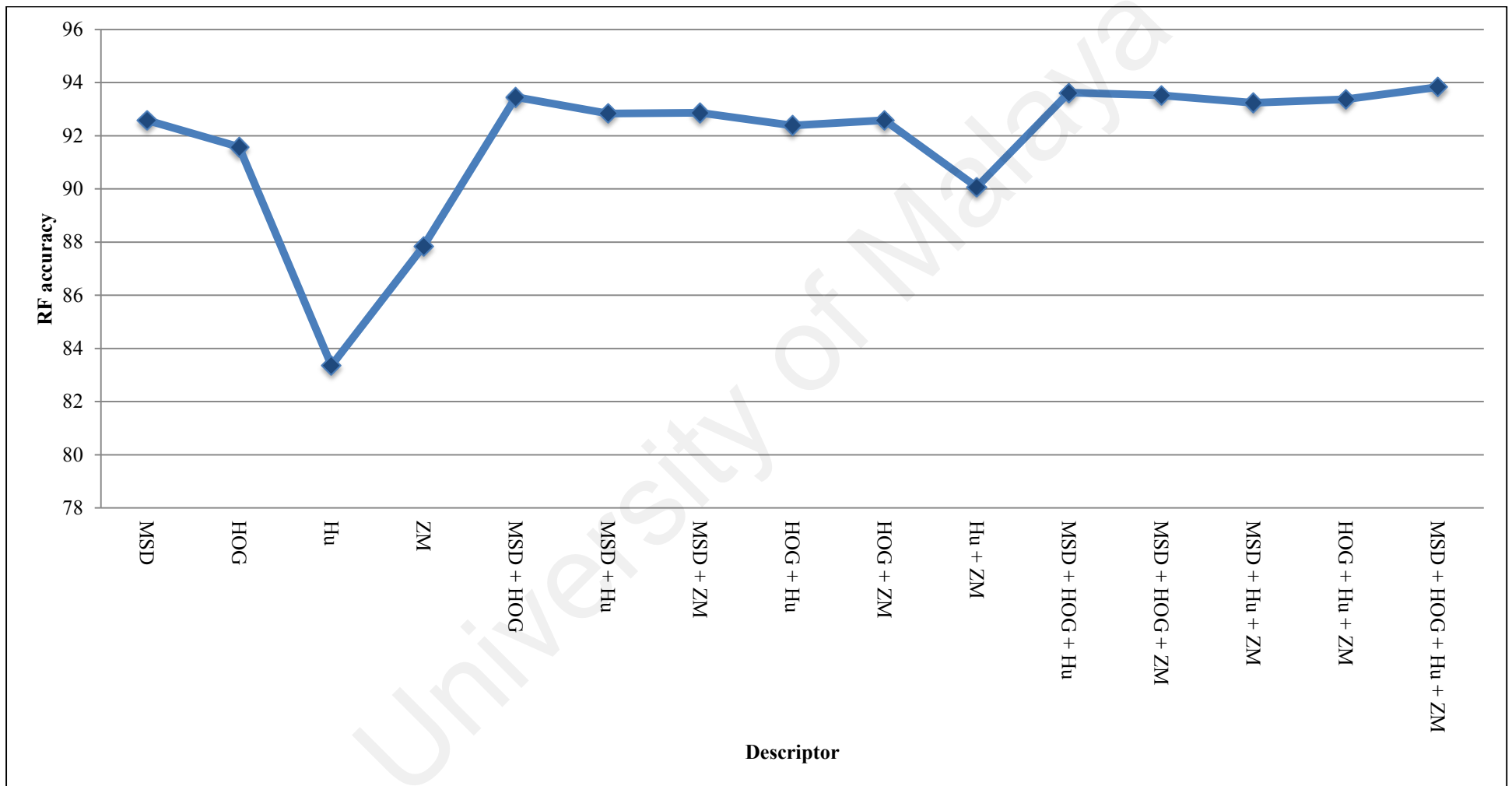


Figure 4.2: Comparison of RF accuracy with various sets of descriptors

4.4.3 Support Vector Machine (SVM)

The SVM tool that was used is the fit multiclass model, which trains an error-correcting output codes (ECOC) multiclass using SVM binary learners. Then trains a one-versus-all ECOC classifier using the ensembles of decision trees as binary learners.

a) Single Descriptor

MSD, HOG, Hu and ZM were tested individually. The average of classification accuracy of SVM using single descriptors was shown in Table 4.8. The SVM accuracy for HOG descriptor achieved the highest result of 84.53% followed by MSD descriptor with 79.78%. Hu obtained the lowest accuracy for single descriptor with 32.74% only.

Table 4.8: Classification accuracy for SVM classifier using single descriptor

Descriptor	Average accuracy of SVM** (%)
MSD	79.78
HOG	84.53
Hu	32.74
ZM	59.34
**Average accuracy = average of 10 runs	

b) Hybrid of Descriptors

The average of classification accuracy of SVM using hybrid of two, three and four descriptors are shown in Table 4.9. The hybridisation of two descriptors of {MSD + HOG} achieved the highest accuracy of 91.01%, followed by hybridisation of {HOG + ZM}, {HOG + Hu}, {MSD + ZM}, {MSD + Hu} and {Hu + ZM} which were 89.93%, 87.72%, 84.61%, 81.99% and 73.45%.

On other hand, the hybridisation of three descriptors by using SVM achieved the highest accuracy of 92.06% for combination of {MSD + HOG and ZM}. Then, followed by hybridisation of {MSD + HOG + Hu}, {HOG, Hu and ZM} and {MSD + HOG + ZM}, which obtained 91.78%, 91.29% and 87.93% respectively. As shown in Table 4.9 the hybridisation of two and three descriptors achieved almost comparable results for SVM classifier.

Then, the average of classification accuracy of SVM using four descriptors, the combination of {MSD +HOG + Hu + ZM} obtained the highest result of 92.74% if compared to the accuracy by using of single, hybrid of two and three descriptors.

By comparing the classification accuracy in Table 4.8 and Table 4.9, it is clear that the hybridisation of descriptors had improved the classification accuracy by using SVM classifier. Figure 4.3 illustrates the graphical comparison of SVM accuracy on myDAUN dataset with various combination sets of descriptors.

Table 4.9: Classification for SVM classifier using hybrid of descriptors

Methods	Descriptor	Average accuracy of SVM ** (%)
Hybrid of two descriptors	{MSD + HOG}	91.01
	{MSD + Hu}	81.99
	{MSD + ZM}	84.61
	{HOG + Hu}	87.72
	{HOG + ZM}	89.93
	{Hu + ZM}	73.45
Hybrid of three descriptors	{MSD + HOG + Hu}	91.78
	{MSD + HOG + ZM}	92.06
	{MSD + Hu + ZM}	87.93
	{HOG + Hu + ZM}	91.29
Hybrid of all descriptors	{MSD + HOG + Hu + ZM}	92.74
**Average accuracy = average of 10 runs		

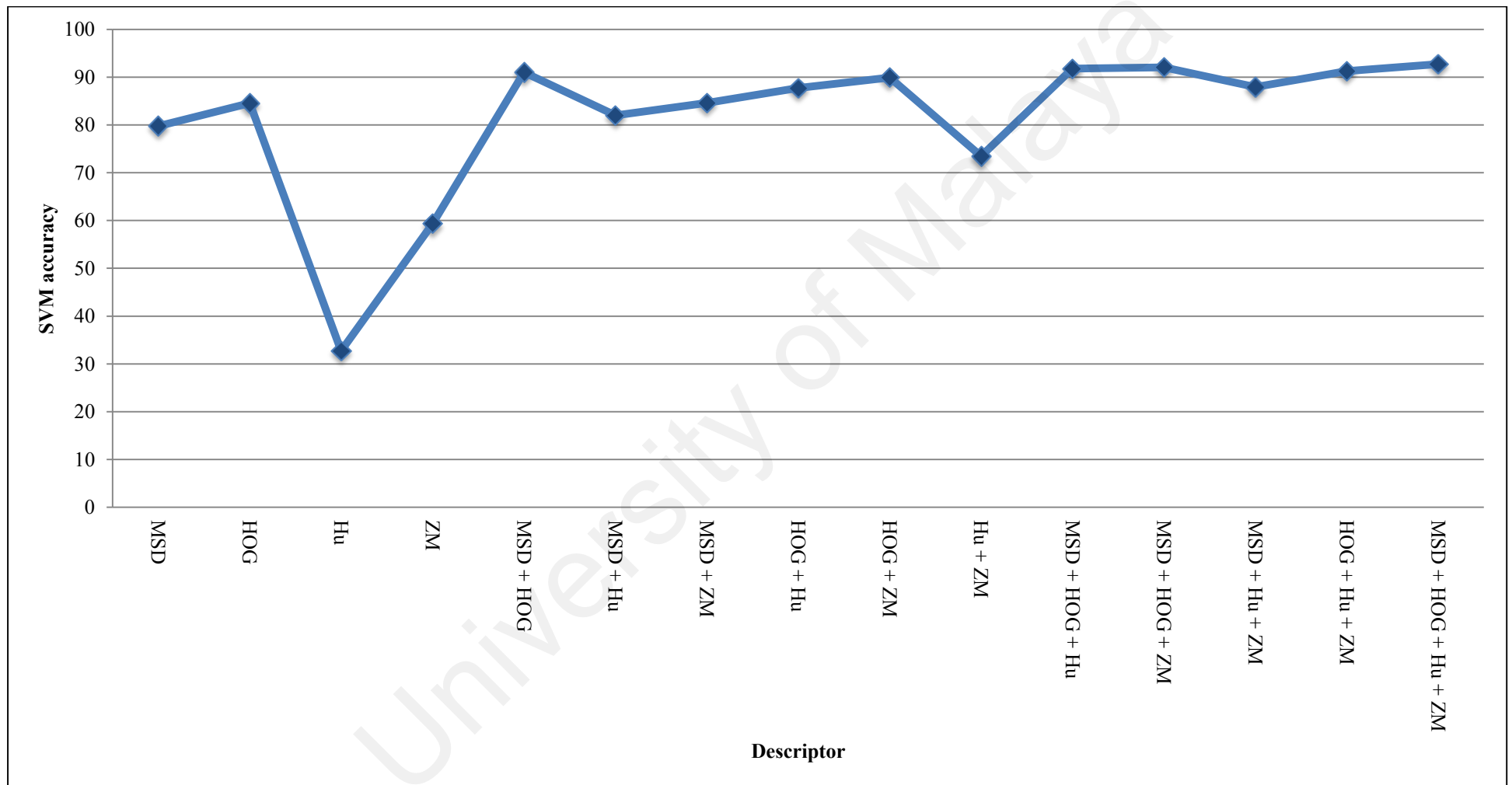


Figure 4.3: Comparison of SVM accuracy with various sets of descriptors

4.4.4 k-Nearest Neighbour (k-NN)

Table 4.16 shows the results generated to determine the best number of neighbour for k-NN classification. In this experiment, k-NN classifier with $k = 1, 2, 3, 4$ and 5 were simulated and as shown in table 4.10, the case of $k = 1$ showed the best classification result. The highest accuracy was obtained with $k = 1$ and the accuracy decreased as the number of neighbours increased.

Table 4.10: Accuracy of number of nearest neighbour

Number of nearest neighbours	Accuracy ** (%)
1	91.55
2	85.19
3	84.74
4	83.56
5	83.48
**Average accuracy = average of 10 runs	

a) Single Descriptor

The descriptors of MSD, HOG, Hu and ZM were tested individually by using k-NN classifier. The average of classification accuracies of k-NN using single descriptors are shown in Table 4.11. The k-NN accuracy for MSD descriptor achieved the highest accuracy of 91.96% followed by HOG descriptor which achieved 90.40%. The lowest accuracy for single descriptor was obtained by Hu with 82.99%.

Table 4.11: Classification accuracy for k-NN classifier using single descriptor

Descriptor	Average accuracy of k-NN** (%)
MSD	91.96
HOG	90.40
Hu	82.99
ZM	87.75
**Average accuracy = average of 10 runs	

b) Hybrid of Descriptors

The average of classification accuracy of k-NN using hybrid of two, three and four descriptors are shown in Table 4.12. The hybridisation of two descriptors by using k-NN classifier, showed that the accuracy of the combination of {MSD + Hu} achieved the highest accuracy of 92.35%, followed by hybridisation of {MSD + HOG} with 92.03%.

The average of classification accuracy of k-NN using hybrid of three descriptors achieved the highest accuracy of 92.42% with combination of {MSD + HOG + Hu}, followed by hybridisation of {MSD + HOG + ZM}, {MSD, Hu + ZM} and {HOG + Hu + ZM}, which obtained 92.17%, 92.10% and 91.56% respectively. As shown in Table 4.12, the hybridisation of three descriptors achieved better results than two descriptors for k-NN classifier.

Next, the accuracy of combination of all descriptors obtained the highest result of 92.60% if compared to the accuracy achieved by using the single, hybrid of two and three descriptors. By comparing the classification accuracy in Table 4.11 and Table 4.12, it is clear that the hybridisation of descriptors improved the classification accuracy by using k-NN classifier. Figure 4.4 illustrates graphical comparison of k-NN accuracy on myDAUN dataset with various combination sets of descriptors.

Table 4.12: Classification for k-NN classifier using hybrid of descriptors

Methods	Descriptor	Average accuracy of k-NN ** (%)
Hybrid of two descriptors	{MSD + HOG}	92.03
	{MSD + Hu}	92.35
	{MSD + ZM}	91.92
	{HOG + Hu}	91.07
	{HOG + ZM}	91.47
	{Hu + ZM}	89.35
Hybrid of three descriptors	{MSD + HOG + Hu }	92.42
	{MSD + HOG + ZM}	92.17
	{MSD + Hu + ZM}	92.10
	{HOG + Hu + ZM}	91.56
Hybrid of all descriptors	{MSD + HOG + Hu + ZM}	92.60
**Average accuracy = average of 10 runs		

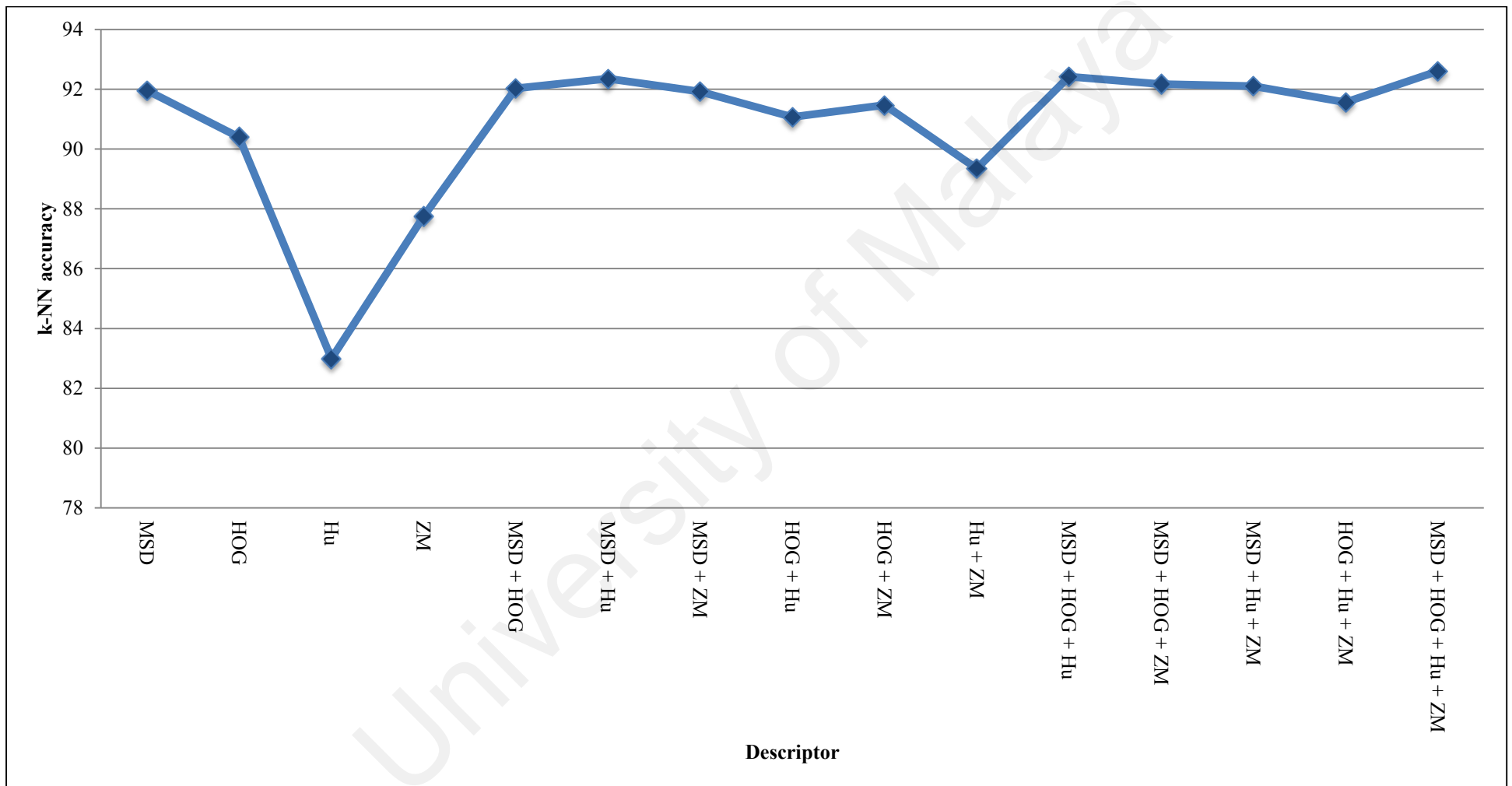


Figure 4.4: Comparison of k-NN accuracy with various sets of descriptors

4.4.5 Linear Discriminant Analysis (LDA)

The LDA is one of the discriminant analysis is a classification method, also known as the Fisher discriminant and it assumes that different classes generate data based on different Gaussian distributions. There are two ways for LDA classification method, first to train a classifier, the fitting function estimates the parameters of a Gaussian distribution for each class and then to predict the classes of new data, the trained classifier finds the class with the smallest misclassification cost

a) Single Descriptor

The single descriptor of MSD, HOG, Hu and ZM were tested individually by using LDA classifier. The average of classification accuracies of LDA using single descriptors is shown in Table 4.13. The LDA accuracy for MSD descriptor achieved the highest result of 82.80% followed by HOG descriptor of 79.76%. The lowest accuracy for single descriptor was obtained by Hu with 37.65% only.

Table 4.13: Classification accuracy for LDA classifier using single descriptor

Descriptor	Average accuracy of LDA** (%)
MSD	82.80
HOG	79.76
Hu	37.65
ZM	56.40
**Average accuracy = average of 10 runs	

b) Hybrid of Descriptors

The average of classification accuracy of LDA using hybrid of two descriptors, showed that the combination of {MSD + HOG} achieved the highest accuracy of 89.56%, followed by hybridisation of {HOG + ZM} with 85.58%. The results of LDA using hybrid of three descriptors are shown in Table 4.23. The accuracy of {MSD + HOG + Hu} achieved the highest accuracy of 90.09%, whereas the hybridisation of {MSD + HOG + ZM}, {HOG + Hu + ZM} and {MSD + Hu + ZM} obtained 89.72%, 87.23% and 86.12% respectively. As shown in Table 4.14, the results of the hybridisation of two and three descriptors achieved are almost comparable for LDA classifier.

In addition, table 4.14 shows the average of classification accuracy of LDA using all four descriptors. By using LDA classifier, the accuracy of combination of all descriptors of obtained the highest result of 90.86%, if compared to the accuracy by using of single, hybrid of two and three descriptors.

By comparing the classification accuracy in Table 4.13 and Table 4.14, it is clear that the hybridisation of descriptors improved the classification accuracy by using LDA classifier. Figure 4.5 illustrates graphical comparison of LDA accuracy on myDAUN dataset with various combination sets of descriptors

Table 4.14: Classification for LDA classifier using hybrid of descriptors

Methods	Descriptor	Average accuracy of LDA ** (%)
Hybrid of two descriptors	{MSD + HOG}	89.56
	{MSD + Hu}	85.14
	{MSD + ZM}	84.31
	{HOG + Hu}	83.47
	{HOG + ZM}	85.58
	{Hu + ZM}	68.31
Hybrid of three descriptors	{MSD + HOG + Hu}	90.09
	{MSD + HOG + ZM}	89.72
	{MSD + Hu + ZM}	86.12
	{HOG + Hu + ZM}	87.23
Hybrid of all descriptors	{MSD + HOG + Hu + ZM}	90.86
**Average accuracy = average of 10 runs		

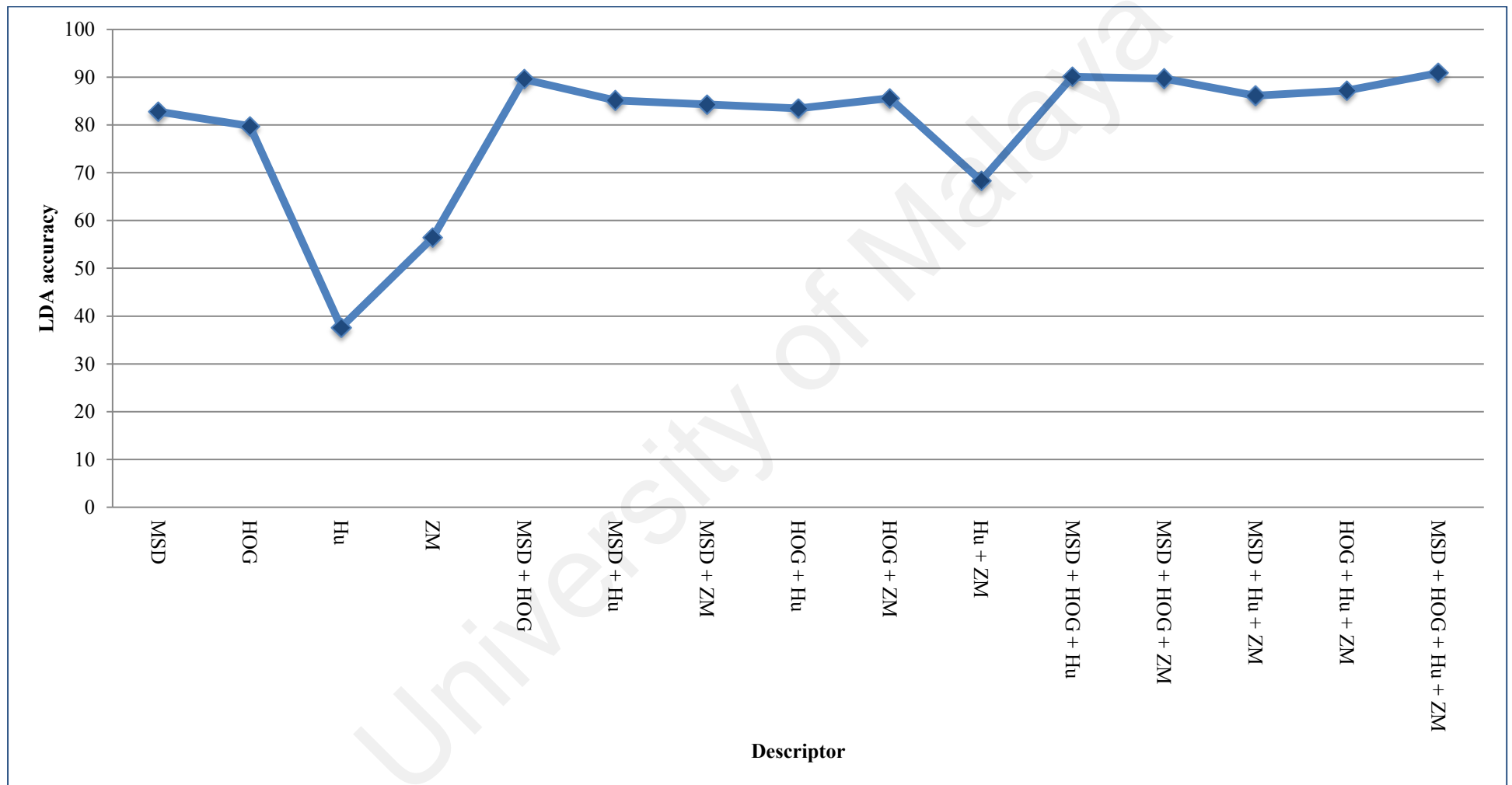


Figure 4.5: Comparison of LDA accuracy with various sets of descriptors

4.4.6 Directed Acyclic Graph Multiclass Least Squares Twin Support Vector Machine (DAG MLSTSVM)

The DAG MLSTSVM is a novel machine-learning algorithm that developed from traditional SVM. It is one of the typical non-parallel SVM. Since the DAG MLSTSVM has superiorities of the simple model, the high training speed and the good performance, it has drawn extensive attention.

a) Single Descriptor

The results of DAG MLSTSVM using single descriptors are shown in Table 4.15. The DAG MLSTSVM accuracy for MSD descriptor achieved the highest accuracy of 95.78% followed by HOG descriptor which achieved 95.40%. The lowest accuracy for single descriptor was obtained by Hu with 85.76%.

Table 4.15: Classification accuracy for DAG MLSTSVM classifier using single descriptor

Descriptor	Average accuracy of DAG MLSTSVM** (%)
MSD	95.78
HOG	95.40
Hu	85.76
ZM	90.54
**Average accuracy = average of 10 runs	

b) Hybrid of Descriptors

The average of classification accuracy of DAG MLSTSVM using hybrid of two descriptors achieved the highest accuracy of 96.94% with {MSD + HOG}, followed by hybridisation of {HOG + Hu} with 96.88%. The results of DAG MLSTSVM using hybrid of three descriptors are shown in Table 4.16. The accuracy of {MSD + HOG + ZM} achieved the highest accuracy of 97.05%, followed by hybridisation of {MSD + HOG + Hu}, {HOG + Hu + ZM}, {MSD + Hu + ZM} which obtained 96.99%, 96.70% and

96.32% respectively. As shown in Table 4.16, the hybridisation of two and three descriptors achieved almost comparable results for DAG MLSTSVM classifier.

The combination of all descriptors achieved the highest accuracy of 97.72% if compared to the accuracy by using of single, hybrid of two and three descriptors. By comparing the classification accuracy in Table 4.15 and Table 4.16, it is clear that the hybridisation of descriptors improved the classification accuracy by using DAG MLSTSVM classifier. Figure 4.6 illustrates graphical comparison of DAG MLSTSVM accuracy on myDAUN dataset with various sets of descriptors.

Table 4.16: Classification for DAG MLSTSVM classifier using hybrid of descriptors

Methods	Descriptor	Average accuracy of DAG MLSTSVM ** (%)
Hybrid of two descriptors	{MSD + HOG}	96.94
	{MSD + Hu}	95.96
	{MSD + ZM}	96.25
	{HOG + Hu}	96.88
	{HOG + ZM}	93.52
	{Hu + ZM}	92.65
Hybrid of three descriptors	{MSD + HOG + Hu}	96.99
	{MSD + HOG + ZM}	97.05
	{MSD + Hu + ZM}	96.32
	{HOG + Hu + ZM}	96.70
Hybrid of all descriptors	{MSD + HOG + Hu + ZM}	97.72
**Average accuracy = average of 10 runs		

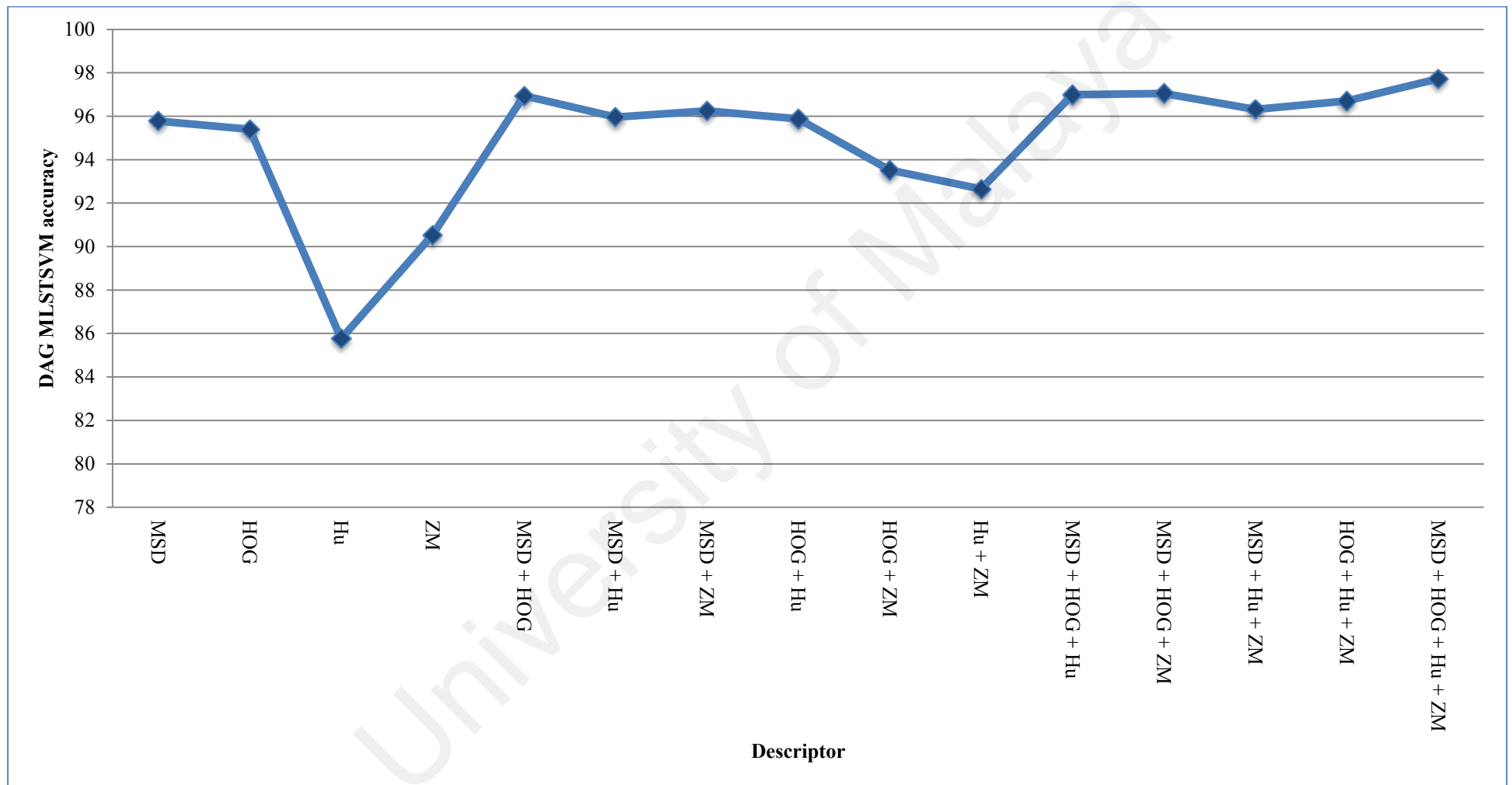


Figure 4.6: Comparison of DAG MLSTSVM accuracy with various sets of descriptors

4.5 Results of Feature Selection

As mentioned in the previous chapter, three feature selection methods were employed, which were Relief, Correlation-based feature selection (CFS) and Pearson's correlation coefficient (PCC). Feature selection was applied in the hybrid of all descriptors since it produced the best result in the classification part. Then, the descriptors were tested with three types of reduced models, which are reduced 50%, 60% and 70% of features from the full model of 133 features. The proposed feature selection methods ranked the feature input based on the categories of reduced models. Table 4.17 presents the total number of the feature input for all types of descriptors.

Table 4.17: Categories of descriptors reduction

Features reduced (%)	Feature inputs			
	MSD	HOG	Hu	ZM
50	10	41	4	13
60	8	32	3	10
70	6	24	2	8

The classification accuracy of the selected feature selection methods are shown in Table 4.18. The classification accuracy obtained using feature selection methods are between 96.98% and 98.13%. For the reduced model of 50%, CFS feature selection method achieved the highest result of 97.79% with 68 features only. Whereas, the Relief method achieved the highest result of 98.13% and 97.64% with reduced models of 60% and 70% with 53 and 40 descriptors respectively. In overall, the accuracy of Relief achieved the best result of 98.13% with reduced model of 60% (with 53 features), which was comparable with the result without feature selection method of 98.23% (with 133 features).

Table 4.18: Classification accuracy for the selected feature selection methods

Descriptor s	Reduced model (%)	*Relief (%)	*CFS (%)	*PCC (%)
Hybrid of all descriptors	50	97.69	97.79	97.33
	60	98.13	96.98	97.10
	70	97.64	97.10	97.15
	None	98.23%		
* = average of 10 runs				

In order to ensure that the efficacy of the feature selection by using Relief method, the computational time was measured for all feature extraction with and without Relief method, and the computational time for the feature extraction are reported in Table 4.19. The computational time for feature extraction of MSD, HOG, Hu and ZM were 84.01 minutes, 334.39 minutes, 225.00 minutes and 1620.00 minutes respectively. In total, the computational time for feature extraction of all descriptors was 2263.40 minutes.

In comparison, the computational time for the feature extraction using Relief method with 60% of reduced feature of MSD, HOG, Hu and ZM were 61.04 minutes, 334.39 minutes, 189.55 minutes and 748.25 minutes respectively. In total, the computational time for feature extraction with Relief was 1033.23 minutes. This result showed that the total computational time was reduced by 1230.17 minutes or 54.35% from the original full model by using feature selection of Relief.

Table 4.19: Running time for features extraction

Descriptors	Time for all feature extraction (minutes)	Time for feature extraction with Relief (minutes)
MSD	84.01	61.04
HOG	334.39	334.39
Hu	225.00	189.55
ZM	1620.00	748.25
Total	2263.40	1033.23

4.6 Cross-validation (CV)

The cross-validation (CV) was conducted in order to over fitting problems since myDAUN dataset consists of 30 samples per species. 5-fold CV and 10-fold CV were implemented with proposed method. The performance of CV was show in Table 4.20. As shown in Table 4.20, the performance of the proposed method without CV obtained 98.23%, which is relatively similar with that of both the 5-fold CV and the 10-fold CV. The 5-fold CV achieved 97.47% and the 10-fold CV achieved 97.84%.

Table 4.20: Validation result with cross-validation

Data partition of CV	Accuracy (%)
Without CV	98.23
5-fold CV	97.47
10-fold CV	97.84

4.7 Validation using Flavia and Swedish Leaf Dataset

The purpose of the validation is to test on the viability and applicability of using hybridisation of four descriptors for the classification of plant species in other datasets. In order to validate the proposed methods, the Flavia dataset Swedish Leaf dataset were used. Flavia dataset and Swedish Leaf dataset are currently the most often and popular benchmark datasets used by researchers to compare and evaluate method across studies. The validation applied the same settings as in myDAUN dataset with 80% for training and 20% for testing and used ANN as the classifier. This classifier was chosen because it obtains the highest accuracy in myDAUN dataset.

4.7.1 Flavia Dataset

The dataset of Flavia dataset contains 1907 leaf images from 32 different plant species with the number ranging from 50 to 77 in each species. In this study, 50 samples for each species were employed. Therefore, in total, there are 1600 images of leaf samples were used, of which 1280 for training and 320 for testing. Table 4.21 shows the average accuracy for various sets of descriptors in Flavia dataset. The accuracy of classification using single descriptor method, MSD and HOG method achieved more than 93% in the Flavia dataset. Whereas, Hu achieved the lowest accuracy, which was only 80.46% in the Flavia dataset. The highest accuracy for the hybrid of two, three and four descriptors increased slightly for Flavia dataset. The combination of all descriptors improved the classification accuracy and produced the best result for classification of plant species in the Flavia datasets. The results achieved in the Flavia datasets were comparable to those obtained in the myDAUN dataset.

Table 4.21: Classification results of Flavia dataset

Methods	Descriptor	* Average accuracy (%)
		Flavia
Single descriptor	MSD	93.30
	HOG	93.49
	Hu	80.46
	ZM	83.22
Hybrid of two descriptors	{MSD + HOG}	95.04
	{MSD + Hu}	93.12
	{MSD + ZM}	93.41
	{HOG + Hu}	93.55
	{HOG + ZM}	93.87
	{Hu + ZM}	88.47
Hybrid of three descriptors	{MSD + HOG + Hu}	94.01
	{MSD + HOG + ZM}	95.14
	{MSD + Hu + ZM}	93.67
	{HOG + Hu + ZM}	94.08
Hybrid of all descriptors	{MSD + HOG + Hu + ZM}	95.25
* = average of 10 runs		

4.7.2 Swedish Leaf Dataset

The Swedish Leaf dataset contains 15 different plant species of Swedish trees with 75 samples per species. Therefore, there are 1125 images of leaf samples in Swedish Leaf dataset that were used, of which 900 were for training and 225 for testing. Table 4.22 shows the average accuracy for various sets of descriptors in Swedish Leaf dataset. The accuracy of classification using single descriptor method, MSD and HOG method achieved more than 98% in the Swedish Leaf dataset. Whereas, Hu achieved the lowest accuracy, which was only 95.20 % in the Swedish Leaf dataset.

The highest accuracy for the hybrid of two, three and four descriptors increased slightly for Swedish Leaf dataset. The combination of all descriptors improved the classification accuracy and produced the best result for classification of plant species in the Swedish Leaf datasets with the accuracy of up to 99.89%. The results achieved in the Swedish Leaf datasets were comparable to those obtained in the myDAUN dataset.

Table 4.22: Classification results of Swedish Leaf dataset

Methods	Descriptor	* Average accuracy (%)
		Swedish Leaf
Single descriptor	MSD	98.65
	HOG	99.15
	Hu	95.20
	ZM	95.95
Hybrid of two descriptors	{MSD + HOG}	99.54
	{MSD + Hu}	98.37
	{MSD + ZM}	99.16
	{HOG + Hu}	99.24
	{HOG + ZM}	99.54
	{Hu + ZM}	98.01
Hybrid of three descriptors	{MSD + HOG + Hu}	99.43
	{MSD + HOG + ZM}	99.64
	{MSD + Hu + ZM}	99.16
	{HOG + Hu + ZM}	99.52
Hybrid of all descriptors	{MSD + HOG + Hu + ZM}	99.89
* = average of 10 runs		

4.8 Summary

This chapter provides the detailed results on the implementation of the proposed methods. Four types of descriptors; MSD, HOG, Hu and ZM were tested. They were grouped into four combination sets of descriptors, which are single, hybrid of two, hybrid of three and hybrid of four descriptors. The classification of tropical shrub species was conducted using six classifiers which are ANN, RF, SVM, k-NN, LDA and DAG MLSTSVM. The results showed that the hybrid of all descriptors of {MSD +HOG +Hu +ZM} stands out to be comparably better than single, hybrid of two and three descriptors with ANN classifier.

The feature selection method of Relief is a highly effective method for feature selection compared to CFS and PCC, which had reduced both in the number of dimensions of the dataset and the running time. The proposed method with 5-CV and 10-CV achieved a comparative accuracy result as compared to the proposed method without cross validation. Finally, this chapter presented the validation results of the proposed methods on the popular Flavia and Swedish Leaf dataset. Both datasets performed well with the proposed methods, in which the results obtained were comparable to those results in myDAUN dataset.

CHAPTER 5: DISCUSSIONS

5.1 Leaves

The topic of automated plant species identification mostly maneuvered by academics specialised in machine learning, computer vision and multimedia information retrieval (Wäldchen & Mäder, 2017). Solely a few researches are performed by interdisciplinary group of computer scientist and botanist. Progressively, research in the field of automated plant species classification and identification is moving towards more interdisciplinary seek. Potent collaboration between people from different background and disciplines is important and necessary in order to gain the benefits of joined and bonded research activities and to evolve widely accepted approaches (Bridle et al., 2013). In the meantime, botanist can learn from computer science approaches and vice versa.

In this study, the proposed approaches for tropical shrub species classification are based on the analysis of leaves. The reasons for focusing on leaves is because of leaves are available and handy for analysis throughout most of the year. Leaves are easy to find and to collect at everywhere in all seasons, and leaves can simply be imaged compared to other plant morphological structures such as flowers, fruits and barks (Cope et al., 2012). These characteristics make the data acquisition process simple and easy.

Generally, the typical ways often utilise flowers and fruits in order to characterise plant species. However, both of these are usually only available and obtainable for a few weeks of the year during the blooming season (Chaki et al., 2015a; Tomar & Agarwal, 2016). Only a few studies proposed to identify species merely based on flowers (Apriyanti et al., 2013; Cho, 2012). Machine learning based on flower classification is one of the challenging tasks in computer vision (Cho, 2012) since flowers and fruits are complex

3D objects and they vary in scale and viewpoint of flower and fruit images compared to leaf images as leaves are virtually 2D in shape.

5.2 Leaf Shape

Among the previous studies, leaf shape analysis has received the most distant attention in automated plant classification. Leaf shape is more heritable although species' leaves contrast in details; the differences across species are usually recognizable to human. Mostly, text-based taxonomic keys include leaf shape for recognition. As stated in Nilsback and Zisserman (2006), they considered shape analysis of flowers for plant species identification, the shape of individual petals, their configuration and the overall shape of a flower is observed and analysed in order to differentiate species. However, the shape of the same flower appears to vary due to the softness and flexibility of the petals making them easy to curl, bend or damaged. Additionally, a flower's shape normally changes where the petals even fall off. Moreover, some of the plant species have a very tiny size of flower and most of the shrub is evergreen and do not have flower.

Although the leaves can be recognised based on their colour, however the colour is not expected to be as discriminative as shape for leaf analysis since most of the leaves are coloured in various shades of colours (Yanikaglo et al., 2014). Example, leaf sample 2 (*Acalypha wilkesiana*) and leaf sample 18 (*Loropetalum chinensis*). *A. wilkesiana* is an evergreen shrub and its pointed oval leaves are coppery green, mottled and streaked with copper, red and purple (see Figure 5.1). Whereas, *L. chinensis* is commonly known as the Chinese fringe flower, has leaves varying from bronze-red when new, to olive green or burgundy when mature, which depends on selection and growing conditions (see Figure 5.2). Thus, colour is not expected to be as discriminative as shape for leaf analysis.



Figure 5.1: A variety of leaf samples of *Acalypha wilkesiana* species



Figure 5.2: A variety of leaf samples of *Loropetalum chinensis* species

The leaf margin can also be used for automated plant species identification, but they may not be present in all of plant species, and the teeth can be easily ruptured or damaged before and after the specimen collection. Thus, it is challenging to obtain quantitative margin measurement automatically (Corney et al., 2012). Therefore, this study proposed the leaf shape features for automated classification of tropical shrub species.

Furthermore, leaf shape is the simplest and easiest aspect for feature extraction because leaf can easily be separated from a uniform background. Based on the primary studies, the analysed leaf images were taken under simplified conditions, which is one mature leaf per image on uniform background. The segmentation of the leaf in the natural background is quite striking because it has an overlapping of the green element in the background.

Thus, the data acquisition for myDAUN dataset followed the standard criteria, which is to select a leaf that has matured, not ruptured or damaged. Then, the samples were compressed and the images were taken in the same standard with uniform background. The shape descriptors are classified into two categories; contour-based and region-based. In this study, region-based methods were implemented as shape descriptors since these methods are more robust as they use the entire shape information.

In addition, these methods have capability to cope well with shape defections that emerge because of missing shape part or occlusion. The major challenge for contour-based methods is the difficulty of self-intersection. Self-intersection happens commonly with lobed leaves for example the leaf samples of species 21 (*Manihot esculenta*) (refer to Figure 5.3).

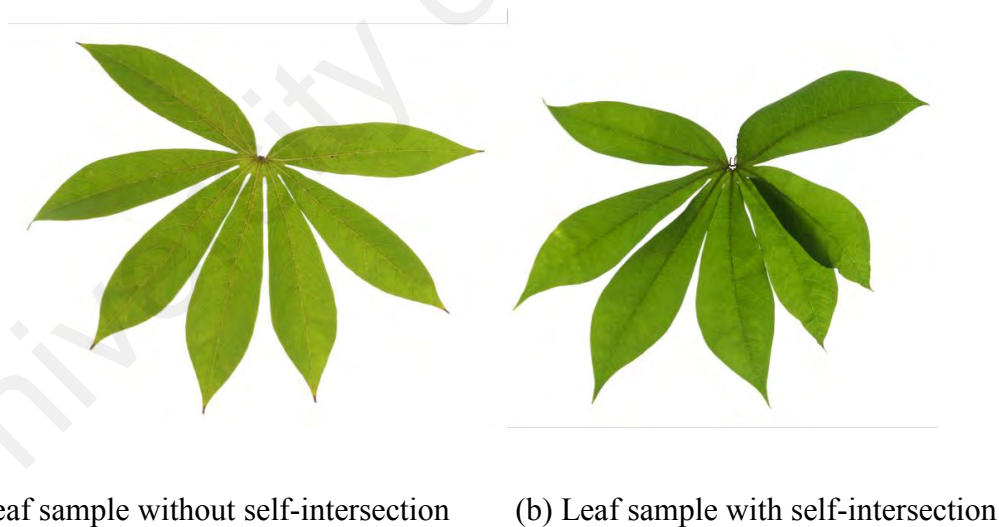


Figure 5.3: A *Manihot esculenta* leaf sample

The self-intersection is a problem where a leaf overlaps with other parts of the same leaf sample and can result in error when tracing the outline. The performance of the contour-based approach is usually sensitive to the quality of the extraction contour during segmentation process, which typically complicates the differentiation between species

with very similar and close in terms of shapes. Thus, it is important for the process of the leaves compression in order to make the leaves not curling and flat. Usually, pressing leaves took a few days for herbarium species, but in this research it took only three to four hours, and the reasons for this was to make sure the leaves were still in fresh condition and to ensure the leaves became delicate and brittle.

5.3 Image Processing

Typically, the segmentation before extracting the features is needed, and the ultimate goal of segmentation is to separate the leaf from its background. The pictures of leaves are required to be against a light, and with an untextured background. Thus, in this research, a light box was used because it helped to diffuse light coming from multiple sources, which allows for even nearly shadow and helps to reduce the time-consuming image pre-processing process.

There are five steps in segmenting the ROI from the original image or RGB image. Firstly, the RGB image or the original image is converted to grey-scale image. Next, Canny edge detection method is applied to the grey-scale images and the image with detected edge is then converted to a binary image. The shape is obtained after the holes of the binary image are filled. The undesired shape of the small particles is removed and the ROI is obtained after this process. Figure 5.4 simplifies all of the steps of the image pre-processing.

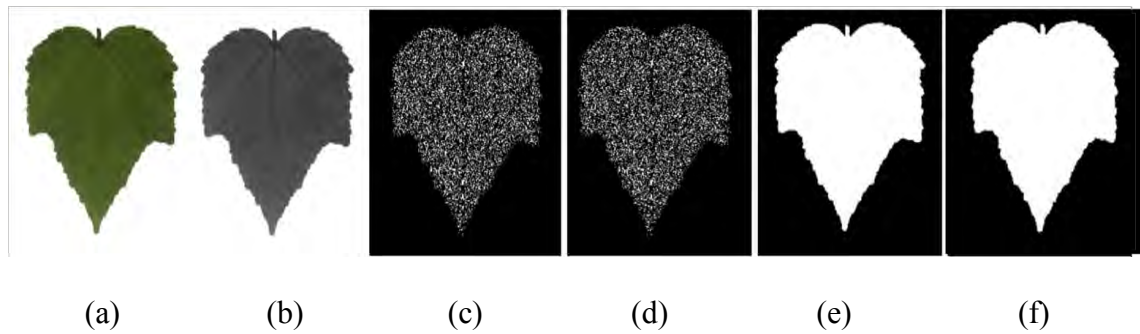


Figure 5.4: The overall steps in image pre-processing. (a) original image, (b) grayscale image, (c) detected edge, (d) binary image, (e) filled binary image, (f) ROI image.

5.4 Single Descriptor

Four types of shape representation were applied in this study which are MSD, HOG, Hu and ZM. MSD is the feature extraction methods that refer to five basic geometric properties of the leaf's shape. Based on these basic descriptors, the morphological descriptors were computed and utilised. In this research, there are 20 simple and morphological descriptors that were commonly applied from the previous studies were selected and implemented.

Table 5.1 shows the classification accuracy for single descriptor. The single descriptor that achieved the best accuracy was MSD with 96.39% using ANN classifier followed by DAG MLSTSVM with 95.78%. Whereas, the lowest accuracy of MSD descriptor obtained 82.80% by using LDA, which is still a good result for classification of tropical shrub species. It can be seen that, 20 descriptors of MSD were sufficient and adequate to give significant analysis for shape descriptor.

There are risks to describe the leaf shape using MSD descriptors only, even though they seem capable to classify a small set of test images. Additionally, many single value descriptors in MSD were highly correlated with each other, thus making the task of

selecting sufficient and relevant independent descriptors more challenging (Cope et al., 2012).

Table 5.1: Classification accuracy for single descriptor

	*Average accuracy (%)					
Descriptor	ANN	RF	SVM	k-NN	LDA	DAG MLSTSVM
MSD	96.39	92.58	79.78	91.96	82.80	95.78
HOG	95.82	91.58	84.53	90.40	79.76	95.40
Hu	82.27	83.36	32.74	82.99	37.65	85.76
ZM	91.79	87.85	59.34	87.75	56.40	90.54
*Average accuracy = average of 10 runs						

It can be seen that in Table 5.1, the HOG descriptors performed better than image moments, which are ZM and Hu. The result of single descriptor of HOG presented that most of the different leaf shapes in myDAUN dataset were correctly identified. The highest accuracy result for the HOG descriptor achieved 95.82%, which is considered as an extremely good result. However, several cases were not well recognised; this is because HOG computes histograms of all images into a block, therefore the local information might be lost.

Furthermore, HOG descriptor is sensitive to the leaf petiole orientation, and in the meantime, the petiole's shape actually carries the species characteristics. Thus, to conquer these drawbacks, a pre-processing step can normalise petiole orientation of all leaf images in a dataset making them applicable to HOG. This purpose had been approached by two studies, (Cope et al., 2012) and (Xiao et al., 2010), and it was shown that HOG achieved a better performance and gave more accurate results when the leaf petiole was cut off before analysis (Xiao et al., 2010). Unfortunately, the HOG descriptor is not invariant to the rotation and the scale changes as image moments operated.

On the other hand, Hu achieved the lowest classification accuracy with 32.74% in SVM and obtained the highest accuracy with 85.76% in DAG MLSTSVM and followed by 82.99% in k-NN. Even though Hu was computationally simple, it was highly sensitive to noise. Seven moments of Hu can represent shape characteristics well, but by using only seven descriptors were not enough for feature extraction because the information carried by their own was very restricted and limited when the image database is large. Generally, they need to be hybridised with other conventional descriptors in order to well describe the actual shape properties of the object.

ZM achieved relatively good results of classification accuracy with ANN and DAG MLSTSVM classifier and it can be a practicable way for classifying structural complex images. ZM gives exceptional invariance features over other moments based solution for instance Hu. However, the limitation and constraint of ZM was the costly computation that made it inapt for some problems. Kadir et al. (2011) found that ZM did not yield better classification accuracy than Hu. As shown in Table 5.1, ZM had the lowest classification accuracy with 56.40% in LDA and obtained the highest accuracy with 91.79% in ANN and followed by 90.54% in MLSTSVM.

5.5 Hybrid of Descriptor

Thus, in order to increase the accuracy of the classification of the tropical shrub species, the feature extraction methods of MSD, HOG, Hu and ZM were combined in to several of set hybridisation. Table 5.2 shows the accuracy of hybridisation of two, three and four descriptors. The hybridisation of the descriptors has successfully proved that the combination of more descriptors increases the accuracy of the classification.

Table 5.2: Classification accuracy of hybrid descriptors

Methods	Descriptor	*Average accuracy (%)					
		ANN	RF	SVM	k-NN	LDA	DAG MLSTSVM
Hybrid of two descriptors	{MSD + HOG}	97.49	93.45	91.01	92.03	89.56	96.94
	{MSD + Hu}	96.67	92.84	81.99	92.35	85.14	95.96
	{MSD + ZM}	96.60	92.86	84.61	91.92	84.31	96.25
	{HOG + Hu}	96.24	92.39	87.72	91.07	83.47	95.88
	{HOG + ZM}	93.70	92.58	89.93	91.47	85.58	93.52
	{Hu + ZM}	93.67	90.07	73.45	89.35	68.31	92.65
Hybrid of three descriptors	{MSD + HOG + Hu}	97.59	93.62	91.78	92.42	90.09	96.99
	{MSD + HOG + ZM}	97.63	93.52	92.06	92.17	89.72	97.05
	{MSD + Hu + ZM}	96.64	93.24	87.93	92.10	86.12	96.32
	{HOG + Hu + ZM}	97.06	93.37	91.29	91.56	87.23	96.70
Hybrid of all descriptors	{MSD + HOG + Hu + ZM}	98.23	93.83	92.74	92.60	90.86	97.72
* = average of 10 runs							

The results in Table 5.2 showed that the classification accuracy increased when combining more descriptors. When only MSD descriptor was used, false classification rate increased for similar and close shaped leaves of some species. The leaf samples of species 24 (*Mussaenda erythrophylla*) and species 25 (*Mussaenda philippica*) (see Figure 5.5) were often misclassified and unrecognised when MSD was used as input descriptor only.



(a) *Mussaenda philippica* leaf

(b) *Mussaenda erythrophylla* leaf

Figure 5.5: A leaf sample of *Mussaenda* sp.

This is due to the shape of the leaves in both species are similar to each other, since both of them belong to the same genus but different species. Although these leaves have similar shapes, but the leaf petiole for both species are obviously different. The leaf petiole of *M. philippica* looks more acute than the leaf petiole of *M. erythrophylla*.

It can be seen that the hybridisation of MSD and HOG descriptor increased the classification accuracy, and this assisted to decrease the misclassification of these tropical shrub species, as HOG descriptor was sensitive to the petiole orientation. The highest accuracy of the classification of hybrid of two descriptor obtained 97.49% with ANN classifier. The hybridisation of two and three descriptors achieved almost comparable results of the classification accuracy.

Subsequently, the hybridisation of all descriptors give the best results compared to single, two or three descriptors. In this research, the highest accuracy of tropical shrub species classification achieved 98.23% with ANN classifier. MSD and HOG descriptors were the leading contributors in the classification of tropical shrub species. On the other hand, Hu invariant moments and Zernike moments helped to improve and enhance the classification accuracy in tropical shrub species in terms of invariant to translation, rotation, and scale.

5.6 Feature Selection

Feature selection does not definitely mean an increase in accuracy. In fact, in all cases, reducing the number of descriptors too drastically will result in a decrease in the accuracy.

However, based on the results obtained after implementation of the feature selection of Relief, CFS, and PCC obtained comparable results compared to when using all 133 descriptors. The Relief method achieved the best result of 98.13% with 53 descriptors, which is 60% reduction in total descriptor and reduced computational time by 1033.23 minutes. This has convinced that feature selection methods are capable to select the optimal descriptors, which are correlated to each other in the classification of tropical shrub species.

5.7 Cross-validation (CV)

The proposed method was tested and implemented using the cross-validation method. The k-fold CV was applied to partition the data due to the small number of sample per species which is 30 samples per species in myDAUN dataset. The 5-fold CV and 10-fold CV achieved 97% of accuracy which was comparable to the proposed method with 98.23% accuracy. The results showed that there is no overfitting occurred in the classification modles

5.8 Validation

Then, the proposed methods were validated and the purpose of the validation is to test on the viability and applicability of using hybridisation of four descriptors on the classification of plant species in other datasets. The combination of all descriptors improved the classification accuracy and produced the best result for classification of plant species in the Flavia and Swedish Leaf datasets. Flavia dataset and Swedish Leaf dataset are currently the most often and popular benchmark datasets used by researchers to compare and evaluate method across studies.

Table 5.3 shows the average accuracy by using ANN classifier for various sets of descriptors in Flavia, Swedish Leaf and myDAUN dataset. The accuracy of classification using single descriptor method, MSD and HOG method achieved more than 93% in the Flavia dataset, 98% in the Swedish Leaf dataset and 95% in the myDAUN dataset.

The highest accuracy for the hybrid of two, three and four descriptors increased slightly for all dataset. The highest accuracy of all dataset were obtained when the hybridisation of all descriptors which were 95.25% in Flavia dataset, 99.89% in Swedish Leaf dataset and 98.23% in myDAUN dataset. The proposed method achieved nearly 100% accuracy on Swedish Leaf dataset. Whereas, when only MSD descriptor was used as input

descriptor in the Flavia dataset, the leaf samples of species *Ilex macrocarpa* and *Chimonanthus praecox* (see Figure 5.6) were often misclassified and unrecognised.

This is due to the shape of the leaves in both species are similar to each other. Hence, the hybridisation of MSD, HOG, Hu and ZM descriptor increased the classification accuracy and assisted to decrease the misclassification of plant identification, which is from 93% to 95%. Therefore, it can be concluded that the results achieved in the Flavia and Swedish Leaf datasets were comparable to those obtained in the myDAUN dataset. The validation results had proved the feasibility of the proposed methods in the automated classification of plant species.



(a) *Ilex macrocarpa* leaf



(b) *Chimonanthus praecox* leaf

Figure 5.6: A leaf sample of *Ilex macrocarpa* and *Chimonanthus praecox* in Flavia dataset

Table 5.3: Classification results of Flavia, Swedish Leaf and myDAUN dataset

Methods	Descriptor	**Average accuracy (%)		
		Flavia	Swedish Leaf	myDAUN
Single descriptor	MSD	93.30	98.65	96.39
	HOG	93.49	99.15	95.82
	Hu	80.46	95.20	82.27
	ZM	83.22	95.95	91.79
Hybrid of two descriptors	MSD + HOG	95.04	99.54	97.49
	MSD + Hu	93.12	98.37	96.67
	MSD + ZM	93.41	99.16	96.60
	HOG + Hu	93.55	99.24	96.24
	HOG + ZM	93.87	99.54	93.70
	Hu + ZM	88.47	98.01	93.67
Hybrid of three descriptors	MSD + HOG + Hu	94.01	99.43	97.59
	MSD + HOG + ZM	95.14	99.64	97.63
	MSD + Hu + ZM	93.67	99.16	96.64
	HOG + Hu + ZM	94.08	99.52	97.06
Hybrid of all descriptors	MSD + HOG + Hu + ZM	95.25	99.89	98.23
** = average of 10 runs				

5.9 Comparison Studies

Finally, the performance of our proposed method compared to other leaf classification studies is shown in Table 5.4. In the study performed by Pham et al. (2013), they compared the HOG and Hu, the results showed that HOG descriptor was more robust than Hu descriptor. The accuracy of HOG and Hu descriptor achieved in this study were 84.70% and 25.31% respectively. In the study presented by Salve et al. (2016), the implementation of HOG and ZM descriptor were proposed. This study used subset Visleaf dataset which contained 50 plant species and 10 samples for each of species, which is a total of 500 images. By using ZM as descriptor, the accuracy achieved 84.66% whereas HOG descriptor achieved 92.67%, and this indicated that ZM had lower accuracy compared to HOG.

Wu et al. (2007) used geometrical descriptors and morphological descriptors in the vein structure. The algorithm was quite simple and provided a good result of 90.31 % of accuracy but it required human intervention for the physiological length width. Moving on, Du et al. (2007) used a combination of morphological and Hu to recognize 20 species of plant and achieved 91% accuracy. Hossain & Amin (2010) used only MSD as part of descriptors set and obtained around 93% of accuracy.

Table 5.4: The performance of our proposed method compared to other leaf classification studies

Reference	Descriptor	Leaf dataset	Accuracy
Pham et al. (2013)	HOG	Flavia	84.70%
	Hu		25.31%
Salve et al. (2016)	ZM	Visleaf	84.66%
	HOG		92.67%
Wu et al. (2007)	MSD	Flavia	90.31%
Du et al. (2007)	MSD, Hu	Own dataset	91.00%
Hossain & Amin (2010)	MSD	Flavia	91.41%
Proposed study	MSD, HOG, Hu, ZM	myDAUN	98.23%
		Flavia	95.25%
		Swedish Leaf	99.89%

CHAPTER 6: CONCLUSION AND FUTURE WORK

6.1 Introduction

This chapter concludes the proposed methods for automated classification of tropical shrub species. The proposed method's strength and limitations are discussed. Finally, future enhancements are proposed.

6.2 Research Summary

The overall aim of this research is to apply a hybrid of leaf shape and machine learning approach for automated classification of tropical shrub species. To this end, a hybrid of all descriptors, which are combination of MSD, HOG, Hu invariant moments and Zernike moments with ANN classifiers, have been tested and the end results are very promising. This section summarises the findings in the development of an automated classification of tropical shrub species in line with the research objectives.

Chapter 3 explains the overviews path of this research and all the methodologies used. There are three objectives for this research. First, is to extract leaf shape features from the images of selected tropical shrub species and second, is to classify tropical shrub species based on leaf shape descriptors. These are described in Chapter 3 where the methods, preparation and procedures for acquiring myDAUN dataset were discussed. Third, is to compare different machine learning methods in the classification of tropical shrub species. The results of the proposed methods were described in Chapter 4 and the discussions were further discussed in Chapter 5.

The dataset of leaf image is named as myDAUN and the images in the myDAUN dataset were sampled from the campus of University of Malaya (UM), Kuala Lumpur, Malaysia.

myDAUN dataset currently focused only on the shrub species that are commonly available. Presently, the sampling took place in four main locations in UM because these locations have more variety of tropical shrubs. The four main locations were Faculty of Science, Tunku Canselor Hall, Varsity Lake and Main Library. With the help and advice from the professional botanists and the staff from the botanical garden, common tropical shrub species that are available at the stated location were selected and chosen.

In regards to the development of tropical shrub species image dataset, due to time and budget limitation, only 45 species of tropical shrubs were selected and 30 leaf samples were collected for each species. Hence, there were a total of 1350 images of tropical shrub leaf images. Before the leaf images were captured, all the samples were compressed and flattened using newspapers or books and the leaf stalks were removed. In order to capture a quality leaf photo, a light box was used and the captured images were taken in the same criteria and standard with uniform background.

Next, pre-processing methods were implemented on the topical shrub species dataset. There are several image processing steps for segmentation process in order to obtain ROI, which is from the RGB image and end with ROI image. Then, feature extractions by using shape representations were applied, which are morphological shape descriptor (MSD), histogram of oriented gradient (HOG), Hu and ZM. Subsequently, feature selection methods were implemented with the objectives to reduce the number of input variables to avoid over fitting, and to find an optimum feature subset for each descriptor. Three feature selection methods were implemented, which are Pearson's correlation coefficient (PCC), Relief, and Correlation-based feature selection (CFS). The number of features selected was reduced from 50%, 60% and 70% feature, and the selected features from each method were tested using proposed classification model.

Six types of classification methods were proposed in this research, which were artificial neural network (ANN), random forest (RF), support vector machine (SVM), k-nearest neighbour (k-NN), linear discriminant analysis (LDA) and directed acyclic graph multiclass least squares twin support vector machine (DAG MLSTSVM). All the classification algorithms were tested using random sampling with 80% for training and 20% for testing. The accuracy of classification result for myDAUN dataset was tested using various sets of descriptors, which are single and hybrid of two, three and four descriptors. The best result from the set of descriptor was compared with selected features selection methods. Furthermore, the proposed method was validated using Flavia and Swedish Leaf dataset, the most popular benchmark for leaf image dataset. The analyses and findings from the proposed automated classification of tropical shrub species using hybrid of leaf shape and machine learning approach are:

- (i) The performance of the hybrid of four descriptors of MSD, HOG, Hu and ZM stands out to be comparably better than single, two and three descriptors.
- (ii) The ANN classification model achieved the best accuracy in classification of tropical shrub species when compared to RF, SVM, k-NN, LDA and DAG MLSTSVM.
- (iii) The accuracy of Relief method with reduced feature of 60% (53 descriptors) showed better result than CFS and PCC, which achieved comparable result with the result without feature selection method.
- (iv) The optimum descriptors are MSD and HOG whereas Hu and ZM help to improve the performance of classification accuracy in terms of translation, rotation and scale.
- (v) The validation result of Flavia and Swedish Leaf dataset for classification using the proposed method obtained comparable results with myDAUN dataset.

As a conclusion, a classification of tropical shrub species using hybrid of four descriptors of MSD, HOG, Hu and ZM achieved the best accuracy compared to single, hybrid of two and three descriptors.

6.3 Research Constraints

The constraint in this study is the use of limited sample of leaf images. In this study, only 45 species of different tropical shrubs were used and each species includes 30 samples of leaf images. Thus, it does not generalise the classification method. The number of tropical shrub species that are currently available in myDAUN dataset is limited and large number of varieties of tropical shrub species has made it difficult to identify and characterise varieties solely on the basis of morphological characters, so that the data selections for the tropical shrub species are limited.

In order to improve the classification performance, the samples of the data for each species should be increased. In the early stage of data acquisition, it took a lot of time because it required advice from professional botanists and expert since plants, especially shrubs, have a variety of species and cultivar, thus the guidance from the professional botanists and expert were crucial and valuable.

All the images that are stored in myDAUN dataset must undergo a pre-processing stage before it can be used as an input for feature extraction part. The intend of pre-processing image is to normalise and standardise all the images in order to identify the main object, which is leaf shape, and to eliminate of all other unrelated information so that the image are in the same standard and are cleared of any noise. In this study, the pre-processing image for the RGB or original image is performed manually using Adobe Photoshop CC software. This step is used to enhance the image quality and to eliminate the illumination

and contrast problem, which would affect the process of object segmentation. Since the pre-processing image is performed manually, it required a lot of time to ensure that the images are in good quality and simplify the next step of image segmentation.

The ZM feature extraction has especially taken a lot of time to extract the features value compared to the other three descriptors. This is because it involves the factorial iterations for each radial polynomial and may cause to numerical instabilities as the order a increase.

6.4 Research Contributions

The contribution of this research can be divided into four parts. First, several of leaf shape features have been extracted, which accomplished Objective 1. It has been proven that the hybrid of feature extraction method using shape representations by combining all descriptors of {MSD + HOG + Hu + ZM} achieved best performance in tropical shrub species classification.

Second, the set of hybrid method had shown to perform better than single, hybrid of two and three descriptors. Thus, the proposed method is feasible to use as a feature extraction method for other leaf image dataset as well. MSD and HOG has been identified as the optimum descriptors for myDAUN dataset whereas Hu and ZM help to improve the performance of the classification accuracy in terms of translation, rotation and scale, which accomplished Objective 2.

Third, the Objective 3 has been attained by the findings of the ANN classification model that achieved high classification ability when compared to RF, SVM, k-NN, LDA and DAG MLSTSVM. Lastly, the Relief method showed better result than PCC and CFS for feature selection which accomplished the Objective 4.

6.5 Future Work

Several suggestions are recommended for future enhancement that could lead to the improvement of the classification of leaf images.

6.5.1 Larger Amount of Data Collection

The best way to increase the performance in terms of the speed of the computational time, and accuracy of the classification of tropical shrub species, is to increase the number of leaves (samples) used for training. This statement is just valid up to a point due to the leaves from a sample of tropical shrub species are dissimilar and different locations, thus, a large number of them can be collected to see all varieties of them.

Currently, the data collection of all tropical shrub species that available in myDAUN dataset only focus on the shrub species, thus in future the data collection in myDAUN dataset can be expanded with more varieties of plant species such as herbs and trees. While creating leaf image database of myDAUN, the research work considered only frontal, fresh and mature leaves. In future, leaves that are wrinkled occulted, immature and dry leaves can be considered and analysed with the proposed method in this study. In addition, discolouration or discoloured leaves is another challenging and difficult task of research that can be considered.

6.5.2 Utilising Other Descriptors

Combining different types of shape descriptors can significantly improve the classification performance, thus by adding more descriptors for feature extraction can be considered to obtain the better performance of classification accuracy. Several types of features can be mixed with weights in the decision rule of the classification. The approach method in this study can be extended to other approaches for tropical shrub species

classification based on plant features of colour, textures and structures of their flower, fruit and bark.

6.5.3 Utilising Other Classifiers

In order to improve the performance of the classifier, this study has utilised six types of classifier. The obtained results were encouraging and promising, as shown in Chapter 4. However, the search for a better classifier for automated classification of tropical shrub species in terms of performance in accuracy must continue because it is highly in demand. Utilising other well-known classifiers such as, but not limited to: Naïve Bayes, Nearest Feature Centre (NFC), and Classification and Regression Tree (CART) may improve the classification performance.

6.5.4 Mobile Apps

The next step in this research line is to develop a mobile app that includes the geo-reference of photos of leaves as an additional element to classify species. Nowadays, a mobile app usually carries everything required for the implementation of a mobile plant identification system, along with a camera, a processor, a user interface and an Internet connection. These make mobile app highly suitable for field use by professional botanists and general public.

REFERENCES

- Agarwal, G., Belhumeur, P., Feiner, S., Jacobs, D., Kress, W. J., Ramamoorthi, R., ... & Russell, R. (2006). First steps toward an electronic field guide for plants. *Taxon*, 55(3), 597-610.
- Aha, D. W., Kibler, D., & Albert, M. K. (1991). Instance-based learning algorithms. *Machine Learning*, 6(1), 37-66.
- Ahmed, N., Khan, U. G., & Asif, S. (2016). An automatic leaf based plant identification system. *Science International*, 28(1), 437-434.
- Apriyanti, D. H., Arymurthy, A. M., & Handoko, L. T. (2013). Identification of orchid species using content-based flower image retrieval. In *Proceeding of the Conference on Computer, Control, Informatics and Its Applications* (pp. 53-57). Jakarta, Indonesia: IEEE.
- Arlot, S., & Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistics Surveys*, 4, 40-79.
- Arora, A., Gupta, A., Bagmar, N., Mishra, S., & Bhattacharya, A. (2012). A plant identification system using shape and morphological features on segmented leaflets: Team IITK, CLEF 2012. In *Proceeding of the CLEF* (pp. 1-14). Florida, USA: India Institute of Technology.
- Beghin, T., Cope, J. S., Remagnino, P., & Barman, S. (2010). Shape and texture based plant leaf classification. In *Proceeding of the Conference on Advanced Concepts for Intelligent Vision Systems* (pp. 345-353). Berlin: Springer.
- Belongie, S., Malik, J., & Puzicha, J. (2002). Shape matching and object recognition using shape contexts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(4), 509-522.
- Bhardwaj, A., Kaur, M., & Kumar, A. (2013). Recognition of plants by leaf image using moment invariant and texture analysis. *International Journal of Innovation and Applied Studies*, 3(1), 237-248.
- Borcherding, K. (1977). Calibration of probabilities: The state of the art/comments. *Decision Making and Change in Human Affairs*, 325-329.
- Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. In *Proceedings of the Annual Workshop on Computational Learning Theory* (pp. 144-152). Pennsylvania, USA: ACM.
- Breiman, L. (2001). Random forest. *Machine Learning*, 45(1), 5-32.
- Bridle, H., Vrieling, A., Cardillo, M., Araya, Y., & Hinojosa, L. (2013). Preparing for an interdisciplinary future: A perspective from early-career researchers. *Futures*, 53, 22-32.
- Bruno, O. M., Plotze, R. D., Falvo, M., & Castro, M. D. (2008). Fractal dimension applied to plant identification. *Information Sciences*, 178(12), 2722-2733.

- Caballero, C., & Aranda, M. C. (2010). Plant species identification using leaf image retrieval. In *Proceedings of the Conference on Image and Video Retrieval* (pp. 327-334). Xi'an, China: ACM.
- Caglayan, A., Guclu, O., & Can, A. B. (2013). A plant recognition approach using shape and color features in leaf images. In Petrosino A. (Ed.), *Image Analysis and Processing: International Conference on Image Analysis and Processing* (pp. 161-170). Berlin, Springer.
- Cerutti, G., Tougne, L., Coquin, D., & Vacavant, A. (2013b). Curvature-scale-based contour understanding for leaf margin shape recognition and species identification. Paper presented at International Conference on Computer Vision Theory and Applications, Spain. Retrieved 7th June 2018, from: <https://hal.archives-ouvertes.fr/hal-00872870/>
- Cerutti, G., Tougne, L., Mille, J., Vacavant, A., & Coquin, D. (2013a). Understanding leaves in natural images – A model-based approach for tree species identification. *Computer Vision and Image Understanding*, 117(10), 1482-1501.
- Chaki, J., Parekh, R., & Bhattacharya, S. (2015a). Plant leaf recognition using texture and shape features with neural classifiers. *Pattern Recognition Letters*, 58, 61-68.
- Chaki, J., Parekh, R., & Bhattacharya, S. (2015b). Recognition of whole and deformed plant leaves using statistical shape features and neuro-fuzzy classifier. In *Proceeding of the Conference on Recent Trends in Information Systems* (pp. 189-194). Kolkata, India: IEEE.
- Chapman, A. D. (2009). *Numbers of living species in Australia and the world* (2nd ed.). Toowoomba, Australia: Australian Biodiversity Information Services.
- Chen, Y., Lin, P., & He, Y. (2011). Velocity representation method for description of contour shape and the classification of weed leaf images. *Biosystems Engineering*, 109(3), 186-195.
- Cho, S. (2012). Content-based structural recognition for flower image classification. In *Proceeding of the Conference on Industrial Electronics Applications* (pp. 541-546). Singapore: IEEE.
- Cope, J. S., Corney, D., Clark, J. Y., Remagnino, P., & Wilkin, P. (2012). Plant species identification using digital morphometrics: A review. *Expert Systems with Applications*, 39(8), 7562-7573.
- Corlett, R. T. (2016). Plant diversity in a changing world: Status, trends, and conservation needs. *Plant Diversity*, 38(1), 10-16.
- Corney, D. P., Tang, H. L., Clark, J. Y., Hu, Y., & Jin, J. (2012). Automating digital leaf measurement: the tooth, the whole tooth, and nothing but the tooth. *PLoS ONE*, 7(8).
- Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3), 273-297.

- Dalal, N., & Triggs, B. (2005). Histograms of oriented gradients for human detection. In *Proceeding of the Computer Society Conference on Computer Vision and Pattern Recognition* (pp. 886-893). California, USA: IEEE.
- Ding, S., Zhao, X., Zhang, J., Zhang, X., & Xue, Y. (2017). A review on multi-class TWSVM. *Artificial Intelligence Review*, 1-27.
- Du, J., Wang, X., & Zhang, G. (2007). Leaf shape based plant species recognition. *Applied Mathematics and Computation*, 185(2), 883-893.
- Du, J., Zhai, C., & Wang, Q. (2013). Recognition of plant leaf image based on fractal dimension features. *Neurocomputing*, 116, 150-156.
- Du, M., & Wang, X. (2011). Linear discriminant analysis and its application in plant classification. In *Proceeding of the Conference on Information and Computing* (pp. 548-551). Phuket, Thailand: IEEE.
- Efroni, I., Eshed, Y., & Lifschitz, E. (2010). Morphogenesis of simple and compound leaves: A critical review. *The Plant Cell*, 22(4), 1019-1032.
- Fern, B. M., Sulong, G. B., & Rahim, M. S. (2014). Leaf recognition based on leaf tip and leaf base using centroid contour gradient. *Advanced Science Letters*, 20(1), 209-212.
- Florindo, J. B., Backes, A. R., & Bruno, O. M. (2010). Leaves shape classification using curvature and fractal dimension. In *Proceeding of the Conference on Image and Signal Processing* (pp. 456-462). Berlin: Springer.
- Fotopoulou, F., Laskaris, N., Economou, G., & Fotopoulos, S. (2011). Advanced leaf image retrieval via Multidimensional Embedding Sequence Similarity (MESS) method. *Pattern Analysis and Applications*, 16(3), 381-392.
- Fu, H., & Chi, Z. (2006). Combined thresholding and neural network approach for vein pattern extraction from leaf images. *Journal of IEE Proceedings - Vision, Image, and Signal Processing*, 153(6), 881.
- Fu, H., Chi, Z., Feng, D., & Song, J. (2004). Machine learning techniques for ontology-based leaf classification. In *Proceeding of the Conference on Control, Automation, Robotics and Vision* (Vol. 1, pp. 681-686). Kunming, China: IEEE.
- Geertsema, M., Highland, L., & Vaugeouis, L. (2009). Environmental impact of landslides. In Sassa K., Canuti P. (Ed.), *Landslides – Disaster Risk Reduction* (pp. 589-607). Berlin: Springer.
- Gennari, J., Langley, P., & Fisher, D. (1988). Models of incremental concept formation. *Artificial Intelligence*, 40(1-3), 11-61.
- Ghiselli, E. E. (1973). *Theory of psychological measurement*. New York: McGraw-Hill.
- Gonzalez, R. C., & Woods, R. E. (2010). *Digital Image Processing* (3rd ed.). Upper Saddle River, New Jersey: Pearson Education, Inc.

- Gu, H. (2014). A directed acyclic graph algorithm for multi-class classification based on twin support vector machine. *Journal of Information and Computational Science*, 11(18), 6529-6536.
- Gupta, P. K. (2007). *Genetics: Classical to Modern* (1st ed.). Meerut, India: Rastogi Publications.
- Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine Learning*, 46(1/3), 389-422.
- Gwo, C., & Wei, C. (2013). Plant identification through images: using feature extraction of key points on leaf contours. *Applications in Plant Sciences*, 1(11), 1-9.
- Hall, M. A. (1999). *Correlation-based feature selection for machine learning* (Doctoral dissertation, The University of Waikato). Retrieved 7th June 2018, from: <https://www.lri.fr/~pierres/donn%E9es/save/these/articles/lpr-queue>
- Hore, A. V., Kehoe, J. G., McMullan, R., & Penton, M. R. (1997). *Construction 2 Environment Science Materials Technology*. London: Macmillan Education UK.
- Hossain, J., & Amin, M. A. (2010). Leaf shape identification based plant biometrics. In *Proceeding of the Conference on Computer and Information Technology* (pp. 458-463). Dhaka, Bangladesh: IEEE.
- Hsiao, J., Kang, L., Chang, C., & Lin, C. (2014). Comparative study of leaf image recognition with a novel learning-based approach. In *Proceeding of the Conference on Science and Information* (pp. 389-393). London, UK: IEEE.
- Hu, M. (1962). Visual pattern recognition by moment invariants. *IEEE Transactions on Information Theory*, 8(2), 179-187.
- Hu, R., Jia, W., Ling, H., & Huang, D. (2012). Multiscale distance matrix for fast plant leaf recognition. *IEEE Transactions on Image Processing*, 21(11), 4667-4672.
- Hussin, N. A., Jamil, N., Nordin, S., & Awang, K. (2013). Plant species identification by using Scale Invariant Feature Transform (SIFT) and Grid Based Colour Moment (GBCM). In *Proceeding of the Conference on Open Systems* (pp. 226-230). Kuching, Malaysia: IEEE.
- Hwang, S., & Kim, W. (2006). A novel approach to the fast computation of Zernike moments. *Pattern Recognition*, 39(11), 2065-2076.
- Jobin, A., Nair, M. S., & Tatavarti, R. (2012). Plant identification based on fractal refinement technique (FRT). *Procedia Technology*, 6, 171-179.
- Joly, A., Bonnet, P., Goëau, H., Barbe, J., Selmi, S., Champ, J., . . . Barthélémy, D. (2015). A look inside the Pl@ntNet experience. *Multimedia Systems*, 22(6), 751-766.

- Joly, A., Goëau, H., Bonnet, P., Bakić, V., Barbe, J., Selmi, S., . . . Barthélémy, D. (2014). Interactive plant identification based on social image data. *Ecological Informatics*, 23, 22-34.
- Kadir, A., Nugroho, L. E., Susanto, A., & Santosa, P. (2011b). Neural network application on foliage plant identification. *International Journal of Computer Applications*, 29(9), 15-22.
- Kadir, A., Nugroho, L. E., Susanto, A., & Santosa, P. I. (2012). Experiments of Zernike moments for leaf identification. *Journal of Theoretical and Applied Information Technology (JATIT)*, 41(1), 82-93.
- Kadir, A., Nugroho, L. E., Susanto, A., & Santosa, P. I. (2013a). Leaf classification using shape, color, and texture features. *International Journal of Engineering Trends and Technology*, 2(1), 225-230.
- Kadir, A., Nugroho, L., Susanto, A., & Santosa, P. (2011a). A comparative experiment of several shape methods in recognizing plants. *International Journal of Computer Science and Information Technology*, 3(3), 256-273.
- Kalyoncu, C., & Toygar, Ö. (2015). Geometric leaf classification. *Computer Vision and Image Understanding*, 133, 102-109.
- Kamavisdar, P., Saluja, S., & Agrawal, S. (2013). A survey on image classification approaches and techniques. *International Journal of Advanced Research in Computer and Communication Engineering*, 2(1), 1005-1009.1
- Kazerouni, M. F., Schlemper, J., & Kuhnert, K. (2015). Comparison of modern description methods for the recognition of 32 plant species. *Signal & Image Processing*, 6(2), 1-13.
- Kebapci, H., Yanikoglu, B., & Unal, G. (2010). Plant image retrieval using color, shape and texture features. *The Computer Journal*, 54(9), 1475-1490.
- Kellogg, E. A. (2016). *Flowering plants, Monocots: Poaceae*. Switzerland: Springer International Publishing.
- Kira, K., & Rendell, L. A. (1992). A practical approach to feature selection. *Elsevier*, 249-256.
- Kononenko, I. (1994). Estimating attributes: Analysis and extensions of RELIEF. In *Proceeding of the European Conference on Machine Learning* (pp. 171-182). Berlin: Springer.
- Kononenko, I. (1995). On biases in estimating multi-valued attributes. In *Proceeding of the Annual International joint Conference on artificial Intelligence (IJCAI)* (Vol. 95, pp. 1034-1040). Montreal, Canada: Morgan Kaufmann.
- Kononenko, I., Robnik-Sikonja, M., & Pompe, U. (1996). ReliefF for estimation and discretization of attributes in classification, regression, and ILP problems. *Artificial intelligence: methodology, systems, applications*, 31-40.

- Kulkarni, A. H., Rai, H. M., Jahagirdar, K. A., & Upparamani, P. S. (2013). A leaf recognition technique for plant classification using RBPNN and Zernike moments. *International Journal of Advanced Research in Computer and Communication Engineering*, 2(1), 984-988.
- Kumar, N., Belhumeur, P. N., Biswas, A., Jacobs, D. W., Kress, W. J., Lopez, I. C., & Soares, J. V. (2012). Leafsnap: A computer vision system for automatic plant species identification. In *Proceeding of the conference on Computer Vision 2012* (pp. 502-516). Berlin: Springer.
- Kuzyakov, Y., & Xu, X. (2013). Competition between roots and microorganisms for nitrogen: Mechanisms and ecological relevance. *New Phytologist*, 198(3), 656-669.
- Lamit, L. J., Lau, M. K., Næsborg, R. R., Wojtowicz, T., Whitham, T. G., & Gehring, C. A. (2015). Genotype variation in bark texture drives lichen community assembly across multiple environments. *Ecology*, 96(4), 960-971.
- Larese, M. G., Bayá, A. E., Craviotto, R. M., Arango, M. R., Gallo, C., & Granitto, P. M. (2014). Multiscale recognition of legume varieties based on leaf venation images. *Expert Systems with Applications*, 41(10), 4638-4647.
- Lavania, S., & Matey, P. S. (2014). Leaf recognition using contour based edge detection and SIFT algorithm. In *Proceeding of the Conference on Computational Intelligent and Computing Research* (pp. 1-4). Coimbatore, India: IEEE.
- Li, W. (2004). Using genetic algorithm for network intrusion detection. Paper presented at the *Proceedings of the United States Department of Energy Cyber Security Group*. Retrieved 7th June 2018, from: <https://bit.csc.lsu.edu/~jianhua/>
- Lin, H., & Peng, H. (2008). Machine recognition for broad-leaved trees based on synthetic features of leaves using probabilistic neural network. In *Proceeding of the Conference on Computer Science and Software Engineering* (Vol. 4, pp. 871-877). Hubei, China: IEEE.
- Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2), 91-110.
- Lowland Forests. (n.d.). Retrieved 7th June 2018, from: http://www.wwf.org.my/about_wwf/what_we_do/forests_main/the_malaysian_rainforest/types_of_forests/lowland_forests/
- Mata-Montero, E., & Carranza-Rojas, J. (2016). Automated plant species identification: challenges and opportunities. In *Proceeding of the Conference on IFIP World Information Technology Forum* (pp. 26-36). Berlin: Springer.
- Mitra, P., Shankar, B. U., & Pal, S. K. (2004). Segmentation of multispectral remote sensing images using active support vector machines. *Pattern Recognition Letters*, 25(9), 1067-1074.

- Mohammadi, M., Raahemi, B., Akbari, A., Nassersharif, B., & Moeinzadeh, H. (2012). Improving linear discriminant analysis with artificial immune system-based evolutionary algorithms. *Information Sciences*, 189, 219-232.
- Mouine, S., Yahiaoui, I., & Verroust-Blondet, A. (2013a). Combining leaf salient points and leaf contour descriptions for plant species recognition. In *Proceeding of the Conference on Image Analysis and Recognition* (pp. 205-214). Berlin: Springer.
- Mouine, S., Yahiaoui, I., & Verroust-Blondet, A. (2013b). A shape-based approach for leaf classification using multiscale triangular representation. In *Proceedings of the Conference on International Conference on Multimedia Retrieval* (pp. 127-134). Dallas, Texas: ACM.
- Nilsback, M., & Zisserman, A. (2006). A visual vocabulary for flower classification. In *Proceeding of the Computer Society Conference on Computer Vision and Pattern Recognition* (pp. 1447-1454). New York, USA: IEEE.
- Novotný, P., & Suk, T. (2013). Leaf recognition of woody species in Central Europe. *Biosystems Engineering*, 115(4), 444-452.
- Our History. (n.d.). Retrieved 7th June 2018, from: <https://www.um.edu.my/about-um/our-history>
- Pauwels, E. J., Zeeuw, P. M., & Ranguelova, E. B. (2009). Computer-assisted tree taxonomy by automated image recognition. *Engineering Applications of Artificial Intelligence*, 22(1), 26-31.
- Pearson, K. (1920). Notes on the history of correlation. *Biometrika*, 13(1), 25.
- Pham, N., Le, T., Grard, P., & Nguyen, V. (2013). Computer aided plant identification system. In *Proceeding of the Conference on Computing, Management and Telecommunications* (pp. 134-139). Ho Chi Minh, Vietnam: IEEE.
- Pimm, S. L., Jenkins, C. N., Abell, R., Brooks, T. M., Gittleman, J. L., Joppa, L. N., . . . Sexton, J. O. (2014). The biodiversity of species and their rates of extinction, distribution, and protection. *Science*, 344(6187), 987-999.
- Prasad, S., Kudiri, K. M., & Tripathi, R. C. (2011). Relative sub-image based features for leaf recognition using support vector machine. *Proceedings of the Conference on Communication, Computing & Security* (pp. 343-346). Odisha, India: ACM.
- Prasad, S., Peddoju, S. K., & Ghosh, D. (2013). Mobile plant species classification: A low computational approach. In *Proceeding of the Conference on Image Information Processing* (pp. 405-409). Shimla, India: IEEE.
- Priyankara, H. A., & Withanage, D. K. (2015). Computer assisted plant identification system for Android. In *Proceeding of the Moratuwa Engineering Research Conference* (pp. 148-153). Moratuwa, Sri Lanka: IEEE.
- Rademaker, C. A. (2000). The classification of plants in the United States Patent Classification system. *World Patent Information*, 22(4), 301-307.

- Rajasekaran, S., & A., V. P. (2012). *Neural networks, fuzzy logic and genetic algorithms: Synthesis and applications*. New Delhi: PHI Learning.
- Raven, P. H., Evert, R. F., & Eichhorn, S. E. (2013). *Biology of plants*. New York: W.H. Freeman and Company.
- RBG Kew (2016). *The state of the world's plants report–2016*. Richmond, UK: Royal Botanic Gardens, Kew.
- Refaeilzadeh, P., Tang, L., & Liu, H. (2009). Cross-validation. In Liu L., ÖZSU M.T. (Ed.). *Encyclopedia of database systems* (pp. 532-538): Berlin: Springer
- Ren, X., Wang, X., & Zhao, Y. (2012). An efficient multi-scale overlapped block lbp approach for leaf image recognition. In *International Conference on Intelligent Computing* (pp. 237-243). Berlin, Springer.
- Rimba Ilmu Botanical Garden. (n.d.). Retrieved 7th June 2018, from: <http://rimba.um.edu.my/>
- Robnik-Šikonja, M., & Kononenko, I. (1997). An adaptation of Relief for attribute estimation in regression. Paper presented at the Conference on Machine Learning. Retrieved 7th June 2018, from: <http://www.clopinet.com/isabelle/Projects/>
- Rosner, B. (2006). *Fundamentals of Biostatistics* (6th ed.). California: Thomson Higher Education.
- Saeys, Y., Inza, I., & Larranaga, P. (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics*, 23(19), 2507-2517.
- Salve, P., Sardesai, M., Manza, R., & Yannawar, P. (2015). Identification of the plants based on leaf shape descriptors. In Satapathy S., Raju K., Mandal J., Bhateja V. (Ed.), *Advances in Intelligent Ssytems and Computing: Proceeding of the International Conference on Computer and Communication Technologies* (Vol. 379, pp. 85-101). New Delhi: Springer.
- Silva, P. F. B. (2013). *Development of a System for Automatic Plant Species Recognition* (Master's thesis, University of Porto). Retrieved 7th June 2018, from: <https://repositorio-aberto.up.pt/bitstream/10216/67734/2/24444.pdf>
- Söderkvist, O. (2001). *Computer vision classification of leaves from swedish trees* (Master's thesis, Linköping University). Retrieved 7th June 2018, from: <http://www.diva-portal.org/smash/get/diva2:303038/FULLTEXT01.pdf>
- Song, S. L. (2005). In *Proceeding of the Conference of Control Applications* (pp. 831-836). Toronto, Canada: IEEE.
- Soni, N. K., & Soni, V. (2010). *Fundamentals of botany*. New Delhi: McGraw-Hill.
- Tandon, P., Abrol, Y. P., & Kumaria, S. (2007). *Biodiversity and its significance*. New Delhi: I.K. International Pub. House.

- Teng, C., Kuo, Y., & Chen, Y. (2009). Leaf segmentation, its 3D position estimation and leaf classification from a few images with very close viewpoints. In *Proceeding of the Conference Image Analysis and Recognition* (pp. 937-946). Berlin: Springer.
- The Malaysian Rainforest. (n.d.). Retrieved 7th June 2018, from: http://www.wwf.org.my/about_wwf/what_we_do/forests_main/the_malaysian_rainforest/
- Thepade, S., Das, R., & Ghosh, S. (2014). Feature extraction with ordered mean values for content based image classification. *Advances in Computer Engineering*, 1-15.
- Tilman, D., Cassman, K. G., Matson, P. A., Naylor, R., & Polasky, S. (2002). Agricultural sustainability and intensive production practices. *Nature*, 418(6898), 671-677.
- Tomar, D., & Agarwal, S. (2015). A comparison on multi-class classification methods based on least squares twin support vector machine. *Knowledge-Based Systems*, 81, 131-147.
- Vapnik, V. N., & Chervonenkis, A. Y. (1968). Algorithms with complete memory and recurrent algorithms in the problem of learning pattern recognition. *Avtomat. i Telemekh*, (4), 95-106.
- Viscosi, V., & Cardini, A. (2011). Leaf Morphology, taxonomy and geometric morphometrics: A simplified protocol for beginners. *PLoS ONE*, 6(10), 1-20.
- Wäldchen, J., & Mäder, P. (2017). Plant species identification using computer vision techniques: a systematic literature review. *Archives of Computational Methods in Engineering*, 25(2), 507-543.
- Wang, B., Brown, D., Gao, Y., & Salle, J. L. (2013). Mobile plant leaf identification using smart-phones. In *Proceeding of the Conference on Image Processing* (pp. 4417-4421). Melbourne, Australia: IEEE.
- Wang, G., & Wang, S. (2006). Recursive computation of Tchebichef moment and its inverse transform. *Pattern Recognition*, 39(1), 47-56.
- Wang, X., Du, J., & Zhang, G. (2005). Recognition of leaf images based on shape features using a hypersphere classifier. In *Proceeding of the Conference on Intelligent Computing* (pp. 87-96). Berlin: Springer.
- Wang, X., Huang, D., Du, J., Xu, H., & Heutte, L. (2008). Classification of plant leaf images with complicated background. *Applied Mathematics and Computation*, 205(2), 916-926.
- Wang, Z., Lu, B., Chi, Z., & Feng, D. (2011). Leaf image classification with shape context and SIFT descriptors. In *Proceeding of the Conference on Digital Image Computing Techniques and Applications* (pp. 650-654). Noosa, Australia: IEEE.

- Whittaker, R. H. (1969). New concepts of kingdoms of organisms. *Science*, 163(3863), 150-160.
- Wiens, J. J. (2016). Climate-related local extinctions are already widespread among plant and animal species. *PLOS Biology*, 14(12).
- Wilby, R. L., & Keenan, R. (2012). Adapting to flood risk under climate change. *Progress in Physical Geography*, 36(3), 348-378.
- Wilf, P., Zhang, S., Chikkerur, S., Little, S. A., Wing, S. L., & Serre, T. (2016). Computer vision cracks the leaf code. In *Proceedings of the National Academy of Sciences*, 113(12), 3305-3310.
- Willis K. J. (2017). *The state of the world's plants report-2017*. Richmond, UK: Royal Botanic Gardens, Kew.
- Wu, S. G., Bao, F. S., Xu, E. Y., Wang, Y., Chang, Y., & Xiang, Q. (2007). A leaf recognition algorithm for plant classification using probabilistic neural network. Paper presented at the International Symposium on Signal Processing using Probabilistic Neural Network, Egypt. Retrieved 7th June 2018, from: <https://ieeexplore.ieee.org/document/4458016>
- Xiao, X., Hu, R., Zhang, S., & Wang, X. (2010). HOG-Based approach for leaf classification. *Advanced Intelligent Computing Theories and Applications with Aspects of Artificial Intelligence*, 149-155.
- Xu, F., Guo, W., Xu, W., Wei, Y., & Wang, R. (2009). Leaf morphology correlates with water and light availability: What consequences for simple and compound leaves? *Progress in Natural Science*, 19(12), 1789-1798.
- Yang, L., & Wang, X. (2012). Leaf image recognition using fourier transform based on ordered sequence. In *Proceeding of the Conference on Intelligent Computing* (pp. 393-400). Berlin: Springer.
- Yanikoglu, B., Aptoula, E., & Tirkaz, C. (2014). Automatic plant identification from photographs. *Machine Vision and Applications*, 25(6), 1369-1383.
- Yu, H., & Yang, J. (2001). A direct LDA algorithm for high-dimensional data — with application to face recognition. *Pattern Recognition*, 34(10), 2067-2070.
- Zajonc, R. B. (1962). A note on group judgements and group size. *Human Relations*, 15(2), 177-180.
- Zhang, D., & Lu, G. (2004). Review of shape representation and description techniques. *Pattern Recognition*, 37(1), 1-19.
- Zhang, L., Kong, J., Zeng, X., & Ren, J. (2008). Plant species identification based on neural network. In *Proceeding of the Conference on Natural Computation* (Vol. 15, pp. 90-94). Washington, USA: IEEE.

- Zhang, S., & Feng, Y. (2010). Plant leaf classification using plant leaves based on rough set. In *Proceeding of the Conference on Computer Application and System Modeling* (Vol. 15, pp. 90-94). Taiyuan, China: IEEE.
- Zhao, C., Chan, S. S., Cham, W., & Chu, L. (2015). Plant identification using leaf shapes - A pattern counting approach. *Pattern Recognition*, 48(10), 3203-3215.
- Zulkifli, Z., Saad, P., & Mohtar, I. A. (2011). Plant leaf identification using moment invariants & General Regression Neural Network. In *Proceeding of the Conference on Hybrid Intelligent Systems (HIS)* on (pp. 430-435). Malacca, Malaysia: IEEE.

University of Malaya

LIST OF PUBLICATIONS AND PAPERS PRESENTED

The list of research papers published and presented in the conferences.

PUBLICATION

Murat, M., Chang, S. W., Abu, A., Yap, H. J., & Yong, K. T. (2017). Automated classification of tropical shrub species: a hybrid of leaf shape and machine learning approach. *PeerJ*, 5, e3792.

CONFERENCES

Miraemiliana Murat, Chang Siow Wee, Arpah Abu, Yap Hwa Jen, and Yong Kien Thai (2017). Automated classification of tropical shrub species: a hybrid of leaf shape and machine learning approach. *The 16th International Conference on Bioinformatics*, 20 – 22th September 2017, Tsinghua University, China.

Miraemiliana Murat, Chang Siow Wee, Arpah Abu, Yap Hwa Jen, and Yong Kien Thai (2017). Automated classification of tropical shrub species: a hybrid of leaf shape and machine learning approach, *22nd Biological Sciences Graduate Congress*, 19 – 21th December 2017, National University of Singapore, Singapore.



Automated classification of tropical shrub species: a hybrid of leaf shape and machine learning approach

Miraemiliana Murat¹, Siow-Wee Chang¹, Arpah Abu¹, Hwa Jen Yap¹ and Kien-Thai Yong¹

¹ Bioinformatics Programme, Institute of Biological Sciences, Faculty of Science, University of Malaya, Kuala Lumpur, Malaysia

² Department of Mechanical Engineering, Faculty of Engineering, University of Malaya, Kuala Lumpur, Malaysia

³ Ecology and Biodiversity Programme, Institute of Biological Sciences, Faculty of Science, University of Malaya, Kuala Lumpur, Malaysia

ABSTRACT

Plants play a crucial role in foodstuff, medicine, industry, and environmental protection. The skill of recognising plants is very important in some applications, including conservation of endangered species and rehabilitation of lands after mining activities. However, it is a difficult task to identify plant species because it requires specialized knowledge. Developing an automated classification system for plant species is necessary and valuable since it can help specialists as well as the public in identifying plant species easily. Shape descriptors were applied on the myDAUN dataset that contains 45 tropical shrub species collected from the University of Malaya (UM), Malaysia. Based on literature review, this is the first study in the development of tropical shrub species image dataset and classification using a hybrid of leaf shape and machine learning approach. Four types of shape descriptors were used in this study namely morphological shape descriptors (MSD), Histogram of Oriented Gradients (HOG), Hu invariant moments (Hu) and Zernike moments (ZM). Single descriptor, as well as the combination of hybrid descriptors were tested and compared. The tropical shrub species are classified using six different classifiers, which are artificial neural network (ANN), random forest (RF), support vector machine (SVM), k-nearest neighbour (k-NN), linear discriminant analysis (LDA) and directed acyclic graph multiclass least squares twin support vector machine (DAG MLSTSVM). In addition, three types of feature selection methods were tested in the myDAUN dataset, Relief, Correlation-based feature selection (CFS) and Pearson's coefficient correlation (PCC). The well-known Flavia dataset and Swedish Leaf dataset were used as the validation dataset on the proposed methods. The results showed that the hybrid of all descriptors of ANN outperformed the other classifiers with an average classification accuracy of 98.23% for the myDAUN dataset, 95.25% for the Flavia dataset and 99.89% for the Swedish Leaf dataset. In addition, the Relief feature selection method achieved the highest classification accuracy of 98.13% after 80 (or 60%) of the original features were reduced, from 133 to 53 descriptors in the myDAUN dataset with the reduction in computational time. Subsequently, the hybridisation of four descriptors gave the best results compared to others. It is proven that the combination MSD and HOG were good enough for tropical shrubs species classification. Hu and ZM descriptors also improved the accuracy in tropical shrubs

Submitted 23 May 2017
Accepted 20 August 2017
Published 12 September 2017

Corresponding author
Siow-Wee Chang,
siowwee@um.edu.my

Academic editor
Jun Pang

Additional Information and
Declarations can be found on
page 19

DOI 10.7717/peerj.3792

© Copyright
2017 Murat et al.

Distributed under
Creative Commons CC-BY 4.0

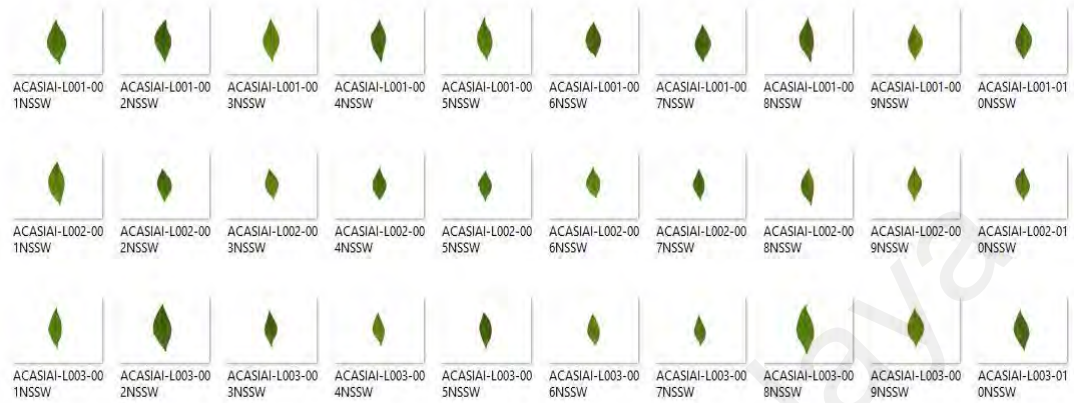
OPEN ACCESS

How to cite this article: Murat et al. (2017), Automated classification of tropical shrub species: a hybrid of leaf shape and machine learning approach. PeerJ 5:e3792; DOI 10.7717/peerj.3792

APPENDICES

Appendix A – Images in myDAUN dataset

Species 1 – *Acalypha siamensis*



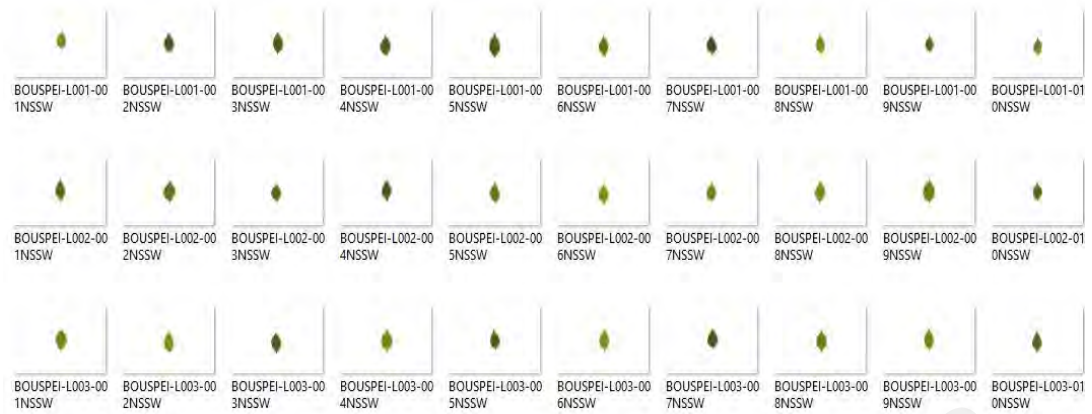
Species 2 – *Acalypha wilkesiana*



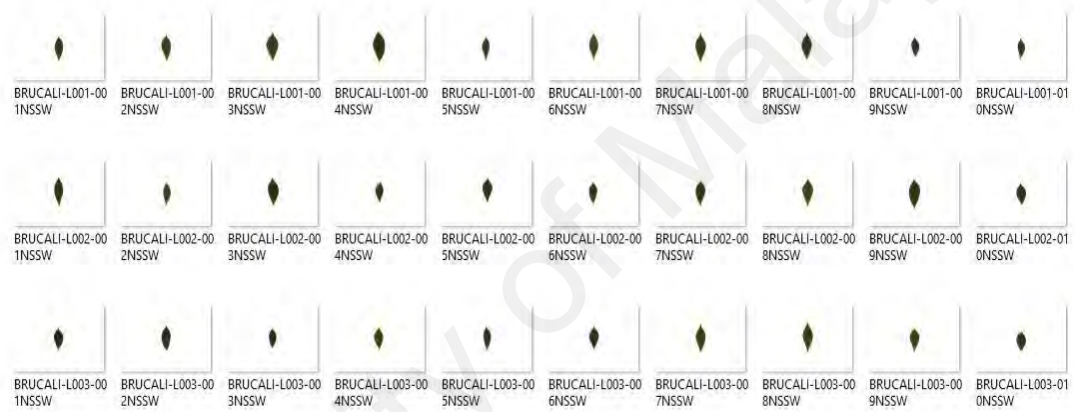
Species 3 – *Allamanda cathartica*



Species 4 – *Bougainvillea spectabilis*



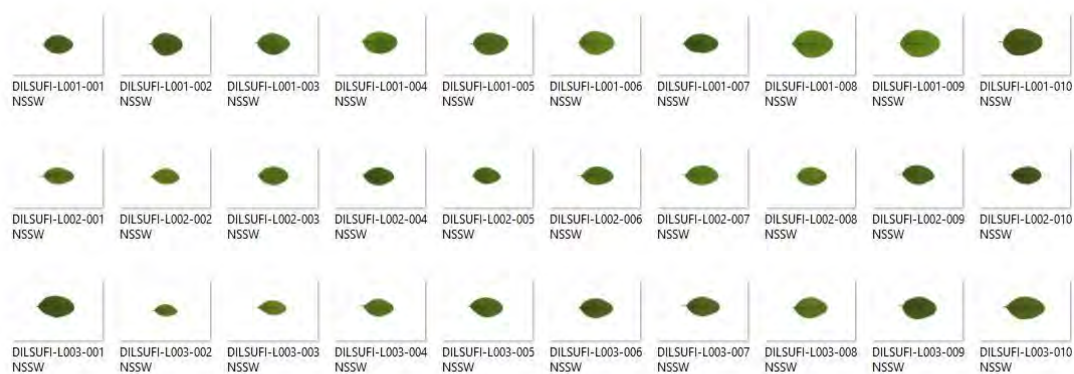
Species 5 – *Bruntelsia calycina*



Species 6 – *Clinacanthus nutans*



Species 7 – *Dillinea suffruticosa*



Species 8 – *Dracaena reflexa*



Species 9 – *Dracaena surculosa*



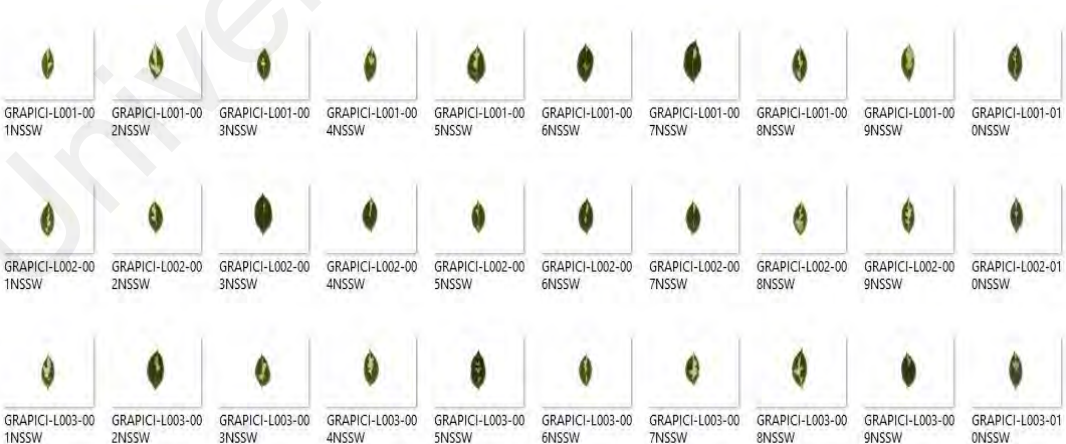
Species 10 – *Duranta erecta*



Species 11 – *Excoecaria cochinchinensis*



Species 12 – *Graptophyllum pictum*



Species 13 – *Hibiscus rosa-sinensis*



Species 14 – *Ixora javanica*



Species 15 – *Lagerstroemia indica*



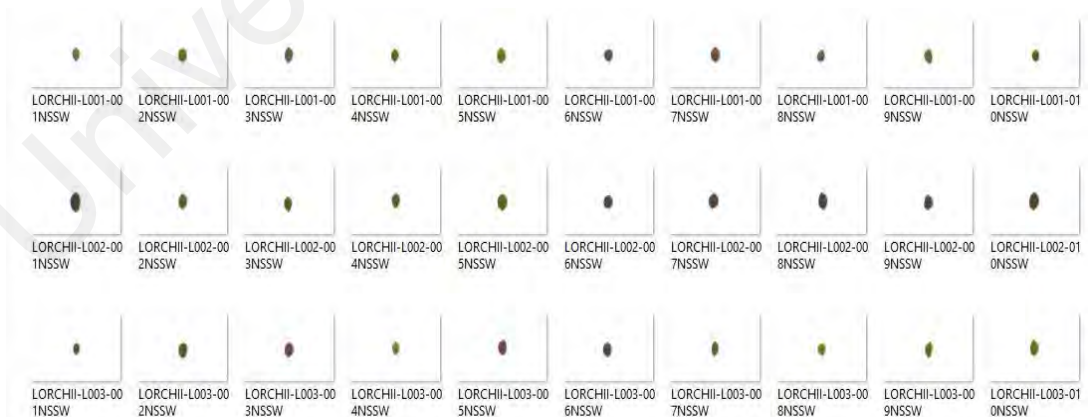
Species 16 – *Lantana camara*



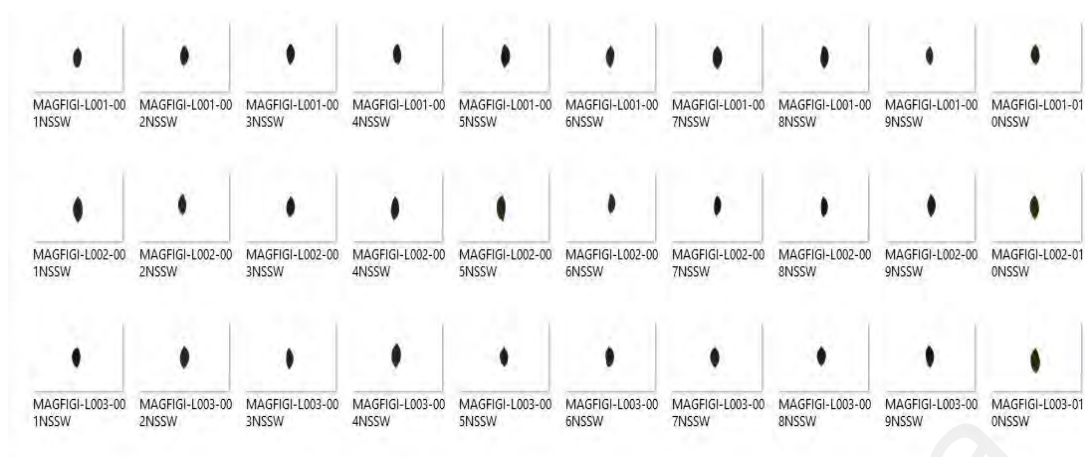
Species 17 – *Lawsonia inermis*



Species 18 – *Loropetalum chinense*



Species 19 – *Magnolia figo*



Species 20 – *Malvaviscus arboreus*



Species 21 – *Manihot esculenta*



Species 22 – *Melastroma malabathricum*



Species 23 – *Murraya paniculata*



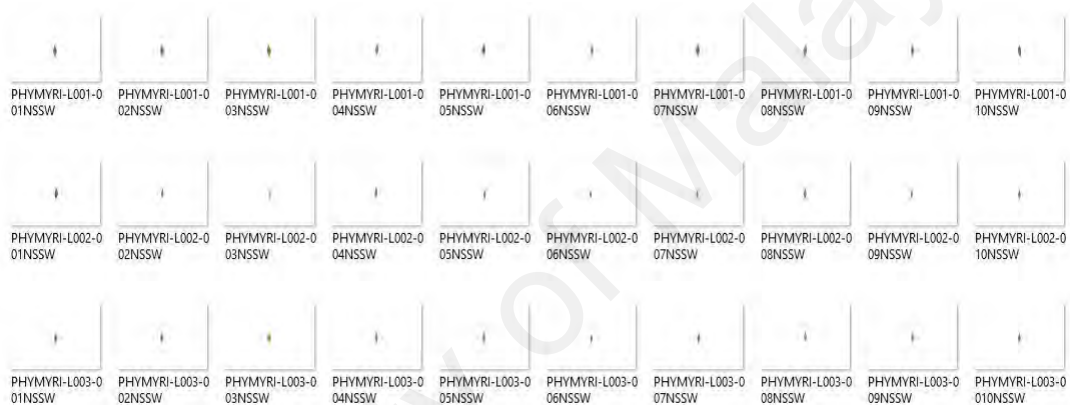
Species 24 – *Mussaenda erythrophylla*



Species 25 – *Mussaenda phillipica*



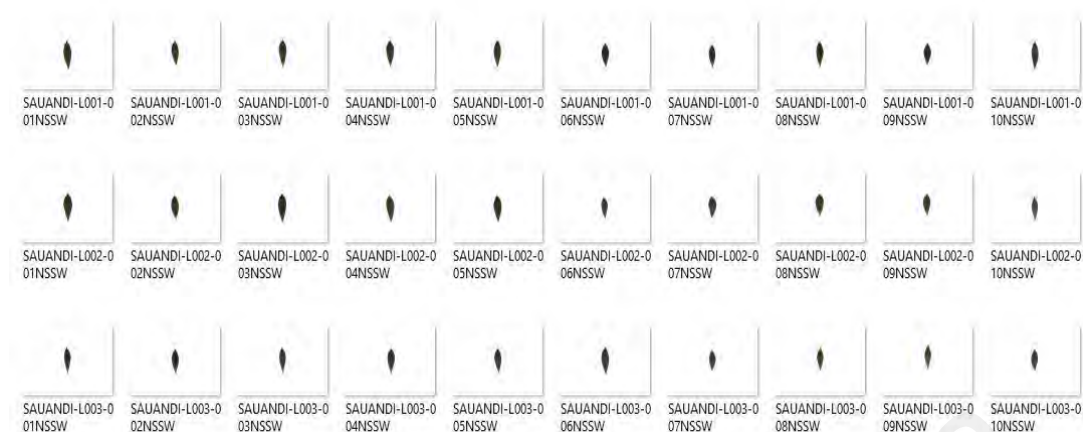
Species 26 – *Phyllanthus myrtifolius*



Species 27 – *Polyscias balfouriana*



Species 28 – *Sauropus androgynus*



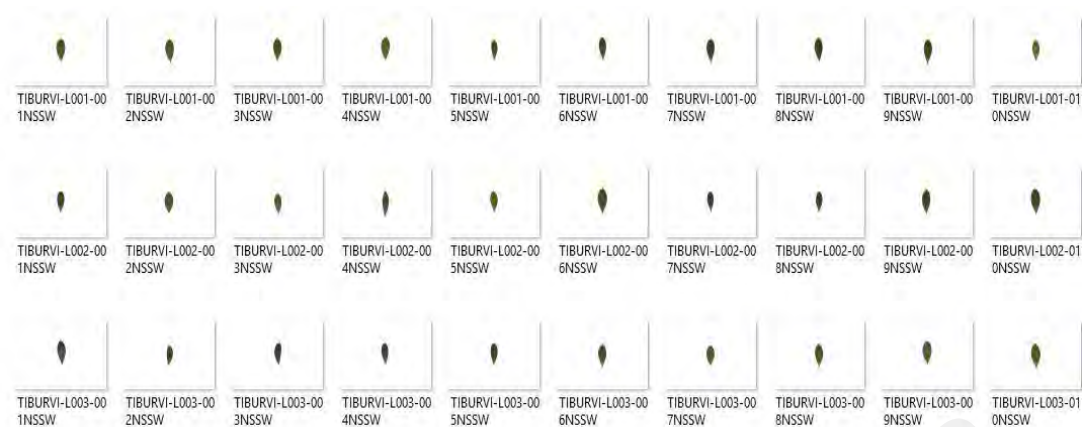
Species 29 – *Strobilanthes crispa*



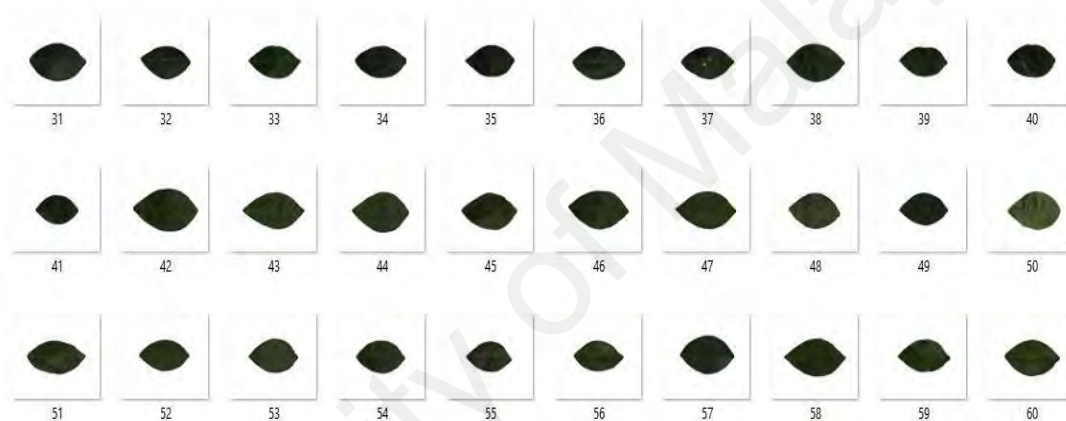
Species 30 – *Tabernaemontana divaricate*



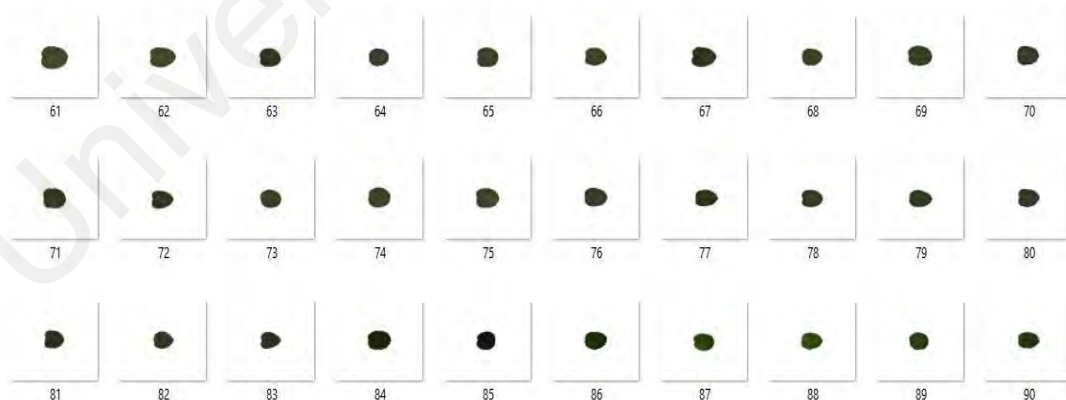
Species 31 – *Tibouchina urvilleana*



Species 32 – *Citrus microcarpa*



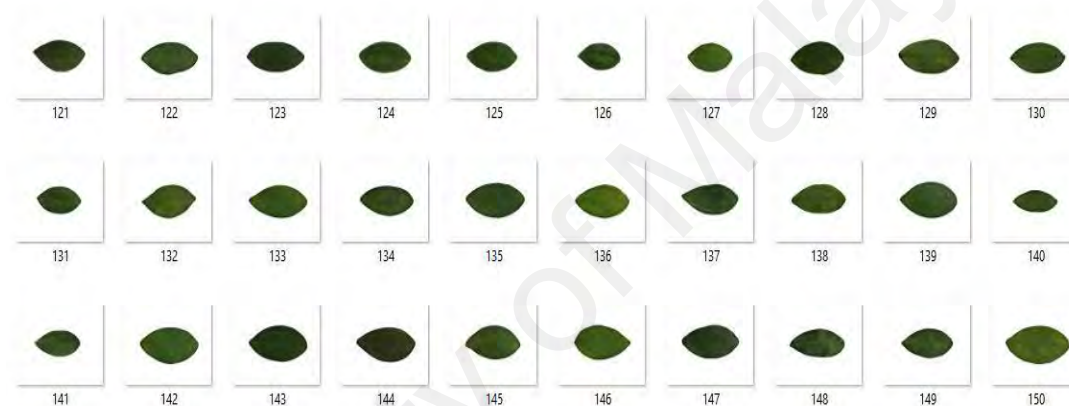
Species 33 – *Mentha piperita*



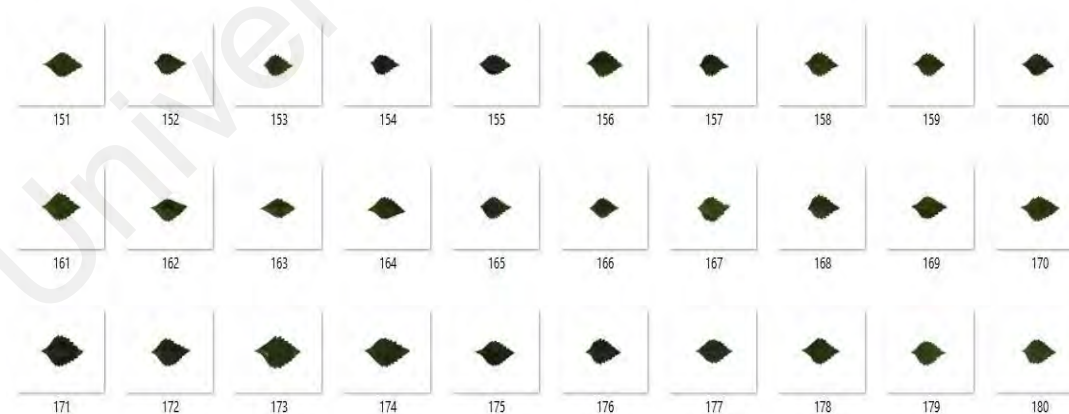
Species 34 – *Andrographis paniculata*



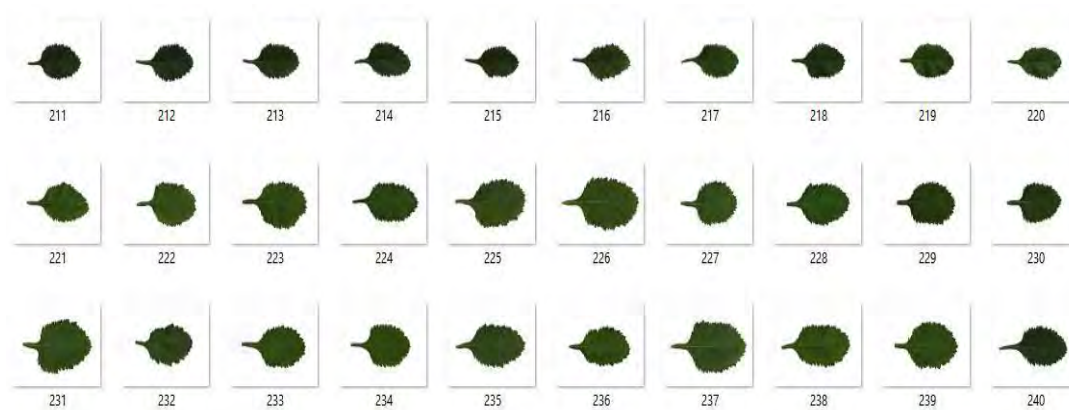
Species 35 – *Rhodomyrtus tomentosa*



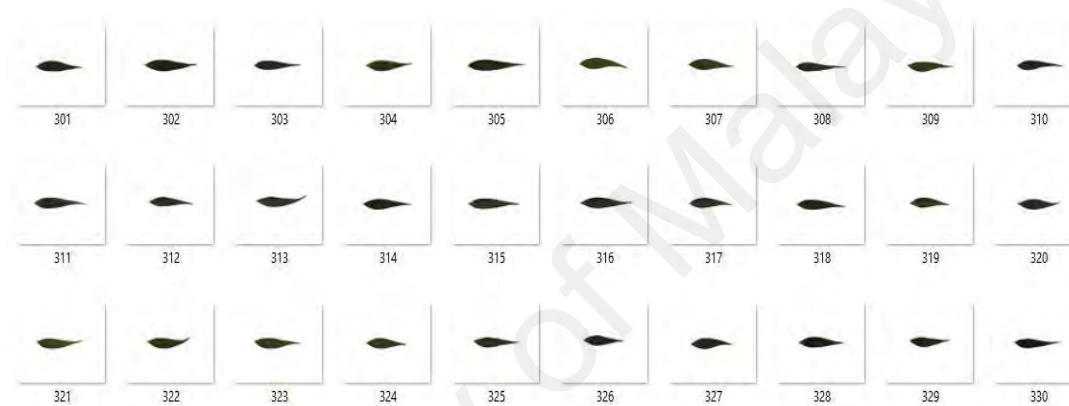
Species 36 – *Orthosiphon aristatus*



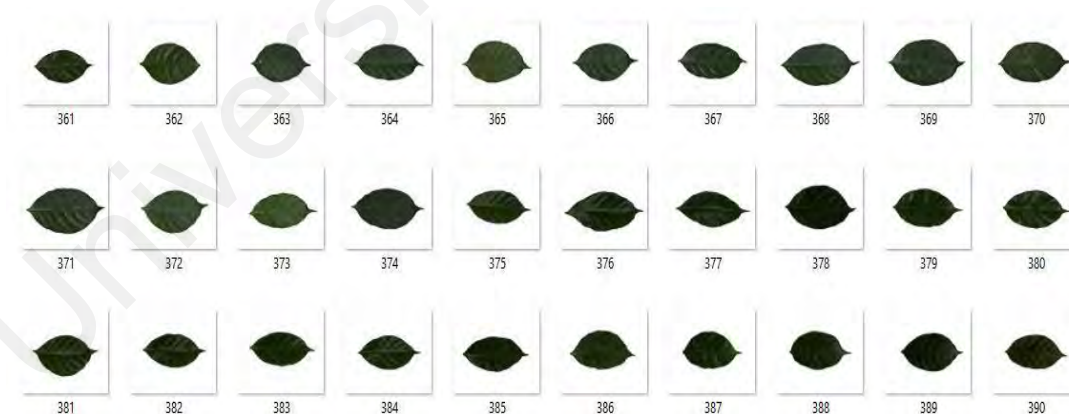
Species 37 – *Centratherum punctatum*



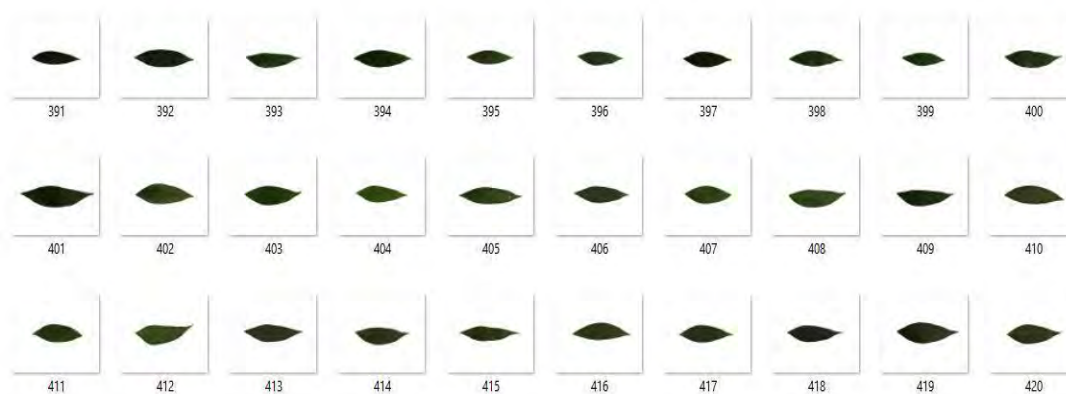
Species 38 – *Polygonum minus*



Species 39 – *Tabernaemontana coronaria*



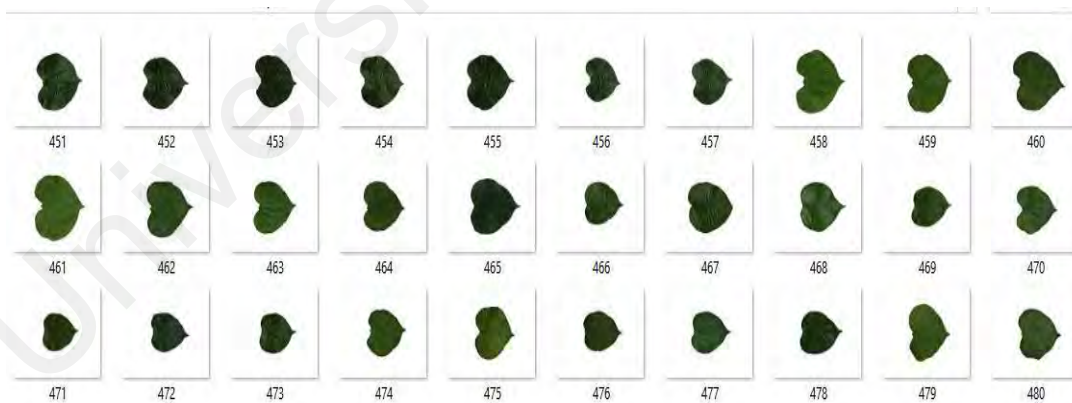
Species 40 – *Justicia gendarusa*



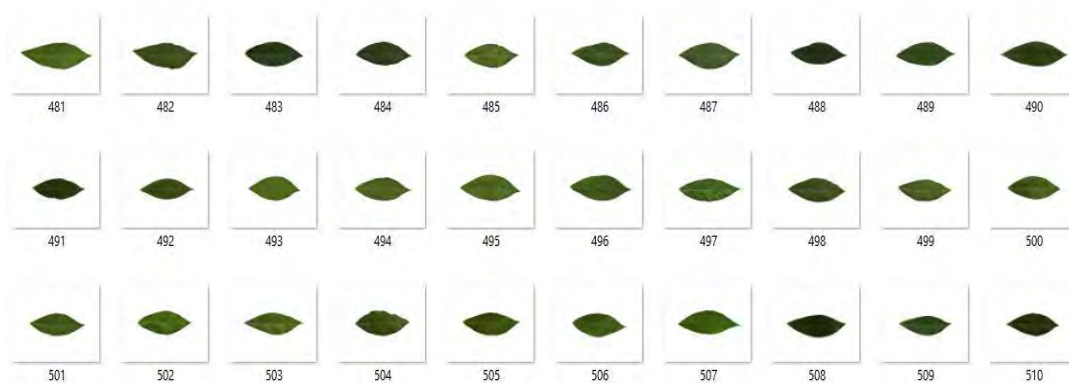
Species 41– *Tetracera scandens*



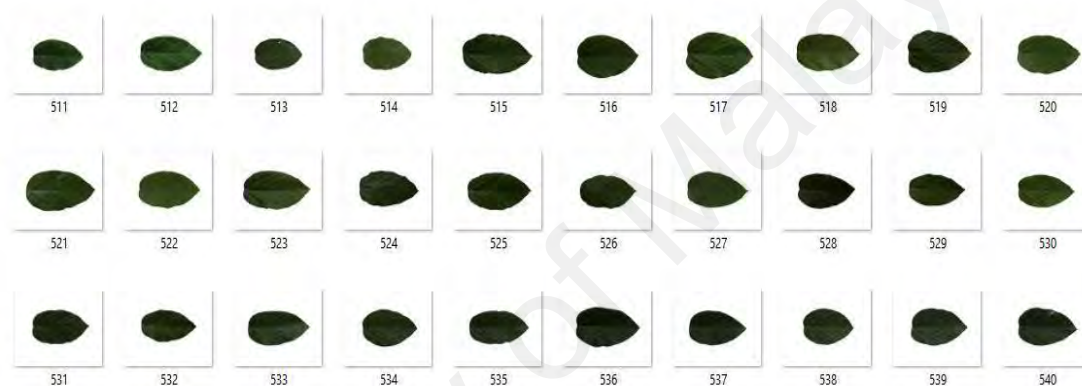
Species 42 – *Piper sarmentosum*



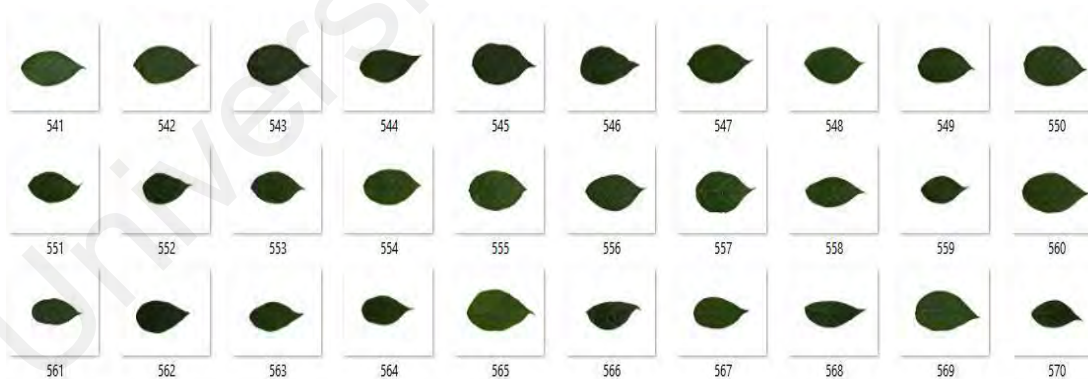
Species 43 – *Rauvolfia serpentina*



Species 44 – *Flemingia strobilifera*



Species 45 – *Cananga odorata*



Appendix B - Parameters of ANN

trainFcn	'trainscg'
trainParam	1 x 1 struct
performFcn	'crossentropy'
derivFcn	1 x 1 struct
divideFcn	'defaultderiv'
divideMode	'divideind'
divideParam	'sample'
trainInd	1 x 1 struct
valInd	1 x 1080 double
testInd	[]
stop	1 x 270 double
num_epochs	'minimum gradient'
trainMask	234
valMask	1 x 1 cell
testMask	1 x 1 cell
best_epoch	1 x 1 cell
goal	234
state	0
epoch	1 x 7 cell
time	1 x 235 double
perf	1 x 235 double
vperf	1 x 235 double
tperf	1 x 235 double
gradient	1 x 235 double
val_fail	1 x 235 double
best_perf	2.8641e-05
best_vperf	NaN
best_tperf	0.0305