# THE DEVELOPMENT OF AUTOMATED IDENTIFICATION SYSTEM FOR SELECTED SPECIES OF MONOGENEANS USING DIGITAL IMAGE PROCESSING, K-NEAREST NEIGHBOUR AND ARTIFICIAL NEURAL NETWORK APPROACHES

**ELHAM YOUSEF KALAFI**

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

**2017**

# THE DEVELOPMENT OF AUTOMATED IDENTIFICATION SYSTEM FOR SELECTED SPECIES OF MONOGENEANS USING DIGITAL IMAGE PROCESSING, K-NEAREST NEIGHBOUR AND ARTIFICIAL NEURAL NETWORK APPROACHES

**ELHAM YOUSEF KALAFI**

**THESIS SUBMITTED IN FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY**

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

**2017**

**UNIVERSITY OF MALAYA**

**ORIGINAL LITERARY WORK DECLARATION**

Name of Candidate: Elham Yousef Kalafi

Matric No: SHC130005

Name of Degree: Doctor of philosophy

Title of Thesis: The Development of Automated Identification System for Selected Species of Monogeneans Using Digital Image Processing, K-Nearest Neighbour and Artificial Neural Network Approaches

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:

## ABSTRACT

One of the key challenges to control diseases in fish population is achieving precise and correct identification of fish parasites. Monogenean parasites are flatworms (Platyhelminthes) that are primarily found on gills and skin of fishes. Organizing and preserving specimens of monogenean is a time consuming and difficult task. In addition, classification and identification of these specimens requires assistance of taxonomy experts. Since last two decades, improvements in developing computational tools made significant motivation to classify biological specimens` images to their correspondence species. These days, identification of biological species are easier for taxonomists and non-taxonomists due to the development of models and methods that are able to characterize species` morphology. Monogeneans have categorical homogeneous morphology, hence, pattern recognition techniques can be used to identify them. In this study, fully automated identification model for monogenean images based on the shape characters of their haptoral organs is developed. The morphological features were extracted from anchors and bars of monogeneans by adoption of digital image processing techniques. The Linear Discriminant Analysis (LDA) method was used to transform extracted feature vector to lower dimension feature vector and the transformed features were put into K-Nearest Neighbour (KNN) and Artificial Neural Network (ANN) classifiers for identification of monogenean specimens of eight species, *Sinodiplectanotrema malayanus*, *Diplectanum jaculator*,*Trianchoratus pahangensis*, *Trianchoratus lonianchoratus*, *Trianchoratus malayensis*, *Metahaliotrema ypsilocleithru*, *Metahaliotrema mizellei* and *Metahaliotrema similis*. Considerably, this is the first fully automated identification system for monogenean with the accuracy of 86.25% using KNN and 93.1% using ANN classification techniques. Images are classified based on monogenean diagnostic organs which are haptoral bars and anchors.

## ABSTRAK

Salah satu cabaran utama bagi mengawal penyakit dalam populasi ikan adalah mengenalpasti parasit ikan secara tepa. Parasit monogenean adalah cacing leper (Platyhelminthes) yang ditemui pada insang dan kulit ikan. Menyusun dan memelihara spesimen monogenean memakan masa yang lama dan merupakan suatu tugas yang sukar. Di samping itu, klasifikasi dan pengenalan spesimen ini memerlukan bantuan daripada pakar-pakar taksonomi. Sejak dua dekad yang lalu, peningkatan dalam penggunaan alatan komputer dijadikan motivasi penting dalam mengklasifikasi imej spesimen biologi berdasarkan spesies. Kini, pengecaman spesies biologi lebih mudah bagi ahli taksonomi dan bukan ahli taksonomi melalui pembangunan model dan kaedah yang dapat mencirikan morfologi species secara teratur. Monogenean mempunyai morfologi homogenan yang mutlak di mana teknik pengecaman corak boleh digunakan bagi mengenalpasti mereka. Dalam kajian ini, model pengecaman automatik sepenuhnya untuk imej monogenean dibangunkan berdasarkan ciri-ciri bentuk organ haptoral mereka. Ciri-ciri morfologi adalah berdasarkan sauh dan bar menggunakan teknik pemprosesan imej digital. Analisis diskriminan linear telah digunakan untuk memilih ciri-ciri terbaik dan dimasukkan ke dalam K-Nearest Neighbour (KNN) dan Artificial Neural Network (ANN) bagi pengecaman lapan spesies monogenean iaitu *Sinodiplectanotrema malayanus, Diplectanum jaculator, Trianchoratus pahangensis, Trianchoratus lonianchoratus, STrianchoratus malayensis, Metahaliotrema ypsilocleithru, Metahaliotrema mizellei dan Metahaliotrema similis.* Sehingga kini, ini merupakan sistem pengecaman automatik sepenuhnya yang pertama bagi monogenean dengan ketepatan 86.25% menggunakan KNN dan 93.1% menggunakan teknik pengelasan ANN. Imej dikelaskan berdasarkan organ diagnostik monogenean iaitu bar haptoral dan sauh.

## ACKNOWLEDGEMENTS

I would like to express my sincere gratitude and appreciation to my supervisor, Associate Prof. Dr. Sarinder Kaur, who has given me the opportunity to pursue my doctoral studies. I am deeply grateful for her guidance, patience, support and constant encouragement throughout the period of my research and preparation of this thesis. She have provided me and other lab members, conducive research environment, which encourages us to be independent in conducting original researches. Also, I wish to express my heartfelt appreciation to my late co-supervisor, Professor Susan Lim Lee Hong, for her invaluable support, guidance and advice over this study.

Besides, I would also like to thank my best friends, Hoda, Ali and Bahar who challenge and push me to improve and be a better researcher. I am deeply grateful for their invaluable support which was instrumental in shaping my approach to this research. Additionally, I would like to thank my fellow research colleagues, Qiqi, Lee Kien, Haris and Najib for their friendship, as well as making the lab a fun and productive environment. I am especially grateful to Dr. Tan Wai Boon for his assistance in lab works. He always came through when I needed help in my research.

Last but not least, I wish to express my deepest love and gratitude to my family. Words are too pale to express how grateful I am to my parents and my brothers for their unconditional love, care, support, encouragement and understanding. Thank you mom and dad for showing faith in me and giving me liberty to choose what I desired. My love and thanks to my beloved husband and best friend, Masoud, my constant companion, who always stood by my side supporting and motivating me all the way through till the end of my PhD journey. Thank you for your sacrifice and embracing my dreams as you do your own.

## TABLE OF CONTENT

Abstract……...……iii

Abstrak……iv

Acknowledgements ..... v

List of Figures ..... ix

List of Tables..... xii

List of Symbols and Abbreviations ..... xiii

List of Appendices ..... xiv

**CHAPTER 1: INTRODUCTION** ..... 1

1.1 Overview ..... 2

1.2 Research questions ..... 3

1.3 Objectives of the study ..... 4

1.4 Scope of the study ..... 4

1.5 Outline of the study ..... 4

**CHAPTER 2: LITERATURE REVIEW** ..... 6

2.1 Monogeneans..... 10

2.2 Image Acquisition ..... 13

2.3 Database ..... 15

2.4 Image processing ..... 16

2.5 Feature Extraction and Selection ..... 19

2.5.1 Feature Extraction ..... 19

2.5.2 Feature Selection..... 24

2.6 Classification ..... 26

2.6.1 K-Nearest Neighbour (KNN)..... 28

2.6.2 Artificial Neural Network (ANN)..... 30

**CHAPTER 3: METHODOLOGY** ..... 32

3.1 Monogeneans Collection ..... 33

3.2 Monogeneans Image Acquisition ..... 33
3.3 Database of Digital Images ..... 34
3.4. Preliminary Identification : Four Species (First Stage) ..... 38
3.4.1 Image Processing ..... 39
3.4.2 Feature Extraction ..... 42
3.4.3 Feature Selection ..... 42
3.4.4 Classification ..... 43
3.4.5 Evaluation ..... 45
3.5. Extended Identification on Eight Species ( Second Stage) ..... 45
3.5.1 Image Processing ..... 45
3.5.2 Extraction of One Anchor ..... 49
3.5.3 Feature Extraction ..... 50
3.5.4 Feature Selection ..... 51
3.5.5 Classification ..... 51
3.5.6 Evaluation ..... 52

**CHAPTER 4: RESULTS** ..... 54
4.1 Preliminary Identification Results (First Stage) ..... 55
4.1.1 Feature Selection ..... 55
4.1.2 Classification ..... 59
4.1.3 Evaluation ..... 65
4.2 Species Identification Results on Eight Species of Monogeneans (Second Stage) ..... 66
4.2.1 Feature Selection ..... 66
4.2.2 Classification ..... 75
4.2.3 Evaluation ..... 80
4.3 Overall Results ..... 81

**CHAPTER 5: DISCUSSION AND CONCLUSION** ..... 82
5.1 Image Acquisition and Database ..... 83

5.2 Monogenean Identification....84
5.3 Comparison with Previous Studies....85
5.4 Constraints and Limitations....86
5.5 Future Works....87
5.6 Conclusions....89
**REFERENCES**....91
**LIST OF PUBLICATIONS AND PAPERS PRESENTED**....102
**APPENDIX A**....103
**APPENDIX B**....104

## LIST OF FIGURES


| | | |
|---|---|---|
| Figure 2.1: | Illustration of a monogenean worm consisting of three main parts | 12 |
| Figure 2.2: | Illustration of image acquisition problems of Euryhaliotrema during digitization | 14 |
| Figure 2.3: | Content based features | 20 |
| Figure 2.4: | The comparison of mean accuracy of SRV and CBIR approaches in categorization of species in five iterations | 23 |
| Figure 2.5: | K Nearest Neighbour classifier in two dimensional feature space | 29 |
| Figure 2.6: | A representation of a simple perceptron | 30 |
| Figure 3.1: | The Scheme of process of proposed identification system for monogeneans | 32 |
| Figure 3.2: | Slides of monogenean specimens | 33 |
| Figure 3.3: | Digitizing the monogenean specimens, using Leica digital camera DFC 320 attached to Leica DMRB microscope | 34 |
| Figure 3.4: | *Sinodiplectanotrema malayanus* | 35 |
| Figure 3.5: | *Trianchoratus pahangensis* | 35 |
| Figure 3.6: | *Metahaliotrema mizellei* | 36 |
| Figure 3.7: | *Metahaliotrema similis* | 36 |
| Figure 3.8: | *Trianchoratus lonianchoratus* | 36 |
| Figure 3.9: | *Trianchoratus malayensis* | 37 |
| Figure 3.10: | *Metahaliotrema ypsilocleithru* | 37 |
| Figure 3.11: | *Diplectanum jaculator* | 37 |
| Figure 3.12: | Image database for training and testing dataset | 38 |
| Figure 3.13: | List of installed toolboxes in MATLAB | 40 |
| Figure 3.14: | Process in image pre-processing, edge detection and image segmentation steps for four species of *Sinodiplectanotrema malayanus, Trianchoratus pahangensis, Metahaliotrema mizellei* and *Metahaliotrema similis* | 41 |
| Figure 3.15: | Neural Network with 10 sigmoid hidden nodes and four output neurons | 44 |
| Figure 3.16: | The illustration of anchors and bars of *Metahaliotrema ypsilocleithrum* | 46 |
| Figure 3.17: | The process of detecting edges from intensity image | 48 |
| Figure 3.18: | The process of converting binary image to segmented image | 49 |
| Figure 3.19: | Extraction of one anchor of each species | 49 |
| Figure 3.20: | Neural Network with 10 sigmoid hidden nodes and four output neurons | 52 |
| Figure 4.1: | 3D scatter plot with different features | 56 |
| Figure 4.2: | 2D scatter plot of first element of selected feature vector by LDA for samples of *Sinodiplectanotrema malayanus, Trianchoratus pahangensis, Metahaliotrema mizellei* and *Metahaliotrema similis* | 57 |

| | | |
|---|---|---|
| Figure 4.3: | 2D scatter plot of second element of selected feature vector by LDA for samples of *Sinodiplectanotrema malayanus, Trianchoratus pahangensis, Metahaliotrema mizellei* and *Metahaliotrema similis* | 58 |
| Figure 4.4: | 2D scatter plot of third element of selected feature vector by LDA for samples of *Sinodiplectanotrema malayanus, Trianchoratus pahangensis, Metahaliotrema mizellei* and *Metahaliotrema similis* | 58 |
| Figure 4.5: | The distinction of feature values before and after LDA feature selection | 59 |
| Figure 4.6: | Illustration of k value in 25 iteration of KNN classification for four species | 60 |
| Figure 4.7: | Neural network training validation performance according to mean square error for four species | 62 |
| Figure 4.8: | Confusion matrix of testing dataset | 63 |
| Figure 4.9: | Illustration of distribution of the neural network errors | 64 |
| Figure 4.10: | The neural network training state showing the progress of the gradient magnitude, the number of validation checks | 64 |
| Figure 4.11: | The Receiver Operating Characteristic (ROC) of training network | 65 |
| Figure 4.12: | 2D scatter plot of first element of selected feature vector by LDA for samples of *Sinodiplectanotrema malayanus, Diplectanum jaculator, Trianchoratus pahangensis, Trianchoratus lonianchoratus, Trianchoratus malayensis, Metahaliotrema ypsilocleithru, Metahaliotrema mizellei* and *Metahaliotrema similis* | 68 |
| Figure 4.13: | 2D scatter plot of second element of selected feature vector by LDA for samples of *Sinodiplectanotrema malayanus, Diplectanum jaculator, Trianchoratus pahangensis, Trianchoratus lonianchoratus, Trianchoratus malayensis, Metahaliotrema ypsilocleithru, Metahaliotrema mizellei* and *Metahaliotrema similis* | 68 |
| Figure 4.14: | 2D scatter plot of third element of selected feature vector by LDA for samples of *Sinodiplectanotrema malayanus, Diplectanum jaculator, Trianchoratus pahangensis, Trianchoratus lonianchoratus, Trianchoratus malayensis, Metahaliotrema ypsilocleithru, Metahaliotrema mizellei* and *Metahaliotrema similis* | 69 |
| Figure 4.15: | 2D scatter plot of fourth element of selected feature vector by LDA for samples of *Sinodiplectanotrema malayanus, Diplectanum jaculator, Trianchoratus pahangensis, Trianchoratus lonianchoratus, Trianchoratus malayensis, Metahaliotrema ypsilocleithru, Metahaliotrema mizellei* and *Metahaliotrema similis* | 70 |

| | | |
|---|---|---|
| Figure 4.16: | 2D scatter plot of fifth element of selected feature vector by LDA for samples of *Sinodiplectanotrema malayanus, Diplectanum jaculator, Trianchoratus pahangensis, Trianchoratus lonianchoratus, Trianchoratus malayensis, Metahaliotrema ypsilocleithru, Metahaliotrema mizellei* and *Metahaliotrema similis* | 71 |
| Figure 4.17: | 2D scatter plot of sixth element of selected feature vector by LDA for samples of *Sinodiplectanotrema malayanus, Diplectanum jaculator, Trianchoratus pahangensis, Trianchoratus lonianchoratus, Trianchoratus malayensis, Metahaliotrema ypsilocleithru, Metahaliotrema mizellei* and *Metahaliotrema similis* | 72 |
| Figure 4.18: | 2D scatter plot of seventh element of selected feature vector by LDA for samples of *Sinodiplectanotrema malayanus, Diplectanum jaculator, Trianchoratus pahangensis, Trianchoratus lonianchoratus, Trianchoratus malayensis, Metahaliotrema ypsilocleithru, Metahaliotrema mizellei* and *Metahaliotrema similis* | 73 |
| Figure 4.19: | 3D scatter plot with combination of LDA selected features: FvLDA1, FvLDA2 and FvLDA3 | 74 |
| Figure 4.20: | Feature vector comparison after and before feature selection | 74 |
| Figure 4.21: | Illustration of k value in 15 iteration of KNN classification for eight species | 75 |
| Figure 4.22: | Neural network training validation performance according to mean square error for eight species | 77 |
| Figure 4.23: | Confusion matrix of testing dataset. The confusion matric shows the classification of eight species of Monogeneans by ANN classifier | 78 |
| Figure 4.24: | Illustration of distribution of the neural network errors | 78 |
| Figure 4.25: | The neural network training state showing the progress of the gradient magnitude, the number of validation checks | 79 |
| Figure 4.26: | The Receiver Operating Characteristic of training network | 80 |

## LIST OF TABLES


| | | |
|---|---|---|
| Table 2.1: | Examples of some automated species identification systems | 9 |
| Table 2.2: | Example of some databases applied in automated identification systems of species images | 16 |
| Table 2.3: | Image processing algorithms used in automated species identification systems | 17 |
| Table 2.4: | Thresholding techniques used in automated species identification systems | 19 |
| Table 2.5: | Overview of shape representation techniques | 21 |
| Table 2.6: | Some of feature selection algorithms used in automated species identification systems | 25 |
| Table 3.1: | Descriptions of shape parameters, used for feature extraction in four species | 42 |
| Table 3.2: | Descriptions of shapes parameters, used for feature extraction in eight species | 50 |
| Table 4.1: | Confusion matrix of KNN classification for four species of: *Sinodiplectanotrema malayanus (Smm), Trianchoratus pahangensis (Tp), Metahaliotrema mizellei (Mmi)* and *Metahaliotrema similis (Mma)* | 60 |
| Table 4.2: | Neural network training performance in terms of mean square error for training, testing and validation sets | 61 |
| Table 4.3: | Confusion matrix of leave one out cross validation for four species of *Sinodiplectanotrema malayanus (Smm), Trianchoratus pahangensis (Tp), Metahaliotrema mizellei (Mmi)* and *Metahaliotrema similis (Mma)* | 66 |
| Table 4.4: | Confusion matrix of KNN classification for eight species | 76 |
| Table 4.5: | Confusion matrix of monogenean Intra-genus KNN classification. A) *Metahaliotrema* samples B) *Trianchoratus* samples | 76 |
| Table 4.6: | Neural network training performance in terms of mean square error for training, testing and validation sets | 77 |
| Table 4.7: | Confusion matrix of leave one out cross validation for eight species of *Sinodiplectanotrema malayanus, Diplectanum jaculator, Trianchoratus pahangensis, Trianchoratus lonianchoratus, Trianchoratus malayensis, Metahaliotrema ypsilocleithru, Metahaliotrema mizellei* and *Metahaliotrema similis* | 81 |
| Table 4.8: | Accuracy of classification techniques in preliminary and extended models | 81 |

## LIST OF SYMBOLS AND ABBREVIATIONS

| | | |
|---|---|---|
| LDA | : | Linear Discriminant Analysis |
| ANN | : | Artificial Neural Network |
| KNN | : | K-Nearest Neighbour |
| DiCANN | : | Dinoflagellate Categorisation by Artificial Neural Network |
| ALIS | : | Automated Leafhopper Identification system |
| DAISY | : | Digital Automated Identification System |
| AIMS | : | Automatic Identification and characterisation of Microbial populationS |
| ABIS | : | Automated Bee Identification System |
| SPIDA-web | : | SPecies IDentification, Automated and web accessible |
| AIICHLA | : | Automated Insect Identification through Concatenated Histograms of Local Appearance |
| STFT-DA | : | Short-time Fourier Transform and Discriminant Analysis |
| MHBI | : | Monogenean Haptoral Bar Image |
| SIFT | : | Scale-invariant feature transform |
| RBC | : | Red Blood Cells |
| SRV | : | Semantically Related Visual |
| CBIR | : | Content-Based Image Retrieval |
| ASM | : | Active Shape Model |
| AAM | : | Active Appearance Model |
| LBP | : | Local Binary Patterns |
| PCA | : | Principal Component Analysis |
| SBS | : | Sequential Backward Selection |
| SFS | : | Sequential Forward Selection |
| SFFS | : | Sequential Forward Floating Selection |
| SMO | : | Sequential Minimal Optimization |
| RBF | : | Radial Basis Functions |
| SVM | : | Support Vector Machine |
| SRV | : | Semantically-Related Visual |
| DT | : | Decision Trees |
| RBF | : | Radial Basis Function |
| BPN | : | Back Propagation |
| QDA | : | Quadratic Discriminant Analysis |
| TIF | : | Tagged Image File |
| GUI | : | Graphical User Interface |
| MSE | : | Mean Square Error |
| LOO | : | Leave-One-Out |

## LIST OF APPENDICES

Appendix A: Leica DFC320 camera specifications …… 103

Appendix B: MATLAB codes……. 104

# CHAPTER 1: INTRODUCTION

Monitoring biodiversity in consonance with the study of biological populations and their growth is important and requires species identification which is time consuming and reliant upon expert ecologists. Hence the demand for automated species identification has increased over the last two decades. Research efforts in identification of species include specimens' image processing, extraction of identical features, followed by classifying them into correct categories. Recently, automation of data classification is primarily focussed on images and incorporated analyse or the images that have become easier due to advance developments in computational technology.

On the other hand, one of the key challenges to control diseases in fish population is achieving precise identification of fish parasites. Parasitic organisms have categorical homogeneous morphology, hence, pattern recognition techniques can be used to identify them (Castañón, Fraga, Fernandez, Gruber, & da F. Costa, 2007). Monogeneans are used in this study because they are worthy taxons for investigation (Brooks & McLennan, 1993). There might be around 25000 species of monogenean in the world while barely 4000 of them are currently known (Whittington, 1998). Monogeneans are flatworm clade that have advanced adaptive radiation (Brooks & McLennan, 1993), with different structural designs in the attachment organs (Kearn, 1994), which are usually used for species identification. In particular, haptoral attachment organ is characterized by sclerotized structures such as anchors, bars and hooks. The morphology of these organs are usually unique to monogenean species (Boeger & Kritsky, 1993) and are used as diagnostic characters in taxonomy (Vignon, 2011a, 2011b).

Automated classification of specimens` images requires development of models and methods that are able to characterize species` images based on the texture or shape of

objects to extract important visual information for classification. In monogenean identification models, all approaches are currently dependent on significant manual input during image processing and feature extraction such as specifying morphological landmark features. The adopted manual methods on each image, substantially slows the process of identification and classification. Hence, it was aimed to develop a fully automated identification model for monogeneans which is robust to variable imaging conditions, damaged specimens and variations within species.

### 1.1 Overview

Monogeneans are flatworms (Platyhelminthes) that are primarily found on gills and skin of fishes. Monogenean parasites have attachment appendages at their haptoral regions that help them to move about the body surface and feed on skin and gill debris. Haptoral attachment organs consist of sclerotized hard parts such as hooks, anchors and marginal hooks. Monogenean species are differentiated based on their haptoral bars, anchors, marginal hooks, reproductive parts` (male and female copulatory organs) morphological characteristics and soft anatomical parts. The complex structure of these diagnostic organs and also their overlapping in microscopic digital images are impediments for developing fully automated identification system for monogeneans (Ali, Hussain, Bron, & Shinn, 2011, 2012; Strona, Montano, Seveso, Galli, & Fattorini, 2014). In this study images of hard parts of the haptoral organs such as bars and anchors are used to develop a fully automated identification technique for monogenean species identification by implementing image processing techniques and machine learning methods.

According to the quality of captured images, images of eight monogenean species namely *Sinodiplectanotrema malayanus, Diplectanum jaculator, Trianchoratus pahangensis, Trianchoratus lonianchoratus, Trianchoratus malayensis, Metahaliotrema ypsilocleithru, Metahaliotrema mizellei* and *Metahaliotrema similis* were selected to

develop an automated technique for identification. Since recognition of monogeneans is based on morphometric features of their hard parts (Lim & Gibson, 2010), images of the hard haptoral organs such as anchors and bars were captured. All acquired images were indexed according to slide tags and stored in image database. One of the biggest challenges of monogenean images were their complexity in terms of messy background and overlapping of anchors and bars. Although many efforts were made to acquire clear images but still some overlapping and clutters were unavoidable. Here, image pre-processing is needed to omit redundant information and to highlight reliable features in order to prepare images for feature extraction. According to features such as: length of bounding box, width of bounding box, centre of bounding box, orientation of bounding box, perimeter, perimeter density, area, area density, Euler number, entropy and major axis length, a feature vector was extracted. By use of Linear Discriminant Analysis (LDA) feature selection technique, the feature vector was transformed to lower dimensional feature vector. The extracted and selected features achieved in previous stages were then used as input to K-Nearest Neighbour (KNN) and Artificial Neural Network (ANN) classifiers to train the system based on training set and test the testing dataset based on trained model.

The presented model in this study empowers fast and accurate fully automated classification of monogeneans to the species level.

### 1.2 Research questions

- How to apply image processing on 2D digitized monogenean specimens` images to prepare them for classification?
- Which classification methods can be used for monogenean species automated identification?
- What is the probability of correct identification and classification of monogenean species?

### 1.3 Objectives of the study

- To prepare a 2D image database of eight selected monogenean species
- To compare the accuracy of two machine learning techniques (i.e. K-Nearest Neighbour and Artificial Neural Network) in identifying/ classifying selected species of monogeneans in Malaysia.
- To develop an automated species identification/ classification model for selected species of monogenean.

### 1.4 Scope of the study

Images of eight monogenean species namely *Sinodiplectanotrema malayanus, Diplectanum jaculator, Trianchoratus pahangensis, Trianchoratus lonianchoratus, Trianchoratus malayensis, Metahaliotrema ypsilocleithru, Metahaliotrema mizellei* and *Metahaliotrema similis* were used to develop an automated model for identification. K-nearest neighbour (KNN) and Artificial Neural Network (ANN) were applied to classify the monogenean specimens based on the extracted features. The automated identification model was implemented in two, preliminary and extended phases. The preliminary automated classification model was implemented by adoption of samples of four species and the extended model was implemented based on samples of eight species.

### 1.5 Outline of the study

Chapter One: In this chapter, the general research framework, which introduces automated identification technique for monogenean species, besides presenting the research questions, objectives and scope of this study was explained.

Chapter Two: This chapter contains the literature review on monogeneans, images acquisition, database for automated systems, images processing, feature extraction, feature selection and classification techniques.

Chapter Three: This chapter contains the materials and methods that have been applied in two preliminarily and extended models, describing specimens' collection and image acquisition, followed by the details about construction of digital images for database and finally explanations about image processing techniques, feature extraction and selection. Finally, this chapter reports how KNN and ANN classification methods were adopted in this study.

Chapter Four: This chapter presents the results of feature selection, classification and evaluation in both preliminarily and extended models.

Chapter Five: This chapter discusses about the results of the development of automated identification model for monogenean images. It also contains comparison of current study with previous studies, confession about constrains and limitations, declaration on the future works for enhancement of automated identification model and finally conclusions.

# CHAPTER 2: LITERATURE REVIEW

Environmental monitoring based on correct identification of specimens according to their correct species or groups is an essential and cost effective task (Larios et al., 2008). The demand for recognition of species has significantly influenced biologists to increase the facilities and proper supply of skills for identification and classification task. In addition, in some cases identification of species group is limited to available human domain experts ( Ali, Hussain, Bron, & Shinn, 2011). Although there was undeniable potential, the development of automated identification systems has been hampered by some taxonomists who hesitated to embrace different methods of species identification (Kiranyaz et al., 2011). The main reason that influenced developing image based identification system was eagerness of taxonomist to reduce the time consumed for analysing samples (Benfield et al., 2007) and to significantly cut down the costs. Culverhouse et al. ( 2003) have shown that categorizing specimens from species which have significant variations in their morphology is taxing. They also demonstrated that the returned accuracy by trained personnel and experts for discriminations and labelling specimens is expected to be in the range of 64% to 95% which is within the performance range of automated methods.

Automated classification of specimens` images to their corresponding species requires development of models and methods that are able to characterize a species` morphology and apply this knowledge for their recognition. These systems should be combined with databases of images or text based information (Martins, Oliveira, Nisgoski, & Sabourin, 2013). Selection of segmentation, feature extraction and classification techniques are dependent on identification taxonomic rank. For example identification and classification at species level require more detail information compare to family level. The aim is discovering semantic concepts from images to identify and

classify objects of interest. For characterization of these objects, efficient features are required to build computational models (Castañón et al., 2007). Object curvature (Riggs, 1973) from respective contour, morphological and geometrical measurements are good examples of different characterization methods.

Previously, many systems have been developed for identification of biological objects at different levels. In 1996, the Dinoflagellate categorization (DiCANN) system, based on neural networks (Culverhouse et al., 1996) was developed. Later, forensic identification of mammals according to their single hair patterns under a microscope was investigated by Moyo et al. (2006), while Yuan et al. (2006) discussed the identification of rats up to the species level from images of their tracks. Automation of species identification systems proved that these tedious tasks could be accomplished in more feasible and efficient manner while minimising sources of errors (Kay, Shinn, & Sommerville, 1999). Examples of such systems are Automated Leafhopper Identification system (ALIS) (Dietrich & Pooley, 1994), Digital Automated Identification System (DAISY) (O'Neill, Gauld, Gaston, & Weeks, 2000), Automatic Identification and characterization of Microbial Populations (AIMS) (Jonker et al., 2000), Automated Bee Identification System (ABIS) (Arbuckle, Schröder, Steinhage, & Wittmann, 2001), BugVisux (Hanqing & Zuorui, 2002), automated identification of bacteria using statistical methods (Trattner, Greenspan, Tepper, & Abboud, 2004), an automated identification system which estimates whiteflies, aphids and thrips densities in a greenhouse (Cho, Choi, Qiao, Ji, & Kim, 2008), species identification, automated and web accessible (SPIDA-web) (Russell, Do, Huff, Platnick, & MacLeod, 2007), But2fly (Liu, Shen, Zhang, & Yang, 2008), Automated Insect Identification through Concatenated Histograms of Local Appearance (AIICHLA) (Larios et al., 2008), an automated identification system for algae (Coltelli, Barsanti, Evangelista, Frassanito, & Gualtieri, 2014), automatic recognition of biological particles in microscopic images

(Ranzato et al., 2007), automatic species identification of live moths (Mayo & Watson, 2007) automated image-based phenotypic analysis in zebrafish embryos (Vogt et al., 2009), automatic recognition system for some cyanobacteria using image processing techniques and ANN approach (Mansoor, Sorayya, Aishah, Mogeeb, & Mosleh, 2011), automatic detection of malaria parasites for estimating parasitemia (Savkare & Narote, 2011), automated weed classification with local pattern-based texture descriptors (Ahmed, Kabir, Bhuyan, Bari, & Hossain, 2014), automated processing of imaging data through multi-tiered classification of biological structures illustrated using caenorhabditis elegans (Zhan et al., 2015), automated identification of copepods using digital image processing and artificial neural network (Leow, Chew, Chong, & Dhillon, 2015), automatic plant species identification using sparse representation of leaf tooth features (Jin, Hou, Li, & Zhou, 2015), automated system for malaria parasite identification (Savkare & Narote, 2015), a software system for automated identification and retrieval of moth images based on wing attributes (Feng, Bhanu, & Heraty, 2016), automatic wild animal monitoring by identification of animal species in camera-trap images using very deep convolutional neural networks (Gomez & Salazar, 2016), automated identification of anastrepha fruit flies in the fraterculus group (Perre et al., 2016) and automated identification of fish species based on otolith contour, using short-time Fourier transform and discriminant analysis (STFT-DA) (Salimi, Loh, Dhillon, & Chong, 2016). Automated systems for biological species are summarized in Table 2.1.

**Table 2.1: Examples of significant automated species identification systems.**

| System | No. of classes | Classification method | Accuracy (%) | Reference |
|---|---|---|---|---|
| Automated Object Recognition Of Blue-Green Algae | 9 | Discriminant Analysis | 98 | (Thiel, Wiltshire, & Davies, 1996) |
| Automatic Classification Of Field-Collected Dinoflagellates | 23 | ANN : Radial Basis Function (RBF) & Back Propagation of error variant (BPN) | 83 | (PF et al., 1996) |
| Automatic Identification Of Human Helminth Eggs | 12 | ANN | 86 - 90 | (Yang, Park, Kim, Choi, & Chai, 2001) |
| Automate Identification Of Bees | 13 | Linear Discriminant Analysis | 98 - 99.8 | (Schroder et al., 2002) |
| Automatic Diatom Identification | 43 | Decision trees and k-nearest neighbour | 82-84 | (Jalba, Wilkinson, Roerdink, Bayer, & Juggins, 2005) |
| Automatic Identification Of Whiteflies, Aphids And Thrips | 50 | ANN | 93-100 | (CHO et al., 2008) |
| Automatic Identification Of Live Moths | 35 | WEKA: Naïve Bayes, Instance-based learning, Decision trees, Random forests and SVM | 85 | (Mayo & Watson, 2007) |
| Automatic recognition system for some cyanobacteria | 4 | ANN | 95 | (Mansoor et al., 2011) |
| Automated weed classification | 2 | Template matching & SVM | 88-98.5 | (Ahmed et al., 2014) |
| Automated Insect Identification | 4 | Automated insect identification, Kadir entropy detector and PCBR | 82-95 | (Larios et al., 2008) |
| Automated Taxon Identification Of Teleost Fishes | 420 | k-nearest neighbour | 72 | (Parisi-Baradad et al., 2010) |
| Automated Real-Time Dynamic Identification Of Flying And Resting Butterfly | 10 | Random tree | 85 | (Loke, Egerton, Cristofaro, & Clementson, 2011) |

**Table 2.1: Continued.**

| | | | | |
|---|---|---|---|---|
| Automatic Identification Of Diatoms | 12 | BP neural networks | 94 | (Luo et al., 2011) |
| Automatic Insect Classification | 10 | SVM | > 90 | (Le-Qing & Zhen, 2012) |
| Automated Identification And Retrieval Of Moth Images | 50 | semantically-related visual (SRV) | 85 | (Feng & Bhanu, 2013) |
| Automatic Identification Of Species | 740 | ANN | 91-93 | (Hernández-Serna & Jiménez-Segura, 2014) |
| Water Monitoring - Automated And Real Time Identification And Classification Of Algae | 23 | ANN : Self Organizing Map (SOM) | 98 | (Coltelli et al., 2014) |
| Automatic Identification Of Butterfly Species | 5 | ANN | 98 | (Kaya, Kayci, & Uyar, 2015) |
| Automated System For Malaria Parasite Identification | 2 | SVM | 80 | (Savkare & Narote, 2015) |
| Automatic Plant Species Identification | 8 | KNN & BP Neural Network | 76-79 | (Jin et al., 2015) |
| Automated identification of copepods | 8 | ANN | 93.13 | (Leow et al., 2015) |
| Automated identification and retrieval of moth images | 50 | SRV attributes | 34-70 | (Feng et al., 2016) |
| Automatic wild animal identification | 26 | Convolutional Neural Networks | 88.9-98.1 | (Gomez & Salazar, 2016) |

## 2.1 Monogeneans

Monogeneans are members of the Platyhelminthes and without intermediate host, they have direct life cycle (Woo & Leatherland, 2006). Usually, Monogeneans live on lower aquatic invertebrates or gills, skin or fins of fishes as host. Monogeneans have their greatest diversity on fishes. Currently in Malaysia, over 200 species of monogeneans have been described from 60 species of fishes (35 and 25 species of

freshwater and marine fish, respectively), three species of turtles and one species of frog (Lim, Tan, & Gibson, 2010). Monogeneans commonly move on the body surface and feed from skin mucus and debris on the gill. They have appendage attachments in their anterior and posterior (haptoral) regions (Figure 2.1) that are used to prevent physical dislodgement from the host.

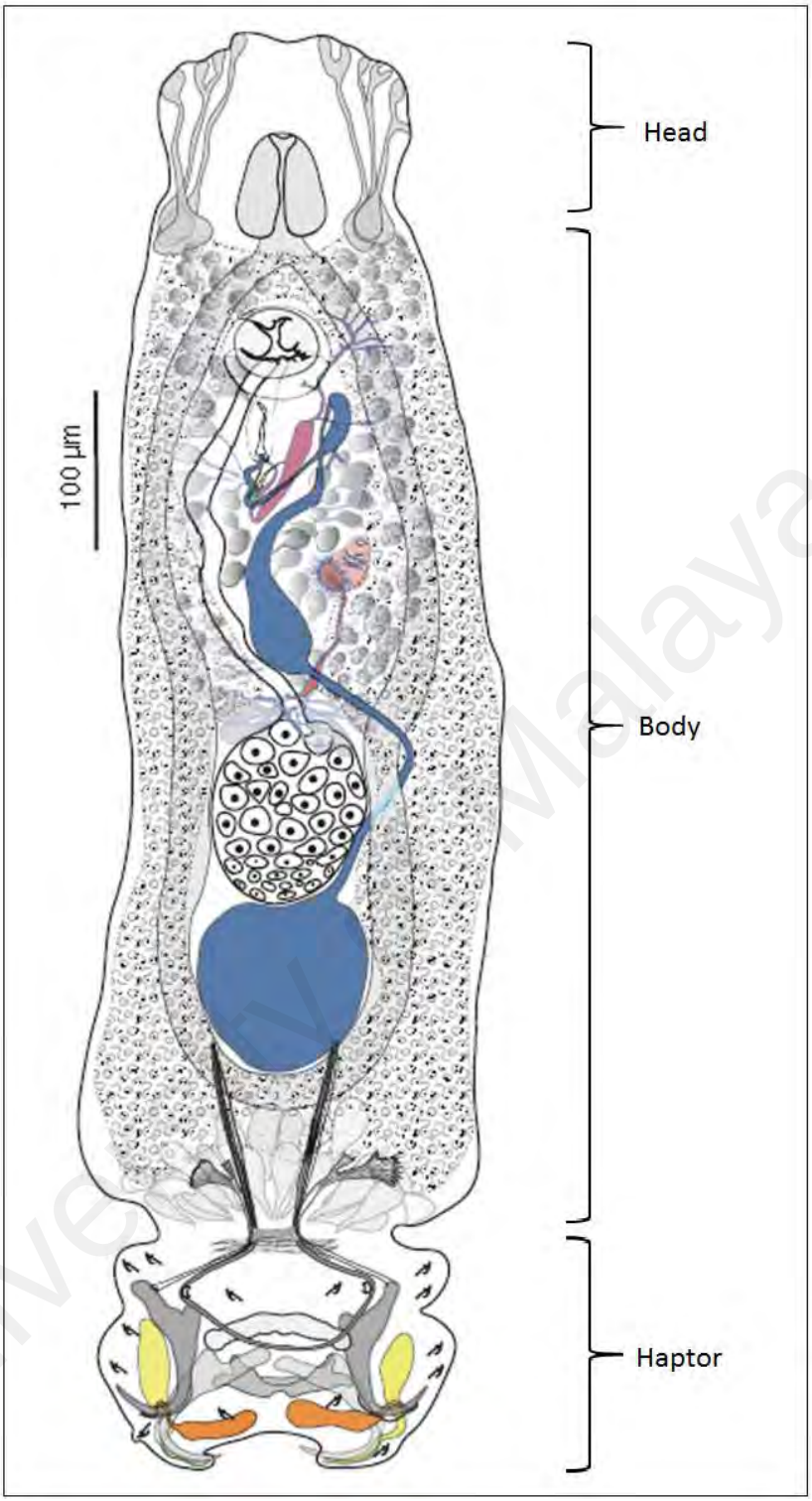


**Figure 2.1: Illustration of a monogenean worm consisting of three main parts (i.e. head, body and haptor parts) (Figure adapted from (Abu, Lim, Sidhu, & Dhillon, 2013)).**

The haptoral organs consist of hard, sclerotized structural parts such as anchors and bars. Since characters that can be extracted from haptoral hard parts of monogenean are

prominent, The features from these characters are significant basis for taxonomic classification and identification of monogenean (Bykhovsky & Nagibina, 1978).

Taxonomists essentially use morphological analysis from sclerotized organs such as anchors and bars in classification of monogenean due to sharp and informative qualitative variation in the latter. Investigations on morphometric characteristics of hard sclerotized organs of monogeneans have been done in terms of evolutionary ecology (Poisot & Desdevises, 2010) and also systematics (Shinn, Gibson, & Sommerville, 2001). Since the form of hard sclerotized organs will not simply change after compression while mounting onto slides, they are ideal for geometric morphometric analysis (Lim & Gibson, 2009). Anchors and bars of monogenean are species specific with respect to their shape and size. To date, in many studies (Pariselle et al., 2011; Rodríguez-González, Míguez-Lozano, Llopis-Belenguer, & Balbuena, 2015; Vignon, 2011a), the data from geometric morphometric analysis of monogenean`s anchors and bars applied in identification and classification of monogenean.

**2.2 Image Acquisition**

Coltelli et al. (2014) believed that image acquisition is the most important step in designing an automated system and capturing images should be well-focused with less complexity. The acquisition condition should be clearly defined and kept equal for all images, later labelled by expert taxonomists. In microscopic images, magnification might be different in the data set and it is important to specify scales in each image to prevent system confusion. Figure 2.2 illustrates three images of *Euryhaliotrema* organs. In Figure 2.2 (a) there is a copulatory organ inside the black circle which is even difficult to be recognised by human eyes. In Figure 2.2 (b) and (c), anchors and bars are illustrated but still the outline of anchors and bars are not recognisable and the organs are not separately distinctive because of overlapping of anchors and bars. All of these

complications in the images might be the result of bad focus, lack of light and contrast settings or other image acquisition factors.

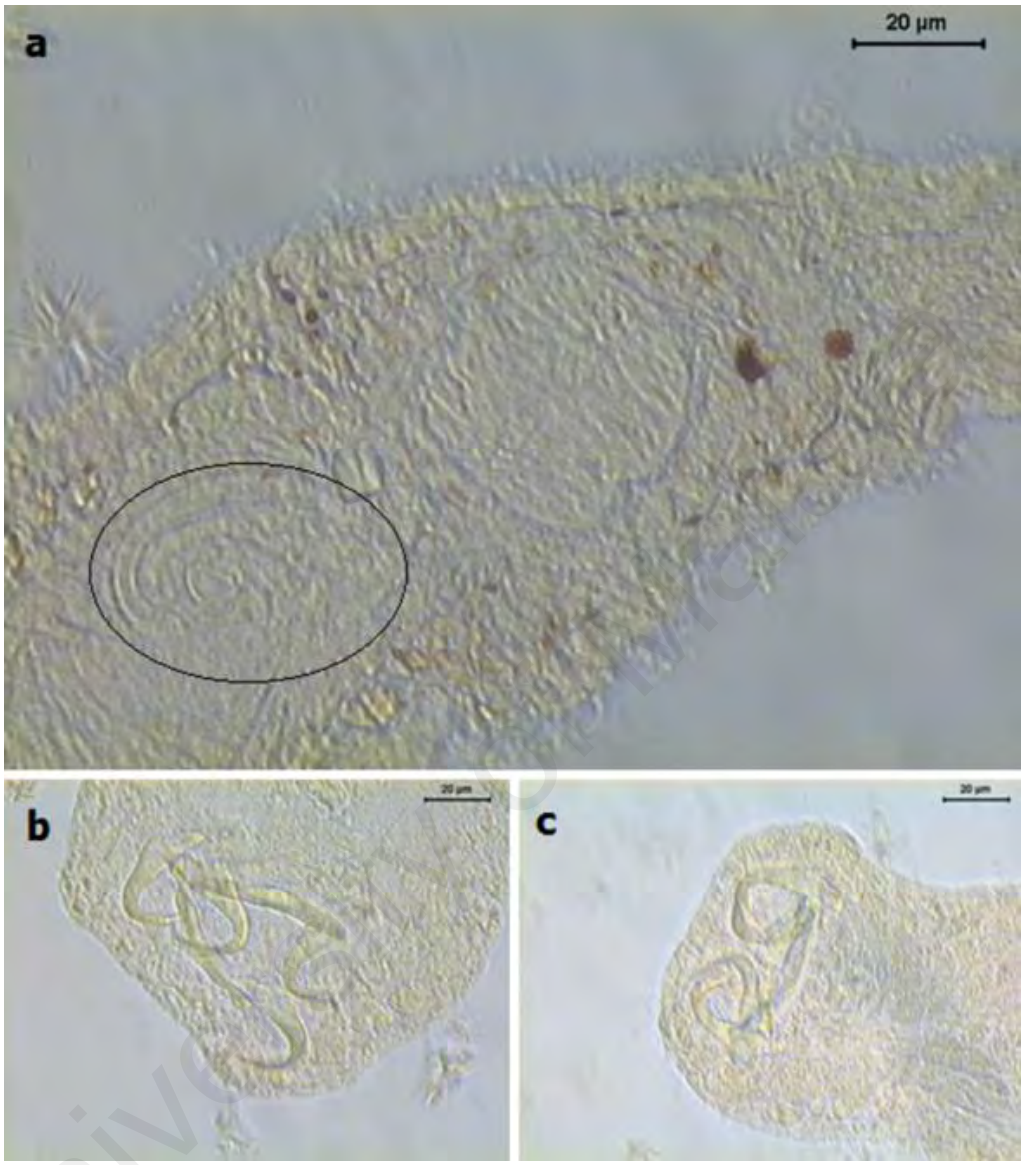


**Figure 2.2: Illustration of image acquisition problems of *Euryhaliotrema* during digitization. a) Noise and debris in the images makes recognition of copulatory organ difficult. b) Bad focus on bars and anchors. c) Messy background of anchors and bars.**

One of the challenges faced in creating image database is the lack of standard imaging condition during image acquisition. A method to control imaging condition for automated identification of stonefly larvae was proposed in which the imaging

apparatus in this system posed and rotated the specimens under the microscope and captured images in standard and consistent conditions (Larios et al., 2008).

### 2.3 Database

In all automated identification systems which are based on species images, the systems are connected to a database of specimens` digital images that contain different number of dominant categories. Data in a database is commonly divided into two sets, one for training the classifiers and the other set for testing the classification. The number of species` images used for training differs widely between systems and is determined according to the applied classifiers. Table 2.2 demonstrates some databases used in automated identification systems. Abu et al. (2013) proposed an image retrieval framework for monogeneans that contains two databases, the monogenean image database and the Monogenean Haptoral Bar Image (MHBI) Fish ontologies. In this study, an ontology framework improves the relevancy of the training set to collect the most relevant images to be used. In the stonefly identification system (Larios et al., 2008), 263 specimens of four species were collected and approximately ten images of each specimen were captured through their imaging apparatus. The database used in the diatom identification system (Jalba et al., 2005) includes two sets of files, the first consists of 120 images of six species from one genera and the second set contains 781 images of 37 species from different genera. The microscopic images in this system varied in terms of quality and noise of contour (contours being noisy) but the system was able to handle the noise. An automated identification systems for classification of tree species (Martins et al., 2013) employed a database of 112 species` images. The microscopic images were acquired with 100× magnification and labelled by dendrologists. The database contained 2240 images, 20 images from each species for training and testing the system. They used 40% of their data for training (8 images for

each species), 20% for validation (4 images for each species) and 40% for testing (4 images for each species).

**Table2.2: Example of some species image databases applied in automated identification systems.**

| Organism | Level of classification | No. of classes | No. of training set | No. of testing data | Total No. of Images | Reference |
|---|---|---|---|---|---|---|
| Monogenean | Species | 6 | 148 | 19 | 167 | (Abu et al., 2013) |
| Stonefly | Species | 4 | 50 of each spp | 50 of each spp | 1240 | (Larios et al., 2008) |
| Diatom | Species | 43 | - | - | 901 | (Jalba et al., 2005) |
| Softwood and Hardwood forest species | Species | 112 | 8 of each spp | 8 of each spp | 2240 | (Martins et al., 2013) |
| Copepods | Genus | 5 | 30 of each spp | 20 of each spp | 400 | (Leow et al., 2015) |

## 2.4 Image processing

The aim of image processing in the system is to transform digital images to a standard pose (Gonzalez & Woods, 2007) and achieving recognizable objects on a uniform background. In this step, image noises should be removed, also contrast and dynamic range of image have to be improved. Image enhancement can be carried out by manual or automatic methods. Manual methods such as the ones carried out using ImageJ (Kiranyaz et al., 2011; Mayo & Watson, 2007) or Photoshop (Larios et al., 2008), may yield better image pre-processing results but it is advisable to use fully automated methods to build systems with large number of images as the manual image processing methods require longer processing time.

Digital images of species, especially microscopic images, usually contain dust or other noise artefacts. Noise makes neighbouring pixel values clutter (Trattner et al., 2004), so it should be reduced by smoothing methods of filtering. The efficiency of removing noise by filtering could be more if it be according to type of noise. Amplifier

or Gaussian, salt and pepper, film grain, non-isotropic, speckle and periodic noise are the most common types of noise. Noise reduction filters can be divided into two categories: linear filters and non-linear filters (Mythili & Kavitha, 2011). Median filtering (Bovik, Huang, & Munson, 1987) is a non-linear filtering which is commonly applied to digital microscopic image (Avci & Varol, 2009; Hernández-Serna & Jiménez-Segura, 2014; Saraswat & Arya, 2014; Weeks, O'Neill, Gaston, & Gauld, 1999). Leow et al. (2015) applied median filtering with 10 × 10 kernel in automated identification system for copepods to suppress the salt and pepper noise created from the water in images.

Image quality is highly affected by illumination, contrast, focus and acquisition resolution (Castañón et al., 2007). Variation in illumination may be caused by different types of lenses (Arce, Wu, & Tseng, 2013) and light sources (Bradbury & Bracegirdle, 1998; Saraswat & Arya, 2014). Histogram equalisation can be applied to reduce variation in illumination (Castañón et al., 2007). Enhancing contrast by stretching the histogram of digital image will spread the brightest and darkest pixel values of grey levels which will later assign to white and black. Table 2.3 shows some image processing algorithms, introduced by Gonzales and Wood (Gonzalez & Woods, 2007).

**Table 2.3: Common image processing algorithms used in automated species identification systems.**

| **Algorithm** | **Comments** | Reference |
|---|---|---|
| Noise Reduction | Linear filtering, Non-linear filtering | (Gonzalez & Woods, 2007) |
| Image Enhancement | Sharpening the image | |
| | Edge highlighting | |
| | Contrast improvement | |
| Image Restoration | Clearing away the blurriness made by linear motion | |
| | Clearing away the optical misrepresentation | |
| | Clearing away the periodic interference | |
| Image Segmentation | Separation of particular shapes from background | |
| | partitioning an image | |

The fundamental step after image processing and before feature extraction and classification is segmentation (Haralick & Shapiro, 1992). Segmentation separates the background from the foreground and is important in computer vision since it finds the location of pixels that can be classified as an object. Pixels with common characteristics (for example texture or colour distribution) are grouped according to the selected segmentation algorithms. Although automated segmentation of specimens from background may still encompass debris and clutter, robust automated systems can categorize species satisfactorily (Culverhouse et al., 1996). Recognition of image parts which belong to an object of interest is often more effective when making use of boundaries and shape information extracted by segmentation methods. The Grabcut algorithm (Rother, Kolmogorov, & Blake, 2004) is a segmentation technique used in automated identification of species systems (Hernández-Serna & Jiménez-Segura, 2014) to remove background. In this technique, hard segmentation made by iterative graph-cut optimization is combined with border matting to get rid of mixed and blurred pixels on boundaries of object. Edge detection (Gonzalez & Woods, 2007) is another common segmentation technique that can be achieved by filters such as Canny's (Canny, 1986) or Sobel's (Gonzalez & Woods, 2007). Both sobel and canny detectors were applied for image segmentation in the automatic algal identification system (Natchimuthu, Natchimuthu, Chinnaraj, Parthasarathy, & Senthil, 2013), due to the significant edges and contours of the objects. There are generally six methods for object segmentations: thresholding (Gonzalez & Woods, 2007), fuzzy theory-based, Patial Differential Equatuin-based, Artificial Neural Network-based, region-based and edge-based methods (Kang, Yang, & Liang, 2009; Khan, 2014).

Thresholding is the most common technique in which binary images are produced according to cut-off value. This method can be mainly subclasses to dynamic, global and local thresholding techniques (Table 2.4) ( Kang, Yang, & Liang, 2009; Singh,

Tomar, & Maurya, 2012). Sometimes, there are specimens overlapping that makes object detection difficult, especially in microscopic images. Distance transforms and watershed transforms can be applied to separate overlapping specimens (Di Ruberto, Dempster, Khan, & Jarra, 2000; Savkare & Narote, 2011).

**Table 2.4: Thresholding techniques used in automated species identification systems.**

| Techniques | Subclasses | Reference |
|---|---|---|
| Dynamic | Watershed thresholding | (Doncic, Eser, Atay, & Skotheim, 2013) |
| Global | Otsu thresholding | (Savkare & Narote, 2015) |
| Local | Adaptive thresholding | (Jin, Hou, Li, & Zhou, 2015) |

## 2.5 Feature Extraction and Selection

Features extracted from digital images are used to train classifiers. Therefore, extraction and selection of best features is important. Classes of features can be grouped into feature vectors which create a representation of objects of interest in the image and should contain taxonomic information. Using all extracted features in classifier will cause heavy computational effort, therefore, selection of effective features is an important task (Sang-Hee, 2010). Optimization of number of features selected for training classifiers is done using feature selection techniques (Choras, 2007). Good performance of both extracted and selected features depends on type of system`s classifiers and the analysing data (Kiranyaz et al., 2011). If employed classifiers are strong enough, even with small number of features, the method may yield successful results (Larios et al., 2008).

### 2.5.1 Feature Extraction

The most salient types of features in images are shape, colour and texture (Islam, Dengsheng Zhang, & Guojun Lu, 2008; Ping Tian, 2013; Shih, Huang, Wang, Hung, &

Kao, 2001). Feature extraction in automated systems may depend on the level of identification, which means features for detection at the order level are different from those at the species level. Some local features such as sparse coding spatial pyramid matching (Lu, Hou, Lin, & Liu, 2010), concatenated feature histogram (Larios et al., 2008) and bag of words (Wen, Guyer, & Li, 2009) which are based on Scale-invariant feature transform (SIFT) (Lowe, 2004; Wang, Lin, Ji, & Liang, 2012) may not extract enough information for identification of high level categories like species. In the automated system for whiteflies, aphids and thrips identification, features such as size, shape of boundary and colour components were considered (CHO et al., 2008) and due to different attached part of each insect, morphological boundary was not used and only three colour components and size were applied as feature. Figure 2.3 illustrates content based features which are common in automated identification systems (Li, Tseng, Hsieh, Yang, & Huang, 2014). Shape representation techniques (Yang, Kpalma, & Ronsin, 2008) are applicable for shape feature extraction (Table 2.5). The techniques in Table 2.5 are classified by their processing approaches.

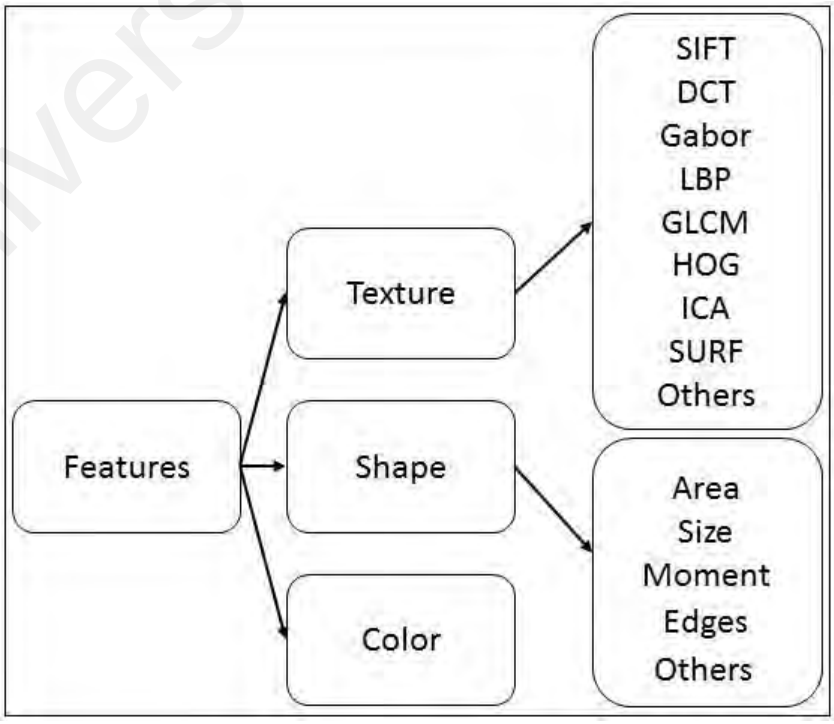


**Figure 2.3: Content based features.**

**Table 2.5: Overview of shape representation techniques.**

| Shape Features | | |
|---|---|---|
| Shape parameters | Center of gravity | |
| | Axis of least inertia | |
| | Average bending energy | |
| | Eccentricity | Principal axes method |
| | | Minimum bounding rectangle |
| | Circularity ratio | |
| | Ellipse variance | |
| | Rectangularity | |
| | Convexity | |
| | Solidity | |
| | Euler number | |
| | Profiles | |
| | Hole area ratio | |
| One dimensional function for shape representation | Complex coordinates | |
| | Centroid distance function | |
| | Tangent angle | |
| | Contour curvature | |
| | Area function | |
| | Triangle-area representation | |
| | Chord length function | |
| Polygonal approximation | Merging methods | Distance threshold method |
| | | Tunnelling method |
| | | Polygon evolution |
| | Splitting methods | |
| Moments | Boundary moments | |
| | Region moments | Invariant moments |
| | | Algebraic moment invariants |
| | | Zernike moments |
| | | Radial Chebyshev moments |
| | | Homocentric polar-radius moments |
| | | Orthogonal Fourier-Mellin moments |
| | | Pseudo-Zernike moments |

**Table 2.5: Continued.**

| | | |
|---|---|---|
| Spatial interrelation feature | Adaptive grid resolution | |
| | Bounding box | |
| | Convex hull | |
| | Chain code | Basic chain code |
| | | Differential chain codes |
| | | Re-sampling chain codes |
| | | Vertex chain code |
| | | Chain code histogram |
| | Smooth curve decomposition | |
| | ALI-based representation | |
| | Beam angle statistics | |
| | Shape matrix | Square model shape matrix |
| | | Polar model shape matrix |
| | Shape context | |
| | Chord distribution | |
| | Shock graphs | |

In the automated system for malaria parasites, area, perimeter, minor and major axis of red blood cells (RBC) were calculated as shape feature components (Savkare & Narote, 2015). Texture features consist of kurtosis, momentum, standard deviation and mean of RBC and intensity values of the green channel were considered as colour features. Local Binary Patterns (Ojala, Pietikainen, & Maenpaa, 2002) were considered as texture descriptors and they are applied in images analysis. Kaya et al (Kaya et al., 2015) extracted four texture features: average, correlation, entropy and energy from the local binary pattern matrix in their automated identification system for butterfly species. In the automated identification and classification system for algae (Coltelli et al., 2014), dissimilarity measurement, centroid distance spectrum, points of contours and some densitometry and morphological features like area, ferret diameters, extinction, centre of gravity coordinates and etc. were calculated. Hernandez-Serna et al. (2014) proposed an automated system which is applicable for identification and classification of plants,

fishes and butterflies. Their approach in this system extended to three different taxonomic groups, therefore, extraction of features should be as general as possible in the way that it could be applied to all species. They used area, perimeter, diameter, compatibility, compactness and solidity as geometrical features, uniformity, median, entropy, variance, inertia, homogeneity and co-occurrence as texture features and Hu invariant set of moments and related moment invariants as morphological features (Ming-Kuei Hu, 1962; Flusser & Suk, 1993).

Feng and Bhanu (2013) developed a system which adopted semantically related visual (SRV) attributes. They claimed that shape, texture and colour may fail in validity if the images are visually complex and have semantic contents. According to the results of their research, it is notable that in all iterations accuracy of using SRV is higher than CBIR. Figure 2.4 illustrates the comparison of mean accuracy of SRV and Content-Based Image Retrieval (CBIR) approaches in categorization of species in five iterations.

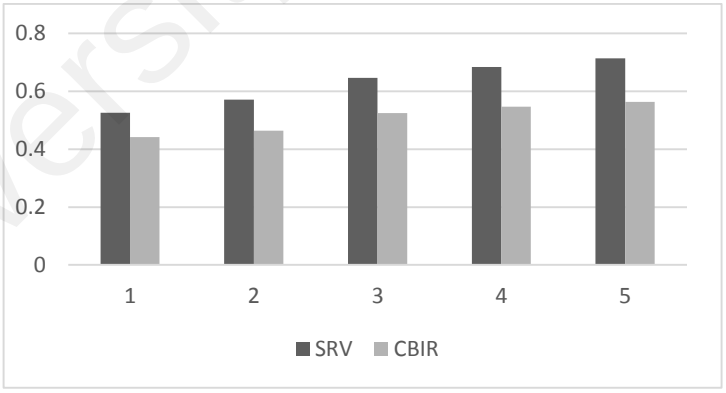


**Figure 2.4: The comparison of mean accuracy of SRV and CBIR approaches in categorization of species in five iterations**

Other features that have been applied in detection and categorization of specimens are classical features such as branch length similarity entropy ( Kiranyaz et al., 2011; Huddar, Gowri, Keerthana, Vasanthi, & Rupanagudi, 2012), corner based features,

edge, ridge, curve, shape descriptors like Fourier descriptors, texture features like co-occurrence, histogram intensity and gradient (Haralick, Shanmugam, & Dinstein, 1973). Also some other feature extraction methods that can be named are Gabor packet based methods (Grigorescu, Petkov, & Kruizinga, 2002), Histogram of Oriented Gradient (Dalal & Triggs, 2005), Scale Invariant Feature Transform (SIFT), Active Shape Model (ASM) (Ali et al., 2012), Active Appearance Model (AAM) and Local Binary Patterns (LBP) (Quivy & Kumazawa, 2011).

### 2.5.2 Feature Selection

Feature selection is a process to identify relevant features while removing irrelevant and redundant features. Relevant features should be informative, fast in computing and also invariant to noise or given transformations. Feature selection is an ideal way in many pattern recognition problems to reduce the dimensions of extracted features. When there are high-dimensional samples but limited incorporated information, the best action is selection of the most informative data (Lei, Liao, & Li, 2012). Now, the decision whether a feature is relevant, redundant or not, are aspects that involves in feature selection operations. The role of selecting features lies in improving the prediction process, correlation coefficient of regression algorithms and comprehensibility of learning results (Karagiannopoulos, Anyfantis, Kotsiantis, & Pintelas, 2007). Table 2.6 shows some feature selection algorithms (Kudo & Sklansky, 2000). Principal component analysis (PCA) (Jolliffe, 2002) is multivariate statistical technique, adopted by DAISY to select important features of images. Due to big amount of detailed information collected by this technique, acquired features are convenient for identification at species level (Wang et al., 2012). Ali et al. (2011) used the assessment of Sequential backward Selection (SBS), Sequential Forward Selection (SFS) and Sequential Forward Floating Selection (SFFS) techniques (Ververidis & Kotropoulos, 2008) for selecting proper features for monogenean classification and the

results indicated that of the 25 features, 21 were the best in classification of *Gyrodactylus* species performances. Feature selection results are dependent on the size of the training data as in (Jain & Zongker, 1997), the quality of feature selection for small data is low and as the training size increases, the quality improves.

**Table 2.6: Example of feature selection algorithms used in automated species identification systems (Table was adapted from Kudo & Sklansky (2000)).**

| Algorithm | Subset | Search Type |
|---|---|---|
| SFS, SBS | Looking for the best subset of given size | Sequential |
| GSFS(g), GSBS(g) | Looking for the best subset of given size | Sequential |
| PTA(l, r) | Looking for the best subset of given size | Sequential |
| GPTA(l, r) | Looking for the best subset of given size | Sequential |
| SFFS, SBFS | Looking for the best subset of given size | Sequential |
| BAB, $BAB^{+}$, $BAB^{++}$ | Looking for the best subset of given size | Sequential |
| RBAB, RBABM | Looking for the smallest acceptable subset | Sequential |
| GA | Looking for optimal combined size and error rate subset | Parallel |
| PARA | Looking for optimal combined size and error rate subset | Parallel |

#### 2.4.2.1 Linear Discriminant Analysis (LDA)

One of the common methods for feature selection is Linear Discriminant Analysis (LDA) (Song, Mei, & Li, 2010). LDA selects independent and most informative features and it can be applied in machine learning, statistics and pattern recognition to detect a linear composition of features that are able to classify classes of objects. The popularity of LDA method is for selecting features that preserves class separation. The goal of LDA is maximising between-classes covariance while minimizing in-class covariance, it means separation between multiple classes by maximizing the component axes (Cai, He, & Han, 2008). Therefore, besides projecting a feature space to smaller subspace, the class-discriminatory information is also maintained. In LDA feature selection, first, d dimensional mean vectors for n classes` dataset are determined. Subsequently, by computing in-between class and within-class scatter matrix, the eigenvectors and corresponding eigenvalues are calculated. Next, sorting eigenvectors

and picking eigenvectors with largest eigenvalues. Finally, the d × n dimension eigenvector is adopted to transform feature space to new subspace. Different elements of features statistically have different effects on the results of feature selection and they can be evaluated by eigenvector elements. Since there are many eigenvectors, LDA chooses some small elements of eigenvectors while evaluating the elements of extracted features (Song et al., 2010).

### 2.6 Classification

The idea of classification is to classify objects of interest based on a specific feature data set to discriminate between distinct classes. Performance of classifiers is highly affected by the segmentation and feature extraction process. Jain et al. (2000) proposed three categories of classifiers: similarity based, probabilistic and decision boundaries. Most of the classification methods are mentioned elsewhere, see (Loncaric, 1998; Zhang & Lu, 2004; Savkare & Narote, 2011), including structural, fuzzy, transform, neural network-based methods and many more. Some automated identification systems such as in copecodes (Leow et al., 2015) employ neural networks or learning algorithms when there are many classes and small number of samples, but some other systems such as in teleost fish (Parisi-Baradad et al., 2010) deal with huge numbers of samples and use other algorithms like K Nearest-Neighbour (KNN) (Duda, Hart, & Stork, 2012). Table 2.1 summarizes some automated identification systems adopting various kind of classification methods. Jalba et al. (2005) used k-nearest neighbour and C4.5 (Quinlan, 2014) algorithms as classification techniques for an automated identification of diatoms. In this system two types of feature vectors were adopted. Both types of feature vectors were constructed for top and bottom curvature spaces. Type-1 feature vector computes the number of peaks, mean curvature and variance for each cluster. Type-2 feature vector computes the mean curvature and variance of the points with the highest curvature for each cluster and the extent. The result with type-2 feature vectors was

84% and better than type-1 feature vectors. The average accuracy of this system when using C4.5 decision trees is higher compared to the rate of identification with human experts (43% to 86.5%). Mayo & Watson (2007) employed methods from the WEKA (Witten & Frank, 2005) machine learning toolkit such as Naïve Bayes, J48, IB1, IB5, Random forests and Sequential Minimal Optimization (SMO) classifiers. The results demonstrated that random forest and SMO classifiers achieved accuracy of 83%, better that other classifiers and by increasing the number of feature attributes, the accuracy reaches to 85%. In identification of species of *Gyroactylus* genus in fish ectoparasite (Ali et al., 2012), features which were extracted by Active Shape Models (ASM), implemented to two linear classifiers, Linear Discriminant Analysis (LDA) and KNN and two non-linear classifiers, Multilayer Perceptron (MLP) and Support Vector Machine (SVM). According to results of this study, LDA method accuracy was 85.71%, MLP method 95.59% and KNN classification accuracy of 98.75%. KNN was outperforming classifier since the testing dataset in identification of *Gyroactylus* species was 68 images and KNN was capable of classifying with limited number of dataset. Hayat Mansoor et al (Mansoor et al., 2011) proposed a system operating with ANN for identification of cyanobacteria genera images. This system recognized 71 of 80 images correctly and detection accuracy was reported as 95%. In classifying insects, Le-Qing & Zhen (2012) employed two SVM classifiers using radial basis functions (RBF) and polynomial kernels respectively. Comparing the evaluated results of these two classifiers, it is notable that polynomial kernel performs better than RBF in verification (91.96–87.5%) and RBF performs better than polynomial kernel in discrimination (93.35–91.57%). These two Support Vector Machine (SVM) classifiers were also employed in an automated classification system for Erythrocytes infected with malaria (Savkare & Narote, 2012). With combination of both classifiers, an identification accuracy rate of 96.42% was achieved. In automated identification of insects at the

order level (Wang et al., 2012), ANN and SVM were used as classification methods. Since SVM is a binary classifier and for classification of multi-class problem it has to use one over all classification for each class, SVM performs better than ANN. Comparing SVM and ANN results with semantically-related visual (SRV) attributes in an automated identification system for moths (Feng & Bhanu, 2013), SRV classifier outperforms both SVM and ANN classifiers. In the study by Kaya et al. (2015), classification was based on LBP and the accuracy rate in identification depends on variables such as neighbouring and radius values.

### 2.6.1 K-Nearest Neighbour (KNN)

One of simplest methods in classification algorithms is K Nearest Neighbour which is sorted as a lazy learning algorithm (Miller, Gregory, Aspden, Stollery, & Gilbert, 2014) but still has been used as a benchmark and workhorse classifier (Athitsos, Alon, & Sclaroff, 2005; Athitsos & Sclaroff, 2005; Peng, Heisterkamp, & Dai, 2001). In KNN, samples within a dataset cluster with other samples that contain similar properties and classes are determined according to the class of nearest neighbours (Holmes & Adams, 2002; Song, Huang, Zhou, Zha, & Giles, 2007). Based on value of nearest neighbour (k), KNN uses majority vote and appoints the labels of classes. Therefore, the performance of KNN is primarily dependent on value of k and the applied distance metric (Latourrette, 2000). Usually, KNN classifier uses Euclidian distances as the distance metric. In cases which the properties of samples are not uniformly distributed, it is difficult to predetermine the value of k but generally, larger values of k show better resistance to presented noise and distinct the boundaries between classes (Y. Song et al., 2007). Therefore, different applications of KNN require applicable value for k. In each application of KNN, k value has to be checked each time and the one with best performance will be selected. First, the training model is computed and according to

neighbours' class, the similarity of each sample with samples in testing data will be calculated (Cunningham & Delany, 2007).

The basic idea of KNN is shown in Figure 2.5 in which two classes of samples in two dimensional feature space are represented. In this figure, three nearest neighbour classifier has to decide p and q belong to which class of o or x. The decision is made by either distance weighted or majority voting.

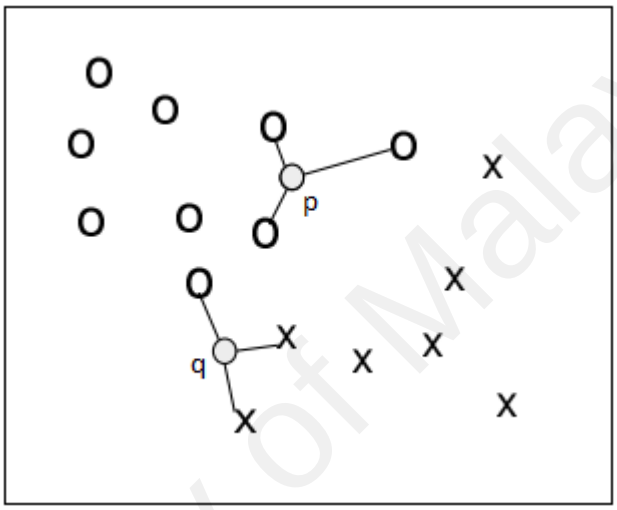


**Figure 2.5: K Nearest Neighbour classifier in two dimensional feature space. There are two classes of X and O and KNN with k value of 3 has to decide q and p belong to which class.**

A disadvantage of using majority voting classification is the tension of classes with more frequent samples to influence the prediction of unknown samples and the idea of weighting the classification according to distance of unknown point to each of nearest neighbours is a way to overcome this problem. Instead, the advantage of KNN is its robustness to noisy training data (Cunningham & Delany, 2007). This is the reason why recognition systems such as analysing received signals (Ault, Zhong, & Coyle, 2005) and offline handwritten signature identification (Soleymanpour, Rajae, & Pourreza, 2010) adopted KNN in their analysis. KNN is a good classification tool for problems with more than two classes (Yazdani, Ebrahimi, & Hoffmann, 2009).

### 2.6.2 Artificial Neural Network (ANN)

Artificial Neural Network (ANN) is a classifier which has been modelled according to human brain. ANN, like human brain, has many nerve cells that are called neurons. Each of neurons are connected to many other neurons and they create a complex network of signal transmission. The inputs from other neurons are collected by each connected neuron. In ANN, the word "perceptron" is mimicked as the neuron. The perceptron (Figure 2.6) receives different weighted inputs and encapsulate them, and the threshold determines if the combined input is exceeded to activate and send an output. Generally, the activation function that is often between 0 and 1 or −1 and 1, determines which output to send. Training network is accomplished by use of derivative of the activation function and it would be better if these derivative expresses according to of the original function value (Priddy & Keller, 2005).

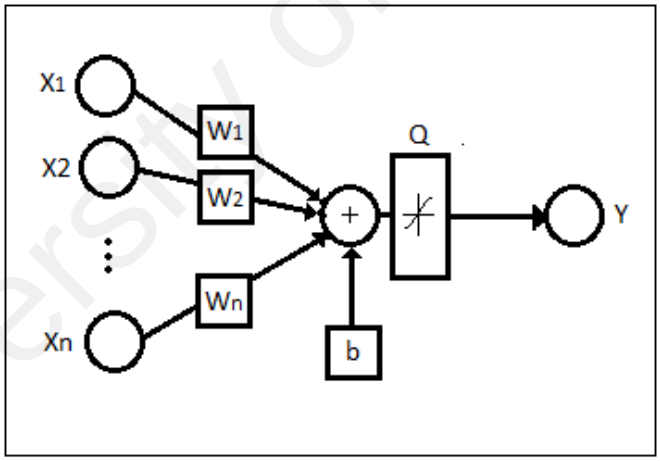


**Figure 2.6: A representation of a simple perceptron. In this illustration Y is the output, Q is the activation function, x is the value of the n connection to the perceptron, w is the weight and b represents the threshold. (Figure was adapted from Priddy & Keller (2005).**

The important aspect of classifier is learning from samples and adapting to them. In ANN, learning archives through updating the weights follow the connections in middle of layers. This can be achieved in several ways which involves initializing the weights. Then output errors by network will be calculated and by the back-propagation process will feed backward. Later the network will learn to categorize classes by updating the

weights through back-propagation. Learning from complicated samples in ANN is easily achievable since it has multilayer structure and multiple inputs can generate single output by simple model.

Artificial neural networks (ANN) have presented fulfilling results in complex classifications and proved capability in selecting proper structure and training techniques for the network (Coltelli et al., 2014; Ginoris, Amaral, Nicolau, Coelho, & Ferreira, 2007; Hernández-Serna & Jiménez-Segura, 2014; Kiranyaz et al., 2011; Culverhouse et al., 1996; Wang et al., 2012; Yang et al., 2001).

In earlier work, ANN performance has been compared with discriminant analysis (DA) and decision trees (DT) techniques ( Ginoris, Amaral, Nicolau, Coelho, & Ferreira, 2007) and ANN outperformed both DA and DT in image classification of protozoa and metazoan with overall accuracy rate of 88%. In other study ( Culverhouse et al., 1996) an automated classification system for dinoflagellates was implemented, using ANN classifiers. In this work, Radial Basis Function (RBF) (Lowe & Broomhead, 1988) and back propagation of error variant (BPN) (McClelland, Rumelhart, Group, & others, 1987) classifiers were compared with two statistical classification methods, K-Nearest Neighbour (KNN) and Quadratic Discriminant Analysis (QDA). RBF performance with accuracy rate of 83% was the best category estimation, leading labelling task in the system where BPN, QDA and KNN lag with 66%, 56% and 60% performance respectively.

# CHAPTER 3: METHODOLOGY

In this chapter, the approaches in methodology of this study are detailed as follows: monogeneans collection, monogeneans image acquisition, database of digital images, image processing, extraction of one anchor, feature extraction, feature selection, classification and evaluation. Figure 3.1 illustrates the scheme of process for development of automated identification system for monogenean.

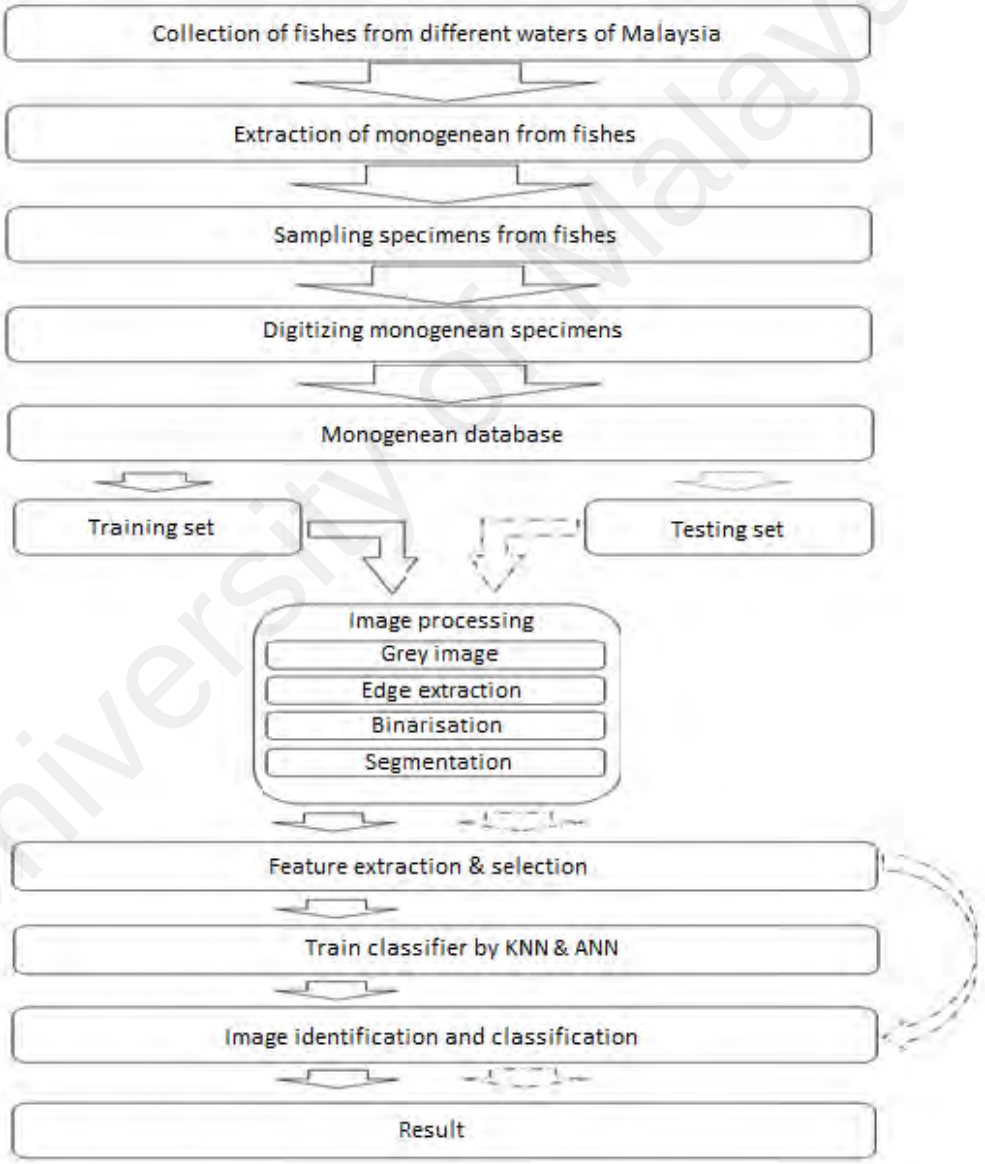


**Figure 3.1: The Scheme of process of proposed identification system for monogeneans.**

### 3.1 Monogeneans Collection

Digital images of anchors and bars of monogeneans were used in this study. Monogeneans were collected from gills of Malaysian fishes. The attached tissues were removed using fine needles and placed on clean slides with a drop of water under a coverslip. Specimens were flattened, so that the hard and soft anatomical structures of their body were exposed. To study monogeneans` specimens under phase contrast microscopy, ammonium pirate glycerine was used to clear and fix the specimens. Later, the specimens in ammonium pirate glycerine were washed, dehydrated by alcohol and firmly fixed in Canada Balsam.

Since some of the slides of monogeneans used in this study were those collected by experts since 1996 (Figure 3.2), Ammonium pirate glycerine was applied to very old specimens' slides to prepare them for image acquisition. Broken and spoiled specimens were discarded during this phase.

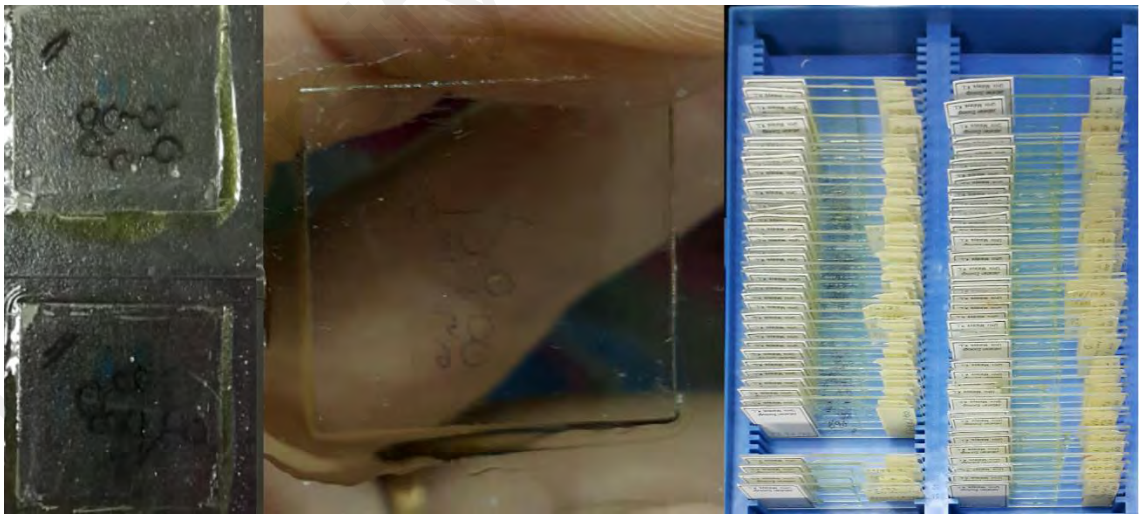

**Figure 3.2: Slides of monogenean specimens**

### 3.2 Monogeneans Image Acquisition

The specimens were investigated by phase contrast microscopy. Recognition of monogeneans is based on morphometric features of their hard parts (Lim & Gibson, 2010), Therefore, images of the hard haptoral organs such as anchors and bars of eight

species were captured using Leica digital camera DFC 320 attached to Leica DMRB microscope (Figure 3.3). The anchors and bars were observable with magnification of 40×. The images of specimens were modified using QWin Plus image analysis module by adding scale of 30 µm to the images. The resolution of images was 1044×772 pixels and saved in Tagged Image File format (TIF).

**Figure 3.3: Digitizing the monogenean specimens, using Leica digital camera DFC 320 attached to Leica DMRB microscope.**

Some of slides of monogenean samples were prepared since 1996 and accordingly, there were variety of species in stored samples. 23 available slides of species were picked and 1060 images of monogenean anchors and bars were captured and 160 images of eight species were selected based on quality of images for developing the automated identification model for monogenean.

### 3.3 Database of Digital Images

In this study, automated identification model for monogenean is connected to a database of specimens` digital images that contain different number of dominant categories. The database consisted of 160 images from eight species. There are

*Sinodiplectanotrema malayanus* (Figure 3.4), *Diplectanum jaculator* (Figure 3.11), *Trianchoratus pahangensis* (Figure 3.5), *Trianchoratus lonianchoratus* (Figure 3.8), *Trianchoratus malayensis* (Figure 3.9), *Metahaliotrema ypsilocleithru* (Figure 3.10), *Metahaliotrema mizellei* (Figure 3.6) *and Metahaliotrema similis* (Figure 3.7).

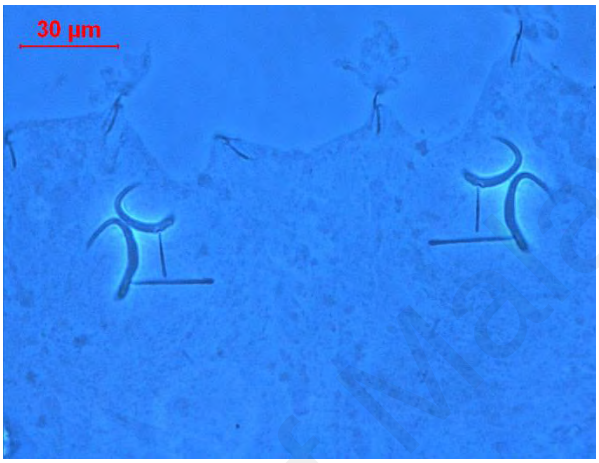


**Figure 3.4: *Sinodiplectanotrema malayanus*.**

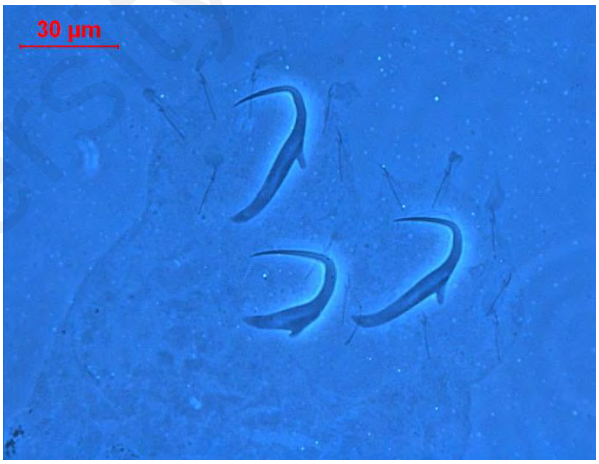


**Figure 3.5: *Trianchoratus pahangensis*.**

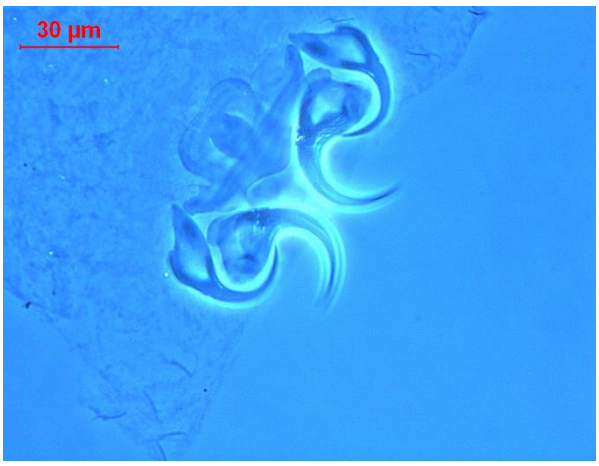


**Figure 3.6:** ***Metahaliotrema mizellei*****.**

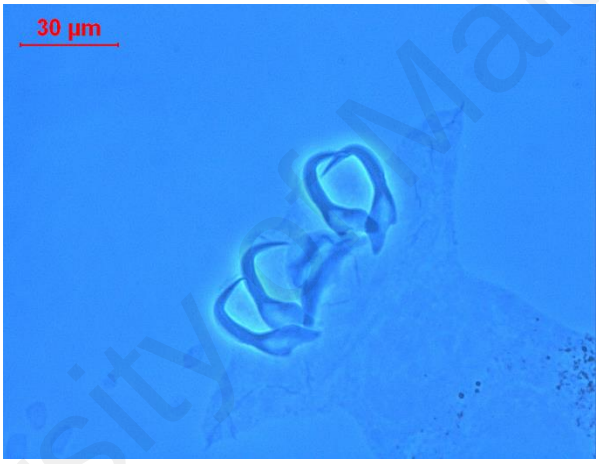


**Figure 3.7:** ***Metahaliotrema similis*****.**

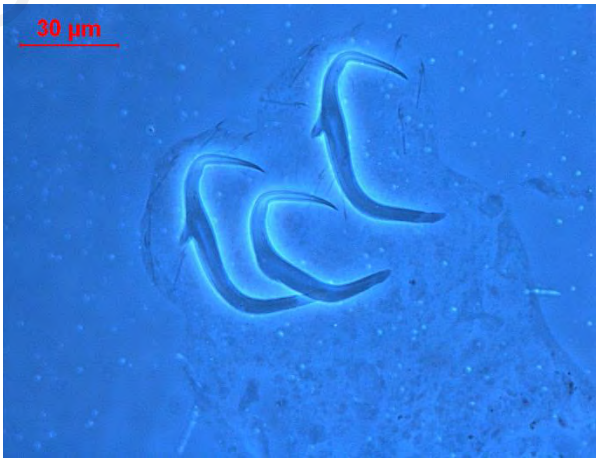


**Figure 3.8:** ***Trianchoratus lonianchoratus*****.**

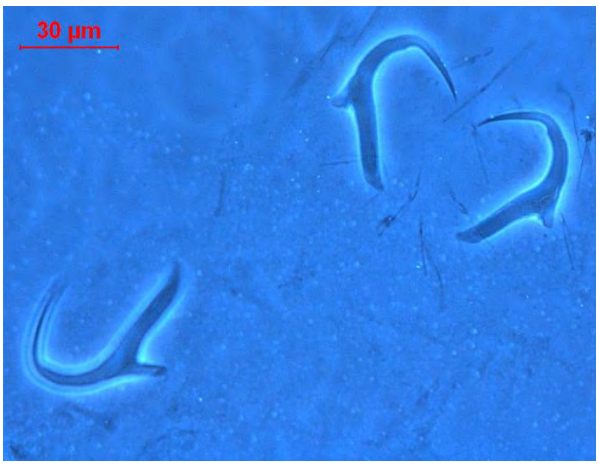


**Figure 3.9:** ***Trianchoratus malayensis*****.**

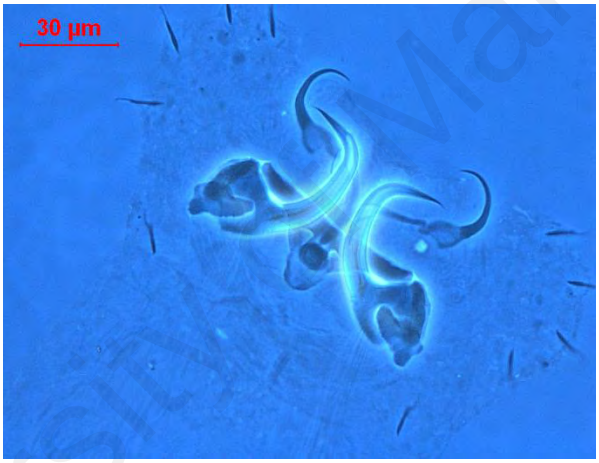


**Figure 3.10:** ***Metahaliotrema ypsilocleithru*****.**

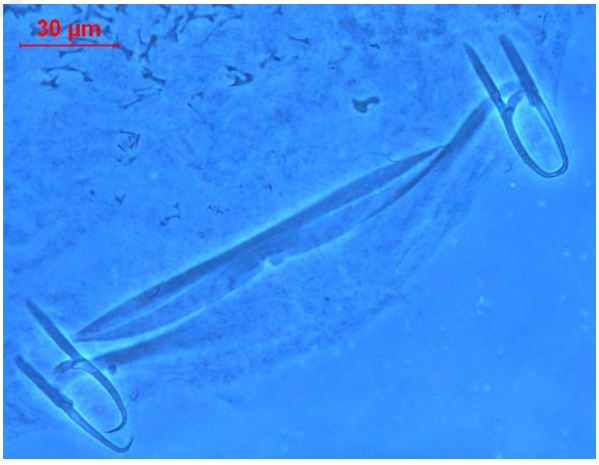


**Figure 3.11:** ***Diplectanum jaculator*****.**

According to successful experiments by Jin et al. ( 2015) and Sang-Hee (2010), 10 images of each species were used for training the KNN classifier and other 10 images as testing set (Figure 3.12). In ANN classification, according to try and errors, the best result were achieved by use of 70% of 160 images for training the system, 15% for testing and 15% for evaluation of system.

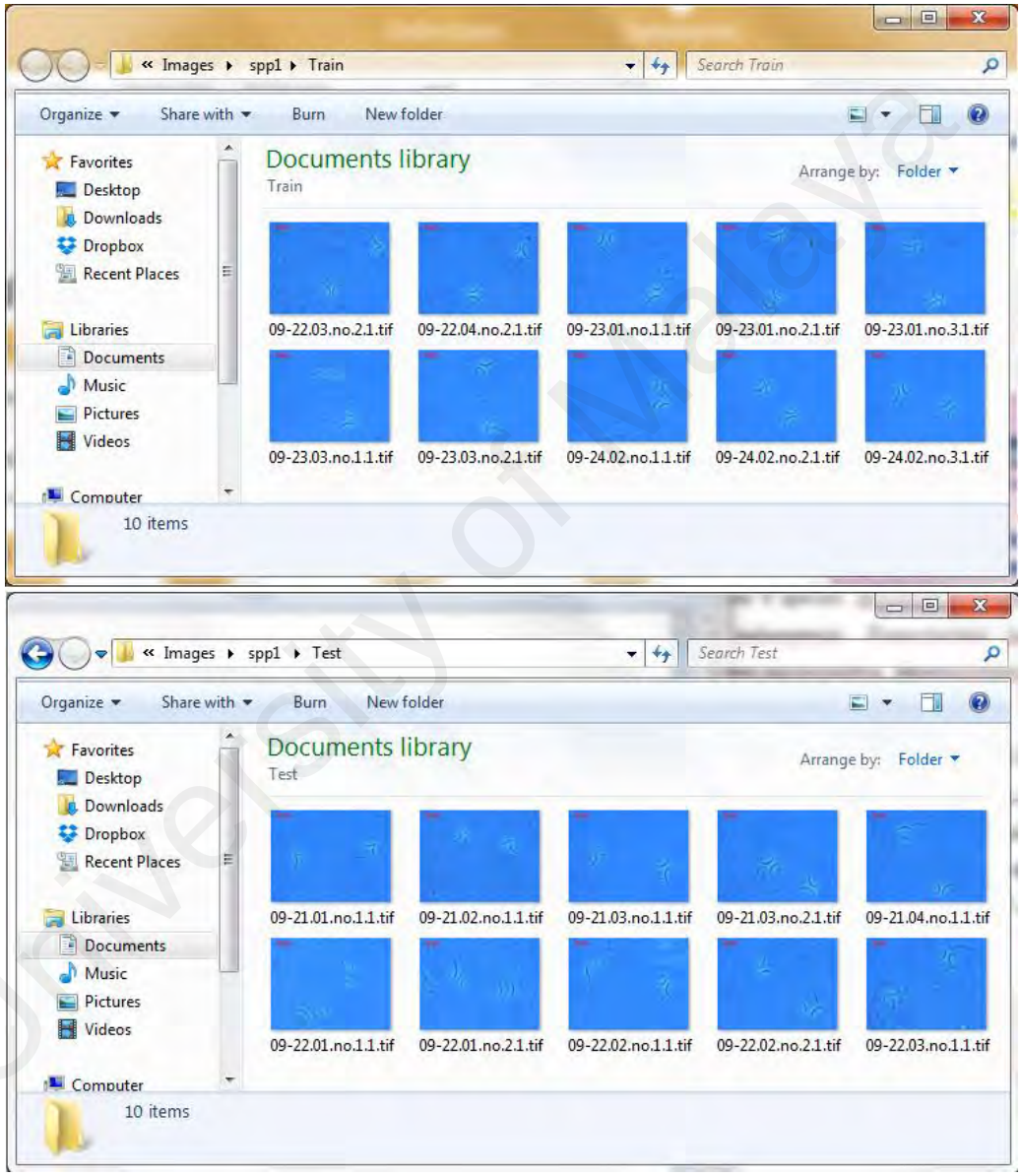


**Figure 3.12: Image database for training and testing dataset.**

### 3.4. Preliminary Identification: Four Species (First Stage)

In first stage of the study, the structure of identification system was made based on four species of monogeneans which were randomly picked from the database of eight species:

*Sinodiplectanotrema malayanus*, *Trianchoratus pahangensis*, *Metahaliotrema mizellei* and *Metahaliotrema similis*. The procedure of development of automated identification model for four species are detailed as follow: image processing, feature extraction, feature selection, classification and evaluation of automated identification model for four species of monogeneans.

### 3.4.1 Image Processing

The Image Processing Toolbox in MATLAB R2013a ("Image Processing Toolbox - MATLAB," n.d.) (Figure 3.13) was adopted for image processing, installed on Intel(R) Xeon (R) CPU E5-1620 v2 @ 3.70GHz, 16.00GB RAM, Windows 7 Professional (64-bit) to conduct this study. The image processing played an important role in this investigation and it was accomplished in two essential steps: First, image pre-processing and second, image segmentation.

```
MATLAB Version: 8.1.0.604 (R2013a)
MATLAB License Number: 262587
Operating System: Microsoft Windows 7 Version 6.1 (Build 7601: Service Pack 1)
Java Version: Java 1.6.0_17-b04 with Sun Microsystems Inc. Java HotSpot(TM) 64-Bi
--------------------------------------------------------------------------------
MATLAB                                                Version 8.1        (R2013a)
Simulink                                              Version 8.1        (R2013a)
Communications System Toolbox                         Version 5.4        (R2013a)
Computer Vision System Toolbox                        Version 5.2        (R2013a)
Control System Toolbox                                Version 9.5        (R2013a)
Curve Fitting Toolbox                                 Version 3.3.1      (R2013a)
DSP System Toolbox                                    Version 8.4        (R2013a)
Data Acquisition Toolbox                              Version 3.3        (R2013a)
Fuzzy Logic Toolbox                                   Version 2.2.17     (R2013a)
Global Optimization Toolbox                           Version 3.2.3      (R2013a)
Image Acquisition Toolbox                             Version 4.5        (R2013a)
Image Processing Toolbox                              Version 8.2        (R2013a)
MATLAB Builder NE                                     Version 4.1.3      (R2013a)
MATLAB Coder                                          Version 2.4        (R2013a)
MATLAB Compiler                                       Version 4.18.1     (R2013a)
Mapping Toolbox                                       Version 3.7        (R2013a)
Neural Network Toolbox                                Version 8.0.1      (R2013a)
Optimization Toolbox                                  Version 6.3        (R2013a)
Parallel Computing Toolbox                            Version 6.2        (R2013a)
Signal Processing Toolbox                             Version 6.19       (R2013a)
SimPowerSystems                                       Version 5.8        (R2013a)
Simscape                                              Version 3.9        (R2013a)
Simulink 3D Animation                                 Version 6.3        (R2013a)
Simulink Coder                                        Version 8.4        (R2013a)
Simulink Control Design                               Version 3.7        (R2013a)
Simulink Design Optimization                          Version 2.3        (R2013a)
Statistics Toolbox                                    Version 8.2        (R2013a)
System Identification Toolbox                         Version 8.2        (R2013a)
Wavelet Toolbox                                       Version 4.11       (R2013a)
xPC Target                                            Version 5.4        (R2013a)
```

**Figure 3.13: List of installed toolboxes in MATLAB.**

#### 3.4.1.1 Image Pre-processing

Background feature minimization is an important pre-processing step in monogeneans classification. Otherwise, soft part features of monogeneans could mix with those from hard parts and the texture analysis will yield unreliable results. The image pre-processing follows as:

(i) Images were converted to intensity images.

(ii) Filtering intensity images with the average correlation kernel of size 20 x 20.

(iii) Detecting the edge of the anchors and bars of monogeneans.

After detecting the edges in the images, image segmentation was performed where bars and anchors were identified and segmented from unwanted particles in the images:

### 3.4.1.2 Image Segmentation

After detecting the edges in the images, image segmentation was done where bars and anchors were identified and segmented from unwanted particles in the images (Figure 3.14):

1) The images were converted to binary images with threshold of zero. After creating an average filter, the image was deducted from filter. The result is an intensity image which contains negative and positive values. Therefore, pixels, greater than 0 will turn to 1(white) and other pixels will turn to 0 (black).

2) Small particles (<1000 pixels) were excluded to ensure only the bars and anchors are segmented for feature extraction.

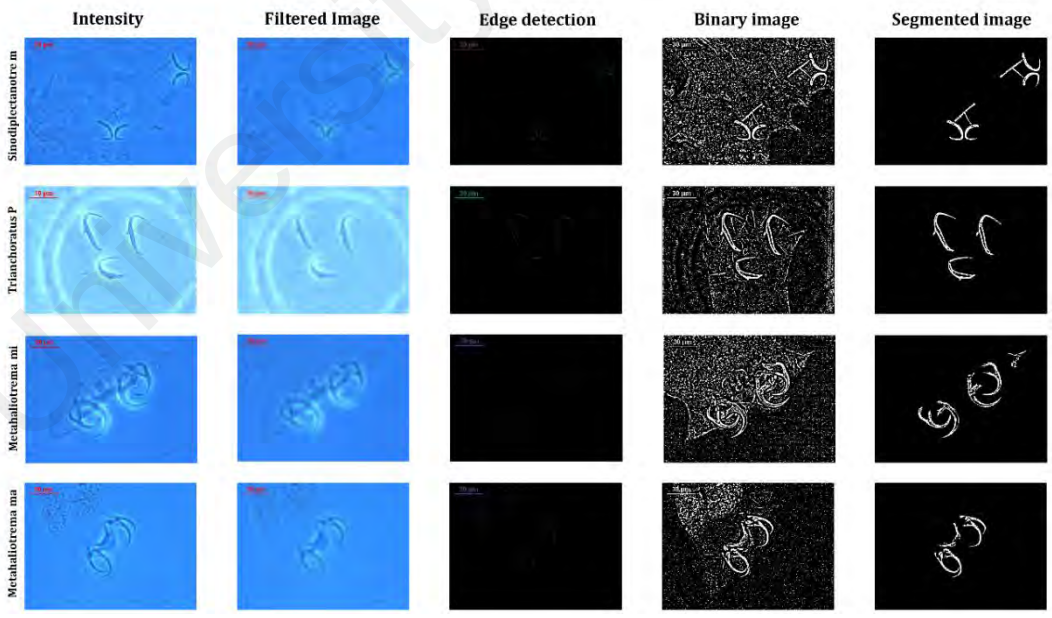


**Figure 3.14: Process in image pre-processing, edge detection and image segmentation steps for four species of *Sinodiplectanotrema malayanus*, *Trianchoratus pahangensis*, *Metahaliotrema mizellei* and *Metahaliotrema similis*.**

### 3.4.2 Feature Extraction

Features were extracted from the shape descriptors represented by the binary images of the bars and anchors, using appropriate functions in MATLAB. The features vector with 10 elements were extracted form the following shape parameters (Table 3.1): Euler number, perimeter, area, area density, perimeter density, centre of bounding box, length of bounding box, width of bounding box and orientation of bounding box.

**Table 3.1: Description of shape parameters, used for feature extraction in four species ( Stage 1).**

| Shape Parameters | Description |
|---|---|
| Area | Actual number of pixels in the region of particular object. |
| Area density | The mass of a substance covering a unit of area. |
| Perimeter | Distance around the boundary of the region. |
| Perimeter density | The measure of length of the perimeter of a set in free boundary. |
| Length of bounding box | Length of smallest rectangle containing the region. |
| Width of bounding box | Width of smallest rectangle containing the region. |
| Center of bounding box | Center point of smallest rectangle containing the region. |
| Orientation of bounding box | The angle between the x-axis and the major axis of the ellipse that has the same second-moments as the smallest rectangle containing the region. |
| Euler number | The number of objects in the region minus the number of holes in those objects. |

### 3.4.3 Feature Selection

To increase the performance of classifiers and decrease the number of unnecessary features, Linear Discriminant Analysis (LDA) was applied for feature selection. Practically, LDA as a feature dimensionality reduction technique would be pre-step for a typical classification task. In this study, for calculation of LDA, 10 dimensional mean vectors for four classes` dataset was calculated. After computing in-between class and within-class scatter matrix, the eigenvectors and corresponding eigenvalues were

calculated. Subsequently, eigenvectors are sorted in line with increasing growth and 3 eigenvectors with largest eigenvalues were picked. Finally, 4×3 dimension eigenvector was adopted to transform feature space to new subspace.

### 3.4.4 Classification

In this study, two classifiers were used to classify the images into the right species.

#### 3.4.4.1 K-Nearest Neighbour (KNN) Training

We applied K-nearest neighbour (KNN) classifier to the same training and test datasets. K-NN, as a non-parametric classifier, identifies the test sample by a majority vote of its neighbours which are assigned to the class that is most common among its K nearest neighbours. The KNN parameter was set to 1 in this study. The three selected features obtained from previous stage were used as input to KNN classifier. Four species of monogeneans were used and the vectors of image labels were prepared according to their features. KNN was used in this study because our dataset was from real world while practical and theoretical data do not follow the same assumptions in KNN. Therefore, no hypothesis was made on the fundamental data distribution. The trained model from KNN classifier was constructed using 40 images and tested with 40 images of monogeneans with 1 nearest neighbour.

The step by step process in KNN classification is as follow:

(i) Compute the distribution of feature values in each class of training dataset.

(ii) Compute Euclidean distance between training and testing feature vectors.

(iii) Sort the Euclidean distance output into ascending order.

(iv) Obtain the first nearest neighbour classes for each of testing feature vectors.

(v) Obtain the hypothesis of the class for each sample by weighted majority voting.

### 3.4.4.2 Artificial Neural Network (ANN) Training

The other pattern recognition tool, used in this study was Artificial Neural Network (ANN) to classify sample specimens to four classes. The ANN classifier structure was a two layer feed-forward network with ten sigmoid hidden nodes and four output neurons and scaled conjugate gradient back propagation was used to train the network (Figure 3.15).

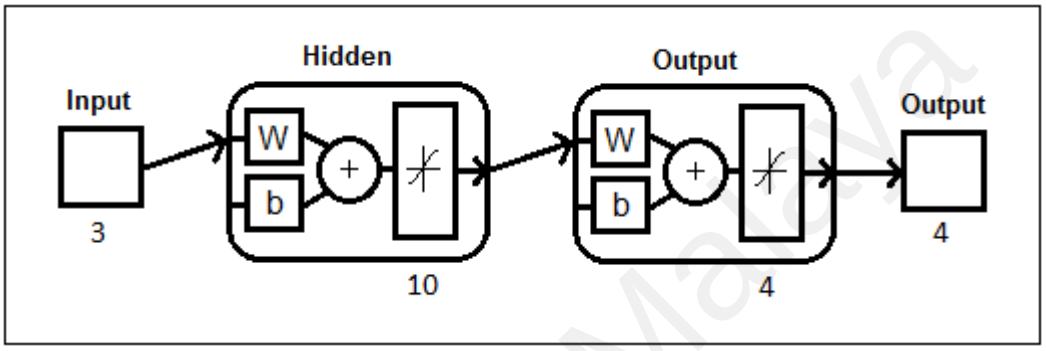


**Figure 3.15: Neural Network with 10 sigmoid hidden nodes and four output neurons.**

Opening the neural network graphical user interface (GUI) in MATLAB was by keying in 'nnstart' function. For ANN classification, pattern recognition tool was adopted and the feature vector as input and target vector were assigned. The whole data (80 images) was divided to three training (56 samples, 70%), testing (12 samples, 15%) and validation (12 samples, 15%) dataset. Training dataset was used for training ANN, testing dataset for performance measurement of the network and validation set to measure generalization of network and terminates training before overfitting.

For evaluating the trained network the confusion matrices and Mean Square Error (MSE) were used. Increasing the value of MSE in samples of validation set imply that the improvement in network generalisation has been stopped and this causes training break. The network was trained several times to obtain best performing train network. Since MSE is the average squared difference between outputs and targets, the lowest value means better performance of train network.

### 3.4.5 Evaluation

The evaluation of the system with both classification techniques were accomplished by correct classification accuracy rate of testing data set. A total number of 40 images from image database were assigned to test the system with KNN classification and 12 images from image database were assigned to evaluate and test the system with ANN classification. Since the sample size was small, Leave-One-Out (LOO) cross validation was used to assess how the results of the system generalize to an independent data set. The result for the evaluation of KNN, ANN and LOO cross validation is recorded in confusion matrices presented in Chapter Four.

## 3.5. Extended Identification on Eight Species (Second Stage)

In second stage of the study, the structure of identification system was extended based on four species of monogeneans to eight species from the database: *Sinodiplectanotrema malayanus, Diplectanum jaculator, Trianchoratus pahangensis, Trianchoratus lonianchoratus, Trianchoratus malayensis, Metahaliotrema ypsilocleithru, Metahaliotrema mizellei* and *Metahaliotrema similis*. The procedure of development of automated identification model for eight species are explained as follow: image processing, extraction of one anchor, feature extraction, feature selection, classification and evaluation of automated identification system for eight species of monogeneans.

### 3.5.1 Image Processing

The Image Processing Toolbox in MATLAB R2013a ("Image Processing Toolbox - MATLAB," n.d.) was adopted for image processing, installed on Intel(R) Xeon (R) CPU E5-1620 v2 @ 3.70GHz, 16.00GB RAM, Windows 7 Professional (64-bit) to conduct this study. The image processing played an important role in this investigation

and it was accomplished in two essential steps: first, image pre-processing and second, image segmentation.

### 3.5.1.1 Image Pre-processing

One of biggest challenges of monogenean specimen images was complexity in terms of messy background and overlapping of anchors and bars. Although many efforts were made to acquire clear images but still some overlapping and clutters were unavoidable (Figure 3.16).

Hence, pre-processing stage played an important role as long as redundant information are omitted and reliable features are highlighted for next process in feature extraction. Pre-processing started with converting three dimensional colour image (RGB images) to two dimensional intensity images using MATLAB function: 'mat2gray'. For filtering the intensity images, average filtering mode as a mask with a 20-by-20 kernel was used to conceal the noise produced by clutters and debris under slides.

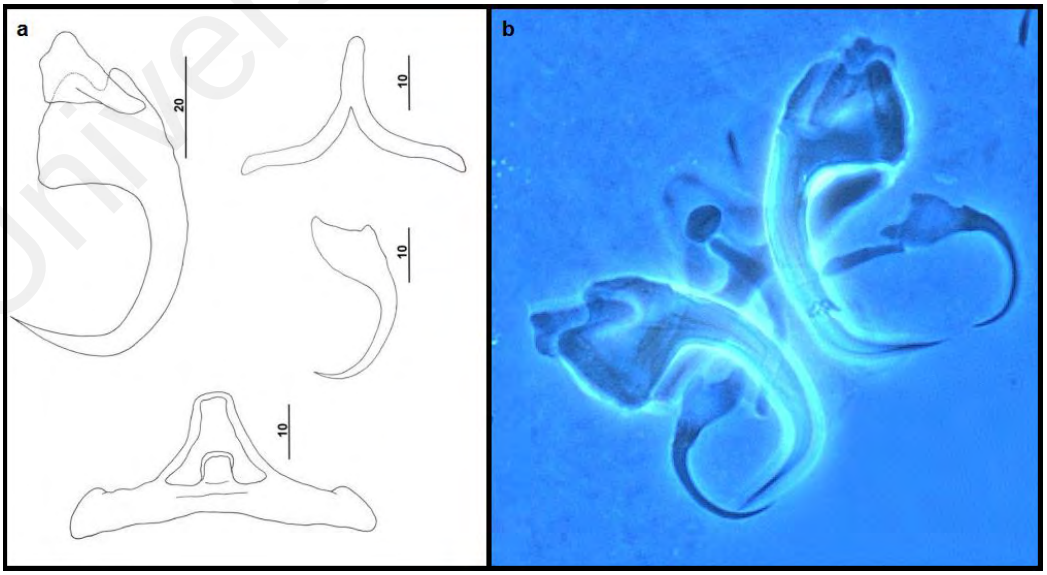


**Figure 3.16: The illustration of anchors and bars of *Metahaliotrema ypsilocleithrum*. a) The illustration of dorsal and ventral anchors and bars. b) The microscopic image of anchors and bars and their overlapping.**

#### 3.5.1.2 Image Segmentation

In order to identify the edge of anchors and bars, the intensity images were deducted from the filtered images (Figure 3.17). The images containing edges of anchors and bars were then converted to binary images. Then, they were binarized with threshold of zero. Then the borders were cleared and objects smaller than 1000 pixels were removed (Figure 3.18). The coordinates of contour pixels for species` anchors were also calculated. Therefore, features were extracted once from all anchors and bars as a united object and the other time only an anchors.

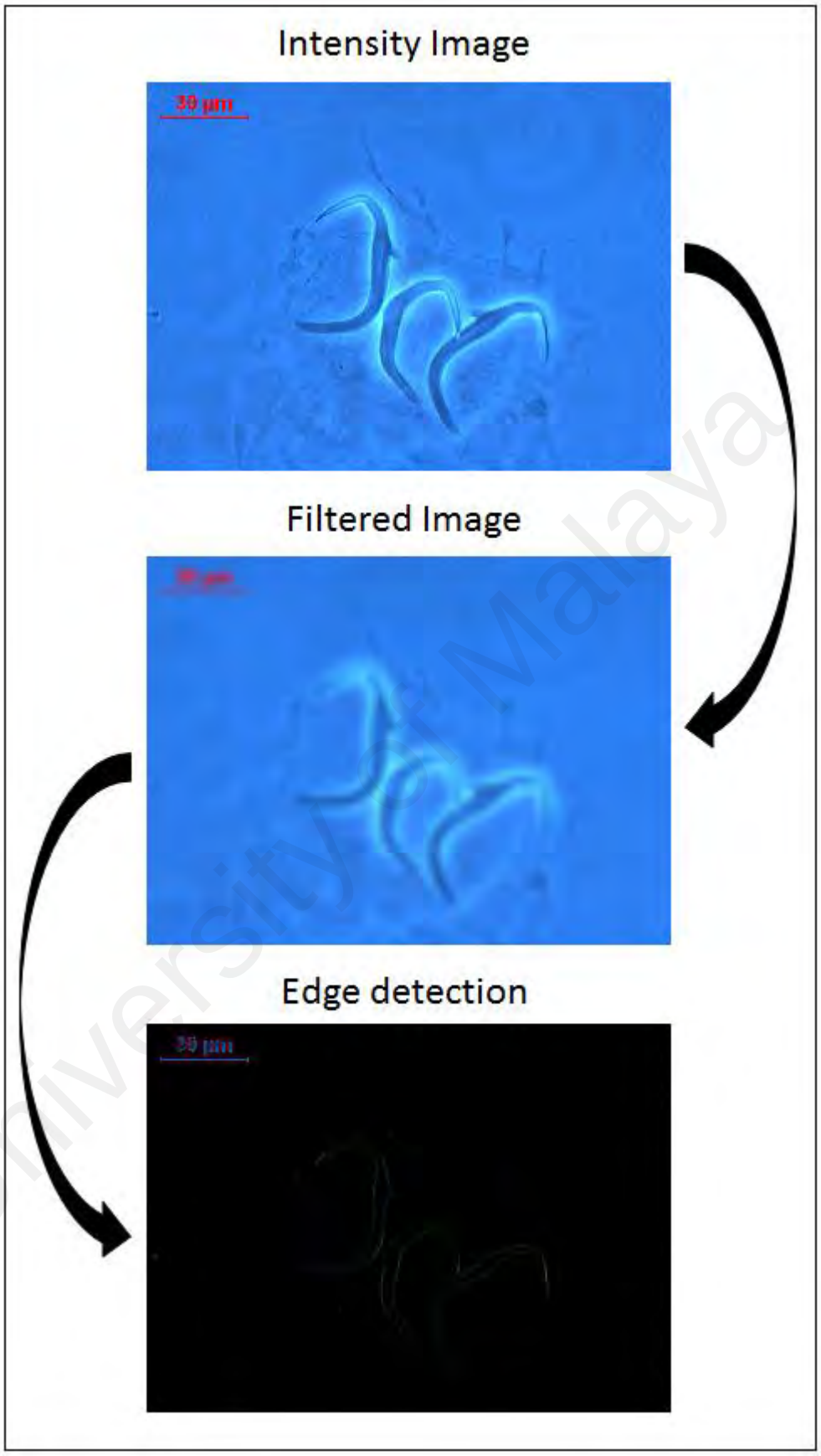


**Figure 3.17: The process of detecting edges from intensity image.**

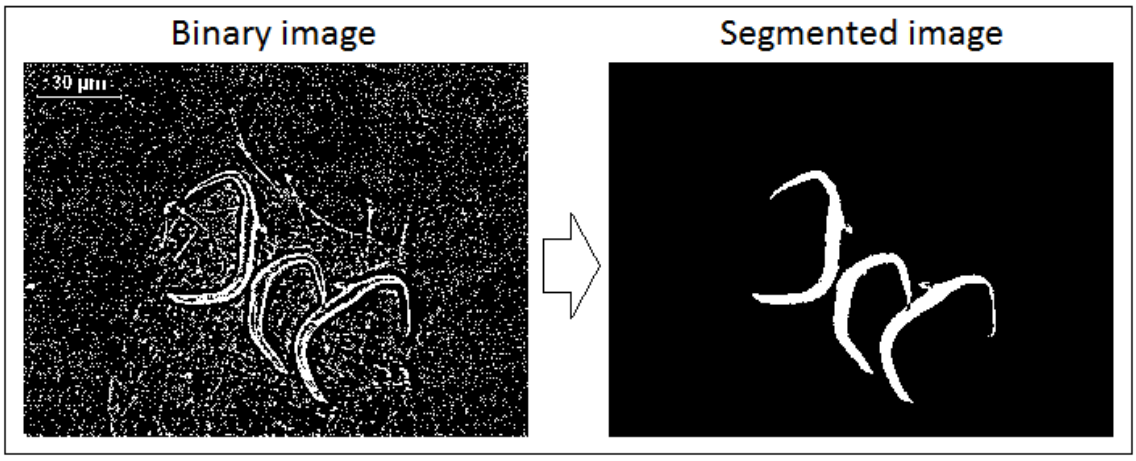


**Figure 3.18: The process of converting binary image to segmented image.**

### 3.5.2 Extraction of One Anchor

The output of image processing stage was segmented images of segmented anchors and bars of monogenean. As a result of dorsal and ventral organ`s overlapping, anchors and bars in some were segmented as one unit of object and that means the computer counted all haptoral organs as one organ. To overcome the misconception of segmented images, one anchor was extracted in each image (Figure 3.19). Therefore, feature extraction was accomplished by extracting features from all anchors and bars as a unit object and also from one anchor.

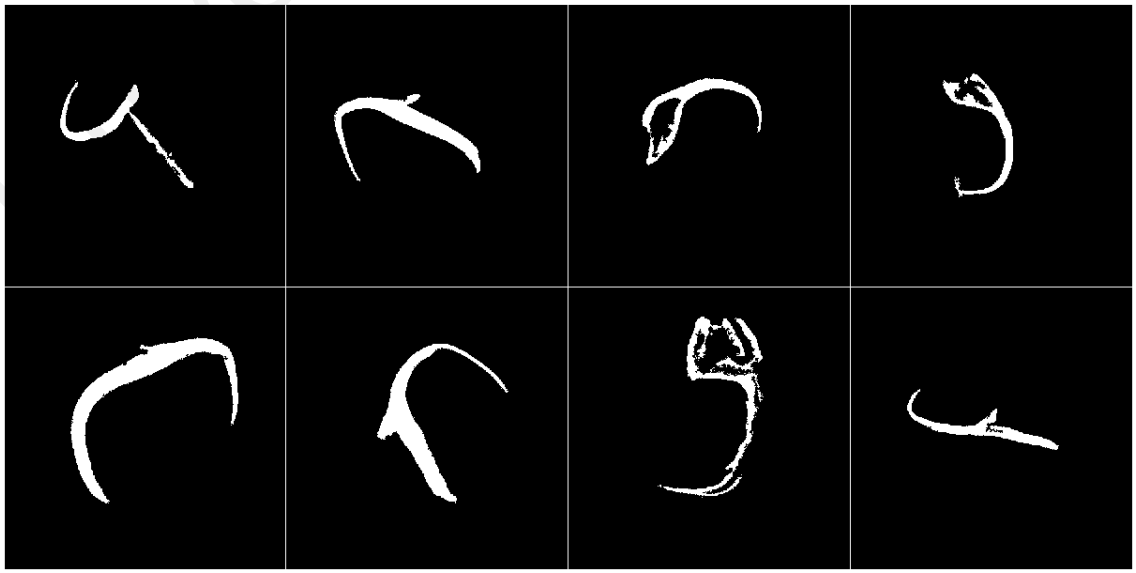

**Figure 3.19: Extraction of one anchor of each species.**

### 3.5.3 Feature Extraction

Binary images were used two times for feature extraction. Once using all anchors and bars as a united object and another time, by calculating coordinates of one anchor and then extracting the features from that anchor. Features were extracted by shape representation techniques from shape descriptors of binary images in MATLAB. According to parameters such as length of bounding box, width of bounding box, centre of bounding box, orientation of bounding box, perimeter, perimeter density, area, area density, Euler number, entropy and major axis length (Table 3.2), a feature vector with 24 elements was extracted.

**Table 3.2: Description of shapes parameters, used for feature extraction in eight species.**

| Shape Parameters | Description |
|---|---|
| Area | Actual number of pixels in the region of particular object. |
| Area density | The mass of a substance covering a unit of area. |
| Perimeter | Distance around the boundary of the region. |
| Perimeter density | The measure of length of the perimeter of a set in free boundary. |
| Length of bounding box | Length of smallest rectangle containing the region. |
| Width of bounding box | Width of smallest rectangle containing the region. |
| Centre of bounding box | Centre point of smallest rectangle containing the region. |
| Orientation of bounding box | The angle between the x-axis and the major axis of the ellipse that has the same second-moments as the smallest rectangle containing the region. |
| Euler number | The number of objects in the region minus the number of holes in those objects. |
| Entropy | The measure of randomness that can be used to characterize the texture of the region. |
| Major axis length | The length (in pixels) of the minor axis of the ellipse that has the same normalized second central moments as the region. |

### 3.5.4 Feature Selection

Feature selection is a technique for reducing the dimensions of feature vector. In this study, the informative and independent features were selected using linear discriminant analysis (LDA) feature selection method (Cai et al., 2008). The goal of LDA is separation between multiple classes by maximizing the component axes. Therefore, besides projecting a feature space to smaller subspace, the class-discriminatory information was also maintained. In this approach, first, 24 dimensional mean vectors for eight classes` dataset was calculated. After computing in-between class and within-class scatter matrix, the eigenvectors and corresponding eigenvalues were calculated. Subsequently, eigenvectors were sorted in line with increasing growth and seven eigenvectors with largest eigenvalues were picked. Finally, the 8×7 dimension eigenvector was adopted to transform feature space to new subspace.

### 3.5.5 Classification

Two classifiers, K-Nearest Neighbour (KNN) and Artificial Neural Network (ANN) were used to classify the images into species.

#### 3.5.5.1 K-Nearest Neighbour (KNN) Training

. In this study, from all 160 images captured from eight different species, trained model from KNN classifier was constructed using 80 images and tested with 80 images of monogeneans with 9 nearest neighbours. The step by step process in KNN classification is as follow:

(vi) Compute the distribution of feature values in each class of training dataset.

(vii) Compute Euclidean distance between training and testing feature vectors.

(viii) Sort the Euclidean distance output into ascending order.

(ix) Obtain the 9 nearest neighbour's classes for each of testing feature vectors.

(x) Obtain the hypothesis of the class for each sample by weighted majority voting.

#### 3.5.5.2 Artificial Neural Network (ANN) Training

The other pattern recognition tool, used in this study to classify sample specimens to eight classes was Artificial Neural Network (ANN). The ANN classifier structure was a

two layer feed-forward network with ten sigmoid hidden nodes and eight output neurons and scaled conjugate gradient back propagation was used to train the network (Figure 3.20).

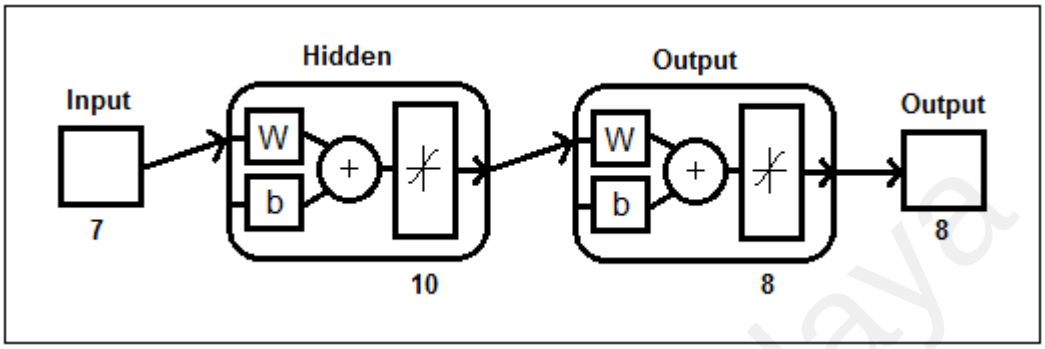


**Figure 3.20: Neural Network with 10 sigmoid hidden nodes and four output neurons.**

The whole data (160 images) was divided to training (112 samples, 70%), testing (24 samples, 15%) and validation (24 samples, 15%) dataset. Training dataset was used for training ANN, testing dataset for performance measurement of the network and validation set to measure generalization of network and terminates training before overfitting.

For evaluating the trained network confusion matrices and Mean Square Error (MSE) were used. Increasing the value of MSE in samples of validation set imply that the improvement in network generalisation has been stopped and this causes training break. The network was trained several times to obtain best performing train network. Since MSE is the average squared difference between outputs and targets, the lowest value means better performance of train network.

### 3.5.6 Evaluation

The evaluation of the system with both classification techniques were accomplished by correct classification accuracy rate of testing data set. A total of 80 images from image database were assigned to test the model with KNN classification and 24 images

to all images were assigned to evaluate and test the model with ANN classification. Also, since the sample size was small, Leave-One-Out (LOO) cross validation was used to assess how the results of the system generalize to an independent data set. The result for the evaluation of KNN, ANN and LOO cross validation is recorded in confusion matrices which are presented in chapter four.

# CHAPTER 4: RESULTS

In this chapter, the results of implementation and empirical considerations are demonstrated. The various approaches, carried out in this study are addressed in detail. First, the results of feature selection, K-Nearest Neighbour and Artificial Neural Network classification and evaluation of classification for four species (first stage) are elaborated. In feature selection, the feature vector with 10 elements was transformed to feature vector with 3 elements. The adoption of selected features had increased the accuracy rate of classification of four monogenean species.

Subsequently, the results of feature selection, KNN and ANN classification and evaluation of classification for eight species (second stage) are explained in detail. In this stage, the model feature extraction was extended to extraction a feature vector with 24 elements which was then transformed to feature vector with seven elements using the LDA technique. The new feature vector employed in KNN and ANN classifications for classifying eight species of monogeneans.

These two main stages follow the original model for automated identification system for monogenean.

## 4.1 Preliminary Identification Results (First Stage)

In this section, the experimental results for preliminary model of automated identification model for four species of *Sinodiplectanotrema malayanus*, *Trianchoratus pahangensis*, *Metahaliotrema mizellei* and *Metahaliotrema similis* are explained in detail. Here, the results reveal the accuracy of proposed model for classification of four species of monogenean in feature selection, classification and evaluation of automated identification system.

### 4.1.1 Feature Selection

A feature vector with 10 elements was extracted from anchors and bars of four species. The features were extracted from shape parameters such as Euler number, perimeter, area, area density, perimeter density, centre of bounding box, length of bounding box, width of bounding box and orientation of bounding box. After LDA feature selection, the feature vector was transformed to feature vector with 3 elements. The 3D scatter plots in Figure 4.1 (a), (b) and (c) show the clustering of four species samples (different colours represent different species) based on features extracted, before LDA feature selection. From the clusters, it is notable that the species of *Sinodiplectanotrema malayanus* and *Trianchoratus pahangensis* and *Metahaliotrema similis* are not well grouped and samples from *Trianchoratus pahangensis* tend to mingle with *Metahaliotrema similis* before feature selection. In Figure 4.1 (d), the clusters of features resulted from LDA feature selection of samples for four species are shown and it is illustrious that the samples are well clustered according to the species.

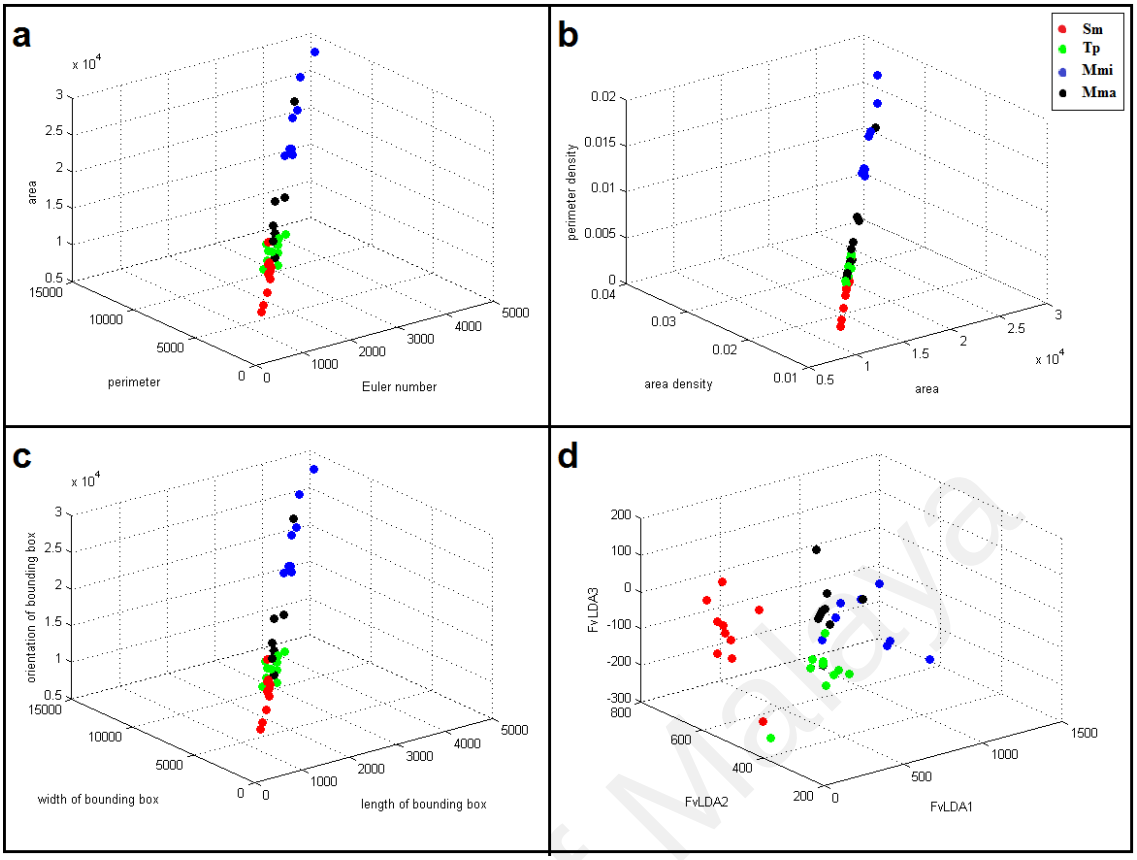


**Figure 4.1: 3D scatter plot with different features. (a) scatter plot with combination of three features which are Euler number, perimeter, area (b) scatter plot with combination of three features which are area, area density, perimeter density (c) scatter plot with combination of three features which are length of bounding box, width of bounding box and orientation of bounding box (d) scatter plot with combination of LDA transformed features: FvLDA1, FvLDA2 and FvLDA3. The data were classified into four species: *Sinodiplectanotrema malayanus* (Smm), *Trianchoratus pahangensis* (Tp), *Metahaliotrema mizellei* (Mmi) and *Metahaliotrema similis* (Mma).**

To study the relationship between the four species according to the features extracted from shape parameters and those transformed by LDA features selection technique, 2D scatter plots were graphed for each selected feature. In 2D scatter plot in Figure 4.2, well separation between species by use of only first element of selected features is shown. *Sinodiplectanotrema malayanus* (represented by red colour dots) is completely separated from *Metahaliotrema mizellei* (represented by blue colour dots) and *Metahaliotrema similis* (represented by black colour dots). Also, samples from *Trianchoratus pahangensis* (represented by green colour dots) mingle with *Sinodiplectanotrema malayanus* and *Metahaliotrema similis*. In Figure 4.3, samples

from *Metahaliotrema mizellei* (represented by blue colour dots) mingle with *Metahaliotrema similis* (represented by black colour dots). Since these two species are from same genera, it is expected that the features resembles. Although third element of the selected feature vector in Figure 4.4 shows well separation of samples between both species of *Metahaliotrema*, still samples from *Sinodiplectanotrema malayanus* (represented by red colour dots) and *Trianchoratus pahangensis* (represented by green colour dots) mingle between all species. However, the combination of three elements for selected feature vectors, achieved acceptable clustering for four species (Figure 4.5).

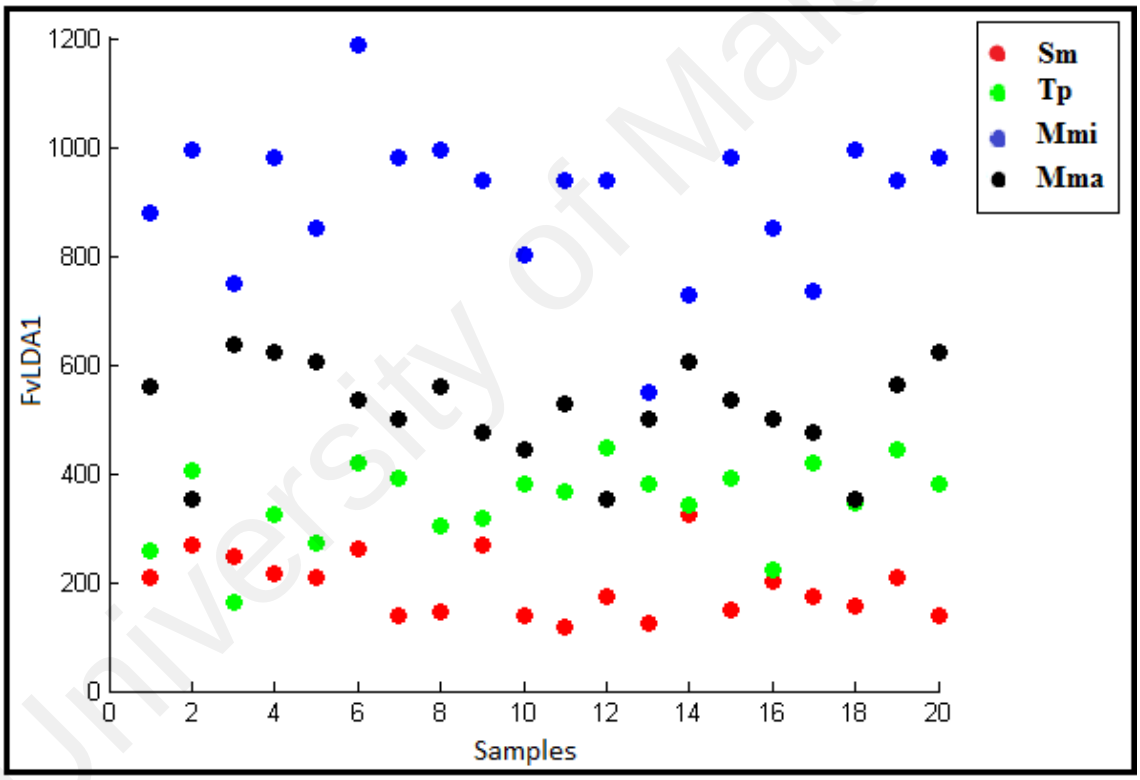


**Figure 4.2: 2D scatter plot of first element of transformed feature vector by LDA for samples of *Sinodiplectanotrema malayanus*, *Trianchoratus pahangensis*, *Metahaliotrema mizellei* and *Metahaliotrema similis*.**

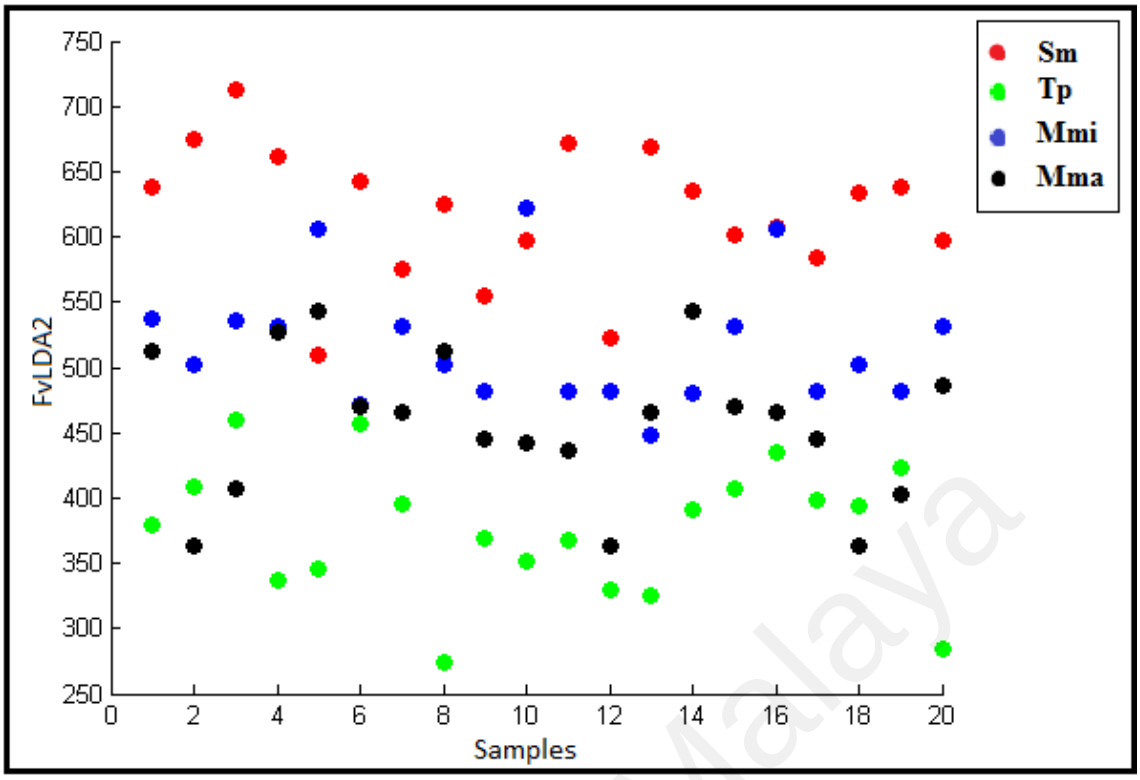


**Figure 4.3: 2D scatter plot of second element of transformed feature vector by LDA for samples of *Sinodiplectanotrema malayanus*, *Trianchoratus pahangensis*, *Metahaliotrema mizellei* and *Metahaliotrema similis*.**

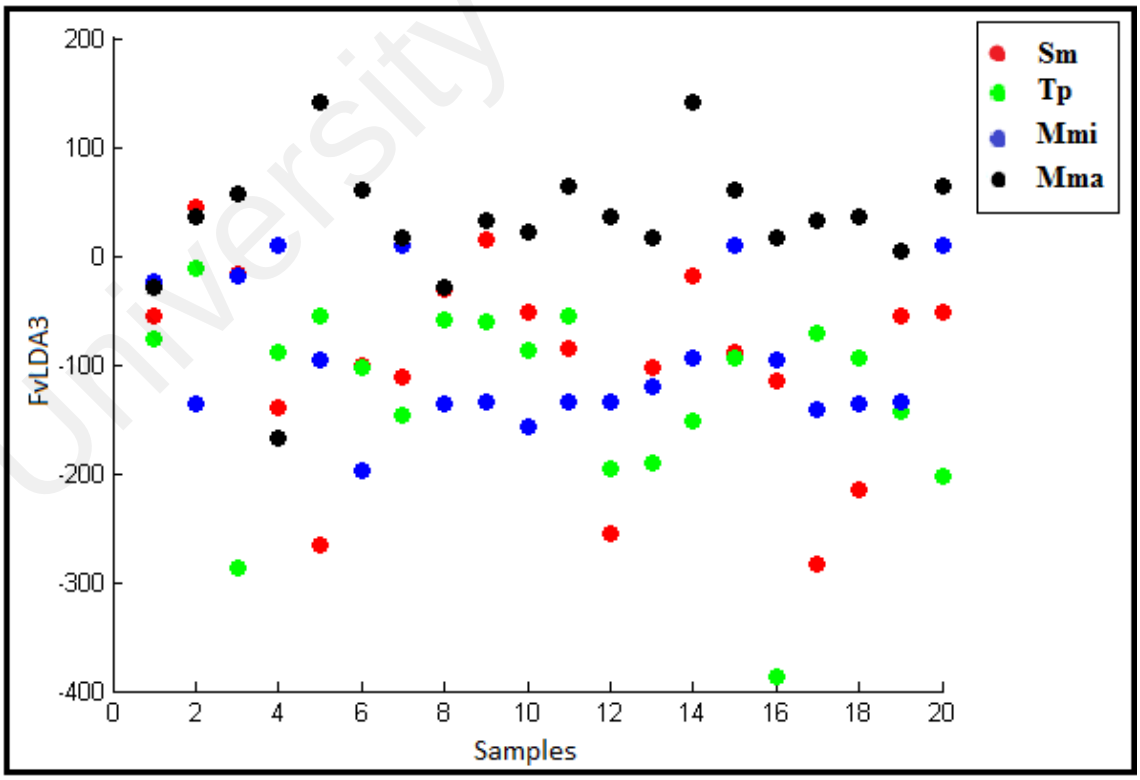


**Figure 4.4: 2D scatter plot of third element of transformed feature vector by LDA for samples of *Sinodiplectanotrema malayanus*, *Trianchoratus pahangensis*, *Metahaliotrema mizellei* and *Metahaliotrema similis*.**

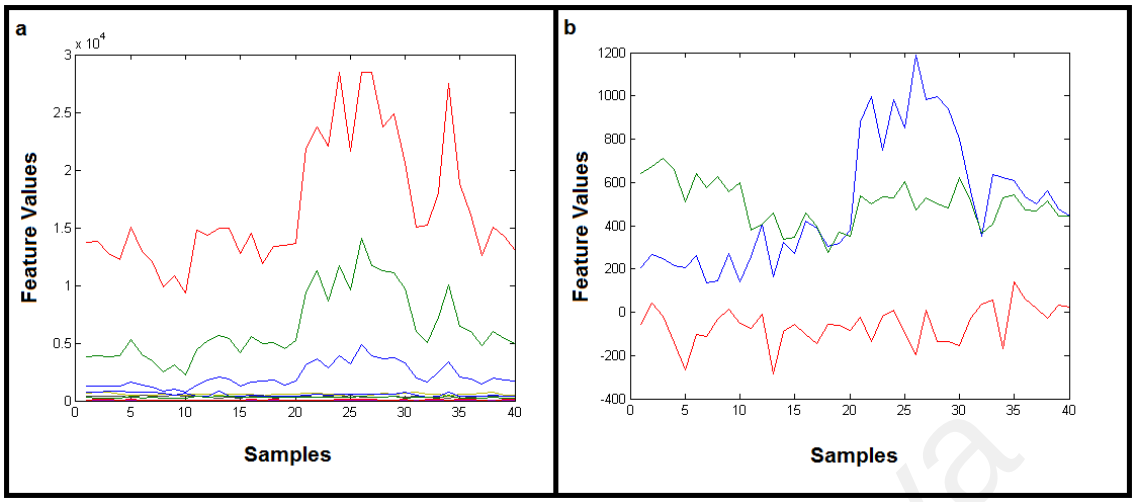


**Figure 4.5: The distinction of feature values before and after LDA feature selection. a) Illustration of discrimination between 10 feature vector elements of *Sinodiplectanotrema malayanus*, *Trianchoratus pahangensis*, *Metahaliotrema mizellei* and *Metahaliotrema similis.* b) Illustration of discrimination between 3 feature vector elements selected by LDA for four species of *Sinodiplectanotrema malayanus, Trianchoratus pahangensis*, *Metahaliotrema mizellei* and *Metahaliotrema similis*.**

### 4.1.2 Classification

The results from feature selection in previous stage were invoked by KNN and ANN. The details of classification approaches and results in both KNN and ANN are indicated in the following sections:

#### 4.1.2.1 K-Nearest Neighbour (KNN)

KNN does not make any hypothesis on the underlying data distribution. This is useful in this study's case since the data is from real world. Generally practical data does not follow the theoretical assumptions like for example Gaussian mixtures or linearly separable made. Non parametric algorithms like KNN come to the rescue here. The trained model was constructed using 10 images of each monogenean species and the model was tested by 10 images of each monogenean species in testing dataset. After 25 iterations of KNN classification with different k values, as reported by the majority voting, the best result was achieved with k=1 nearest neighbour (Figure 4.6). According

to the confusion matrix (Table 4.1), the overall classification score for four species with KNN classification is 95%.

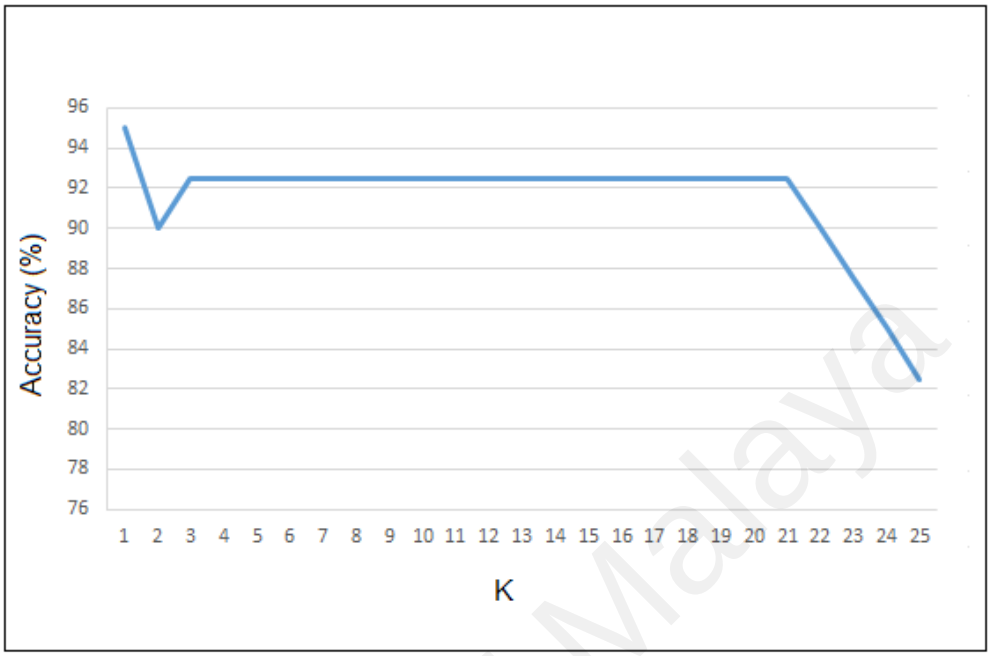


**Figure 4.6: Illustration of k value in 25 iterationss of KNN classification for four species. The best result made by k=1.**

**Table 4.1: Confusion matrix of KNN classification for four species of: Sinodiplectanotrema malayanus (Smm), Trianchoratus pahangensis (Tp), Metahaliotrema mizellei (Mmi) and Metahaliotrema similis (Mma).**

| | | Results | | | Accuracy % |
|---|---|---|---|---|---|
| Species | Smm | Tp | Mmi | Mma | |
| Smm | 10 | 0 | 0 | 0 | 100 |
| Tp | 0 | 10 | 0 | 0 | 100 |
| Mmi | 0 | 0 | 8 | 2 | 80 |
| Mma | 0 | 0 | 0 | 10 | 100 |
| Overall | | | | | 95 |

### 4.1.2.2 Artificial Neural Network (ANN)

The architecture of ANN classification was a two layer feed-forward network with ten sigmoid hidden nodes and four output neurons and scaled conjugate gradient back propagation was used to train the network. The network was trained by 56 samples and

the trained model was tested by 12 samples and validated by 12 samples. Mean Square Error (MSE) was used for evaluating the trained network and incrimination in MSE imply that the improvement in network generalisation has been stopped and this causes training break. In this experiment, the MSE value for training, testing and validation set is reported in Table 4.2, MSE is the average squared difference between output and targets and lower value of MSE means better performance of train network. The percentage of error indicates the fraction of samples which are misclassified.

**Table 4.2: Neural network training performance in terms of mean square error for training, testing and validation sets.**

| | Samples | MSE | Error (%) |
|---|---|---|---|
| Training Set | 56 | 0.00517713 | 0.892857 |
| Validation Set | 12 | 0.00617574 | 0 |
| Testing Set | 12 | 0.00263427 | 0 |

After 52 iterations, best trained network was constructed with MSE of 0.0061757 at epoch 46 (Figure 4.7). According to confusion matrix in Figure 4.8, it is notable that the best overall accomplished classification was 98.8% of all 80 images in training, validation and testing set. The plot for error distribution of neural network is shown in Figure 4.9. The error histogram plot represents that the error of this proposed system is very close to zero. The progress of other variables such as gradient magnitude and validation checks are illustrated in Figure 4.10. On the training state plot, the maximum validation check 6 at epoch 53 and at this point, the neural network halts the training process to give best performance. Receiver Operating Characteristic (ROC) curve of the network which illustrates true positive rate verses false positive rate at various threshold settings of the network, is shown in Figure 4.11. Area under the curve (AUC) shows a maximum perfect result for this proposed system. At the neural network train, test and validation conclusion, this network performs around 93% correct classification of eight classes of monogenean species.

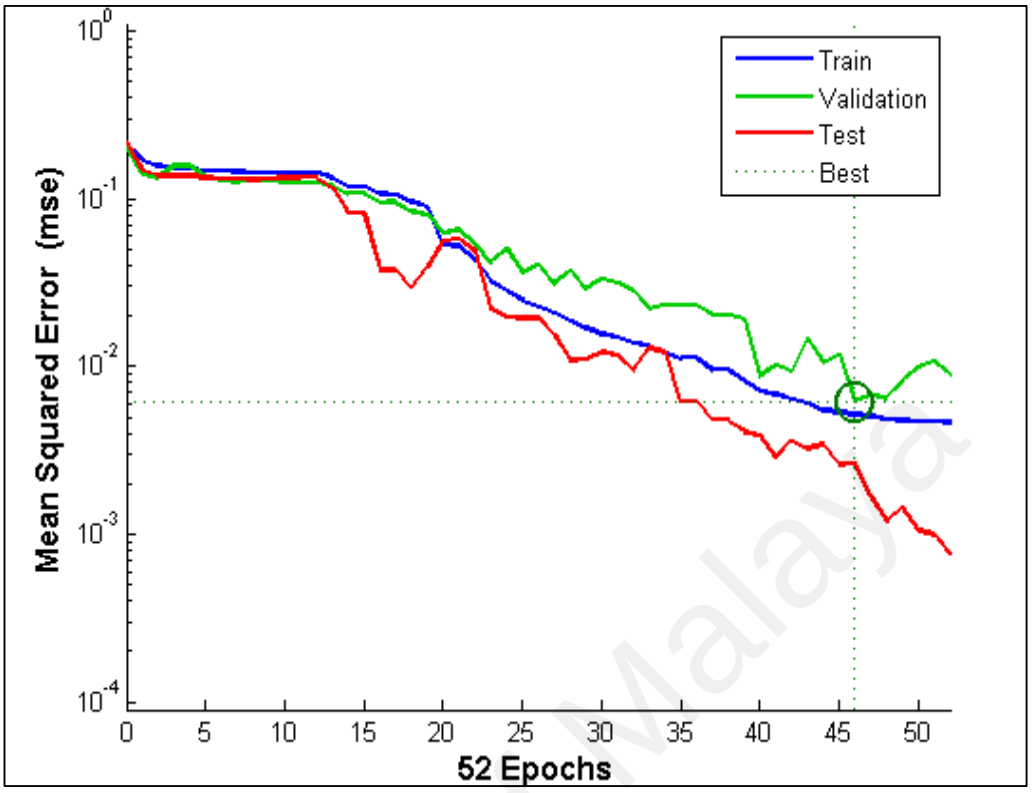


**Figure 4.7: Neural network training validation performance according to mean square error for four species. Best validation performance achieved at epoch 46.**

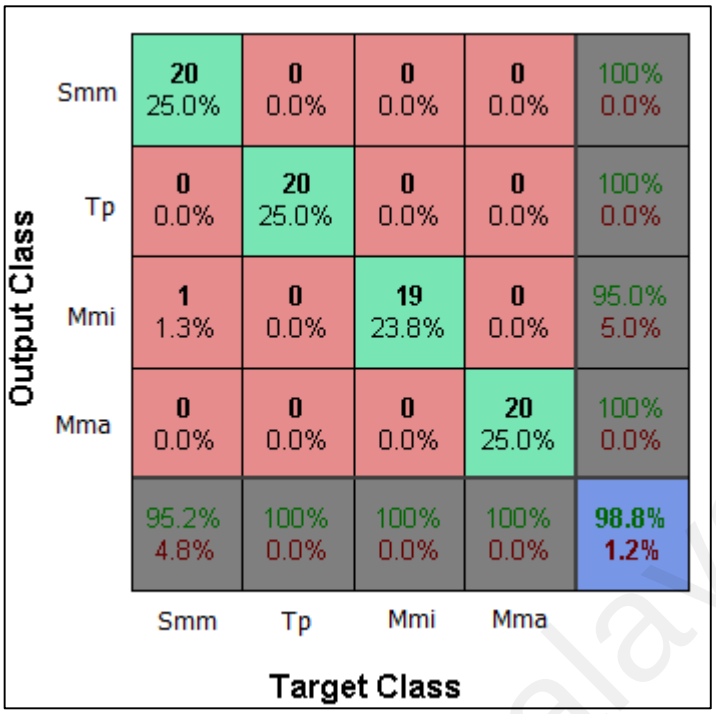


**Figure 4.8: Confusion matrix of testing dataset. The confusion matric shows the classification of four species of monogeneans by ANN classifier. The data was classified into four species: *Sinodiplectanotrema malayanus* (Smm), *Trianchoratus pahangensis* (Tp), *Metahaliotrema mizellei* (Mmi) and *Metahaliotrema similis* (Mma)**

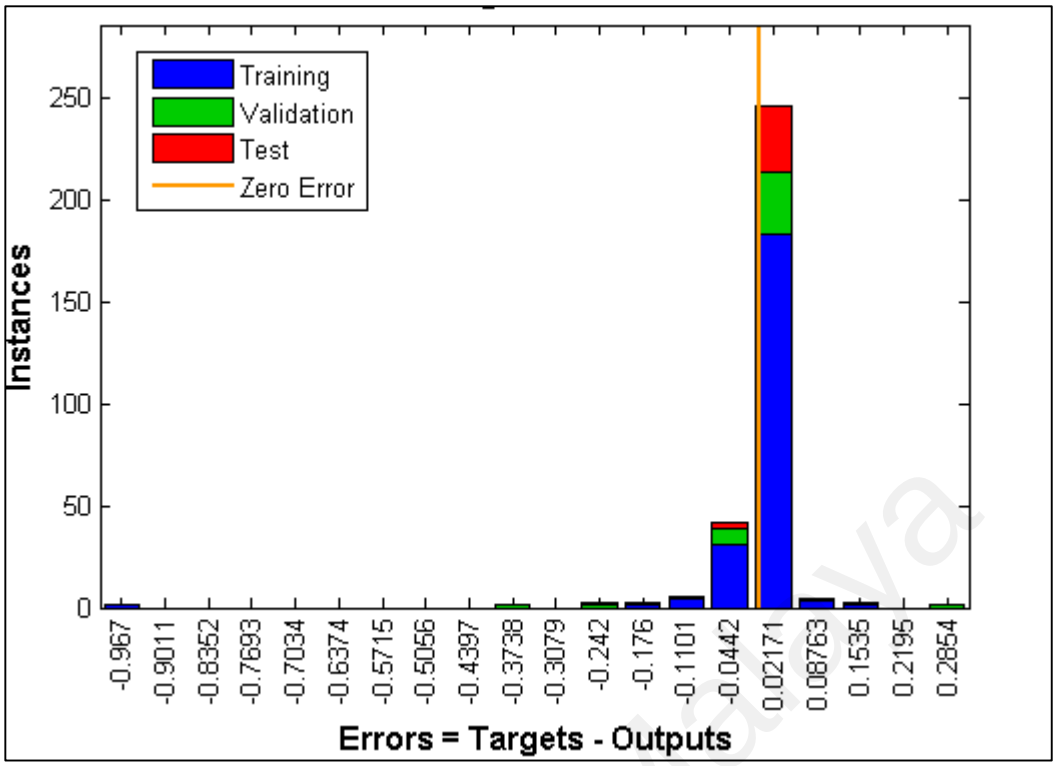


**Figure 4.9: Illustration of distribution of the neural network errors.**

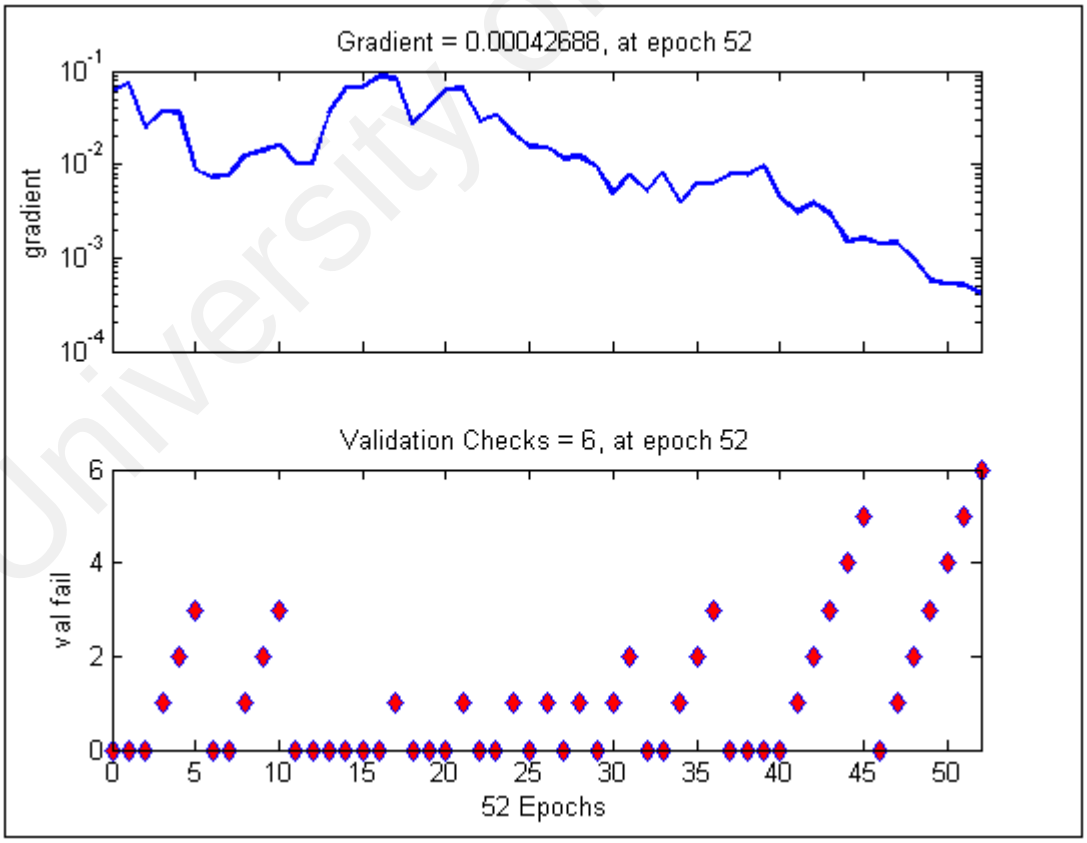


**Figure 4.10: The neural network training state showing the progress of the gradient magnitude and the number of validation checks.**

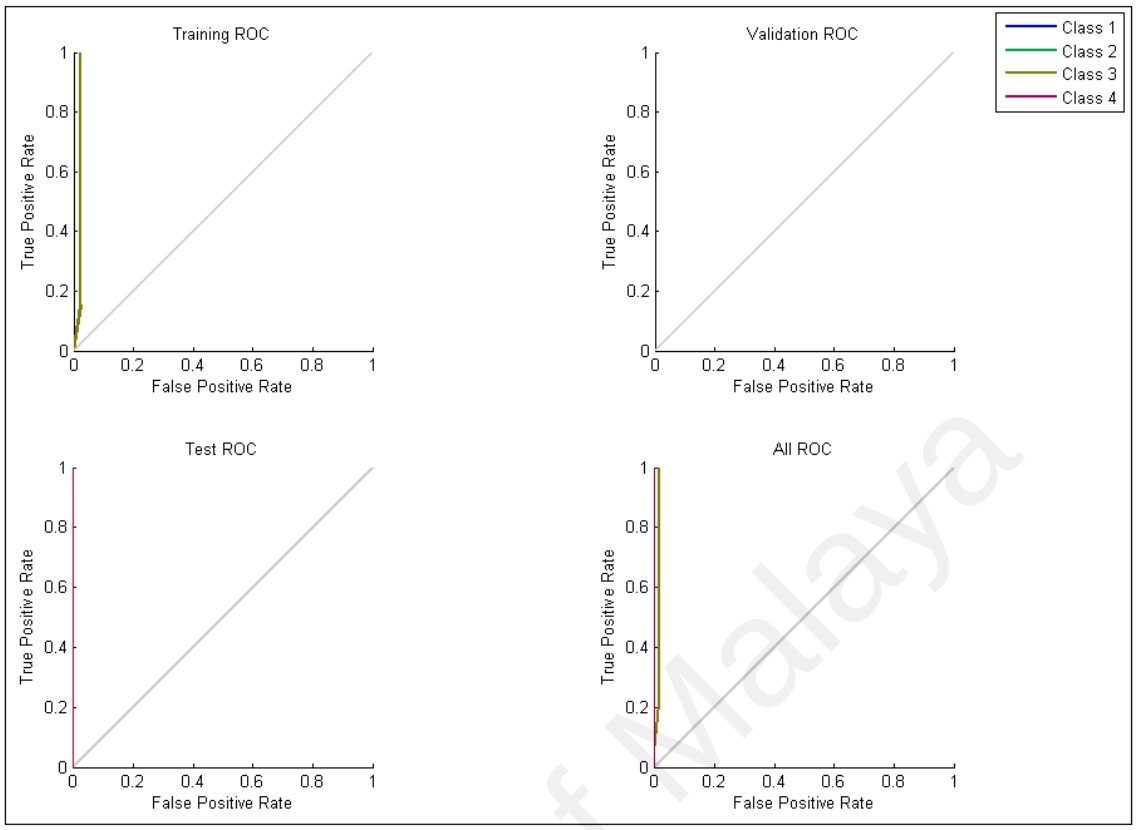


**Figure 4.11: The Receiver Operating Characteristic (ROC) of training network. In the regression plot, a regression between network outputs and network targets is illustrated.**

### 4.1.3 Evaluation

The performance of the system with both classification techniques was evaluated by correct classification accuracy rate of testing data set. A total number of 40 images from image database were assigned to test the system with KNN classification and 12 images to all images were assigned to evaluate and test the system with ANN classification. Additionally, since the sample size was small in this study Leave-One-Out (LOO) cross validation was applied to assess how the results of current system generalize to an independent data set. The results of KNN and ANN reported in confusion matrices in Table 4.1 and Figure 4.8. The result of LOO cross validation is illustrated in Table 4.3 with accuracy score of 91.25%.

**Table 4.3: Confusion matrix of leave one out cross validation for four species of *Sinodiplectanotrema malayanus* (Smm), *Trianchoratus pahangensis* (Tp), *Metahaliotrema mizellei* (Mmi) and *Metahaliotrema similis* (Mma).**

| | | Results | | | Accuracy % |
|---|---|---|---|---|---|
| Species | Smm | Tp | Mmi | Mma | |
| Smm | 20 | 0 | 0 | 0 | 100 |
| Tp | 0 | 18 | 0 | 1 | 90 |
| Mmi | 0 | 0 | 19 | 1 | 95 |
| Mma | 0 | 3 | 1 | 16 | 80 |
| Overall | | | | | 91.25 |

## 4.2 Species Identification Results on Eight Species of Monogeneans (Second Stage)

In this section, the preliminary model for four species is extended to development identification model for eight species of *Sinodiplectanotrema malayanus, Diplectanum jaculator, Trianchoratus pahangensis, Trianchoratus lonianchoratus, Trianchoratus malayensis, Metahaliotrema ypsilocleithru, Metahaliotrema mizellei* and *Metahaliotrema similis.* The experimental results for feature selection, classification and evaluation of automated identification model for eight species are presented. The results are demonstrated to reveal the accuracy of proposed model for classification of eight species of monogenean.

### 4.2.1 Feature Selection

The features extracted for designing preliminary model was not enough to be extended for eight species, therefore, a feature vector with 24 elements was extracted from anchors and bars of eight species. The features were extracted from shape parameters such as Euler number, perimeter, area, area density, perimeter density, centre of bounding box, length of bounding box, width of bounding box, orientation of bounding box, entropy and major axis length. The features were extracted from shape parameters for two times, once, from all anchors and bars of sample as a unit object and the other time from only one anchor of the sample. After LDA feature selection, the feature vector was transformed to feature vector with seve elements. To study the

relationship between the eight species according to the selected features, 2D scatter plots were plotted for each element of selected features. 2D scatter plots in Figure 4.12, Figure 4.13, Figure 4.14, Figure 4.15, Figure 4.16, Figure 4.17 and Figure 4.18 show the discrimination between eight species (represented by eight different colours) by use of only one element of selected features in each plot. The samples from *Metahaliotrema mizellei* (represented by blue colour dots) mingle with *Metahaliotrema similis* (represented by black colour dots) and *Metahaliotrema ypsilocleithru* (represented by brown colour dots). Since these three species are from same genera, it is expected that the features be close. In Figure 4.12, well separation between black, green and red dots is obvious, which shows clustering among *Metahaliotrema, Trianchoratus* and *Sinodiplectanotrema malayanus.* The samples in plots based on one selected feature mingle in different species, but in 3D scatter plot, combination of three selected feature elements (FvLDA1, FvLDA2 and FvLDA3) in Figure 4.19, well separation between samples is illustrated.

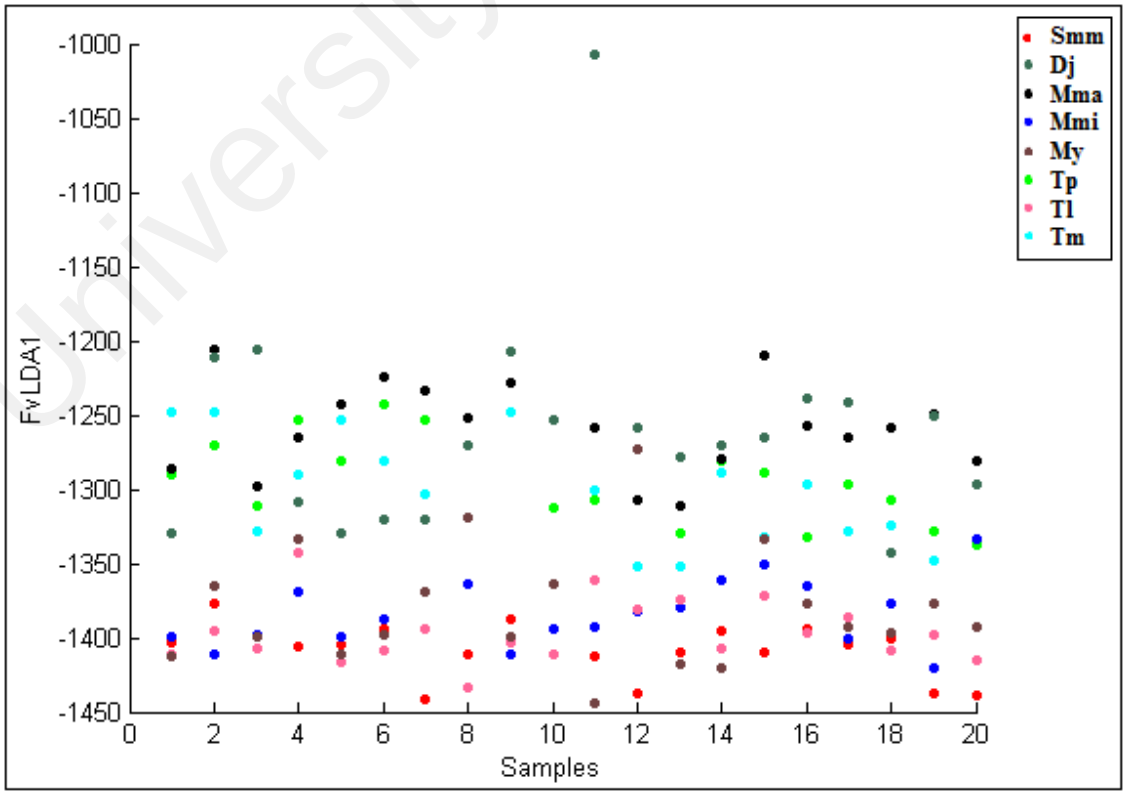

**Figure 4.12: 2D scatter plot of first element of selected feature vector by LDA for samples of *Sinodiplectanotrema malayanus*, *Diplectanum jaculator*,*Trianchoratus pahangensis*, *Trianchoratus lonianchoratus*, *Trianchoratus malayensis*, *Metahaliotrema ypsilocleithru*, *Metahaliotrema mizellei* and *Metahaliotrema similis*.**

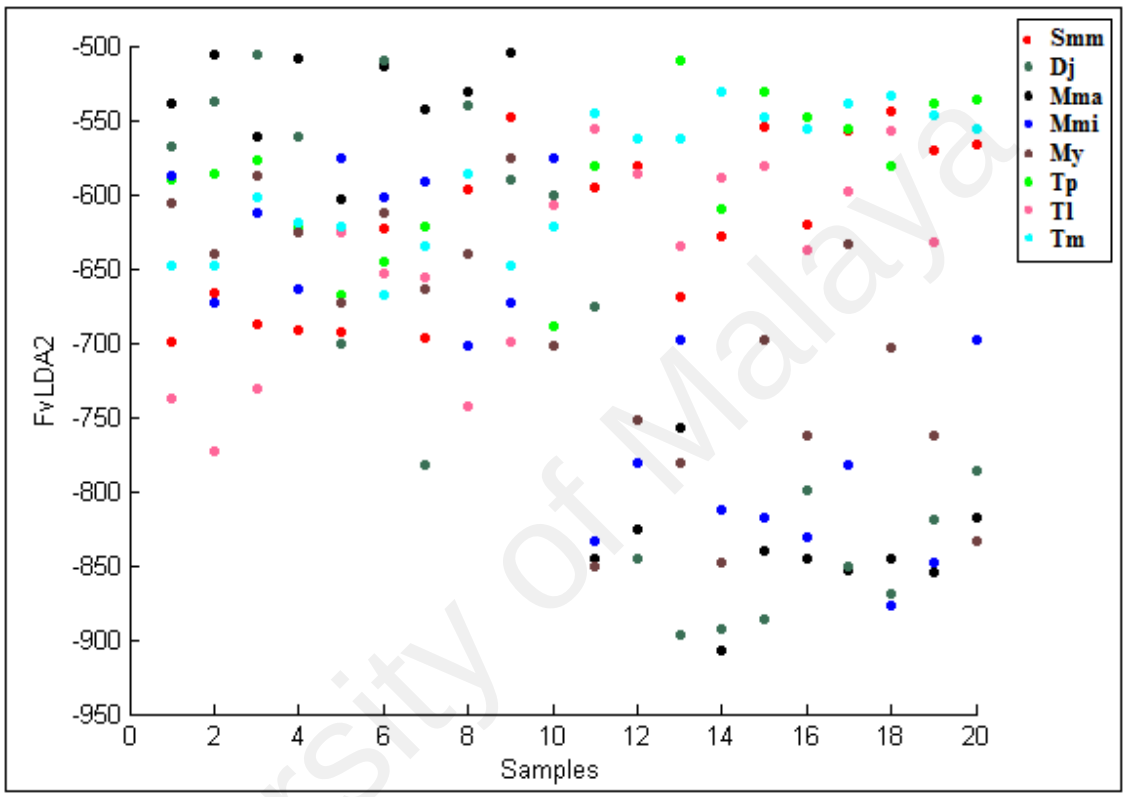


**Figure 4.13: 2D scatter plot of second element of selected feature vector by LDA for samples of *Sinodiplectanotrema malayanus*, *Diplectanum jaculator*,*Trianchoratus pahangensis*, *Trianchoratus lonianchoratus*, *Trianchoratus malayensis*, *Metahaliotrema ypsilocleithru*, *Metahaliotrema mizellei* and *Metahaliotrema similis*.**

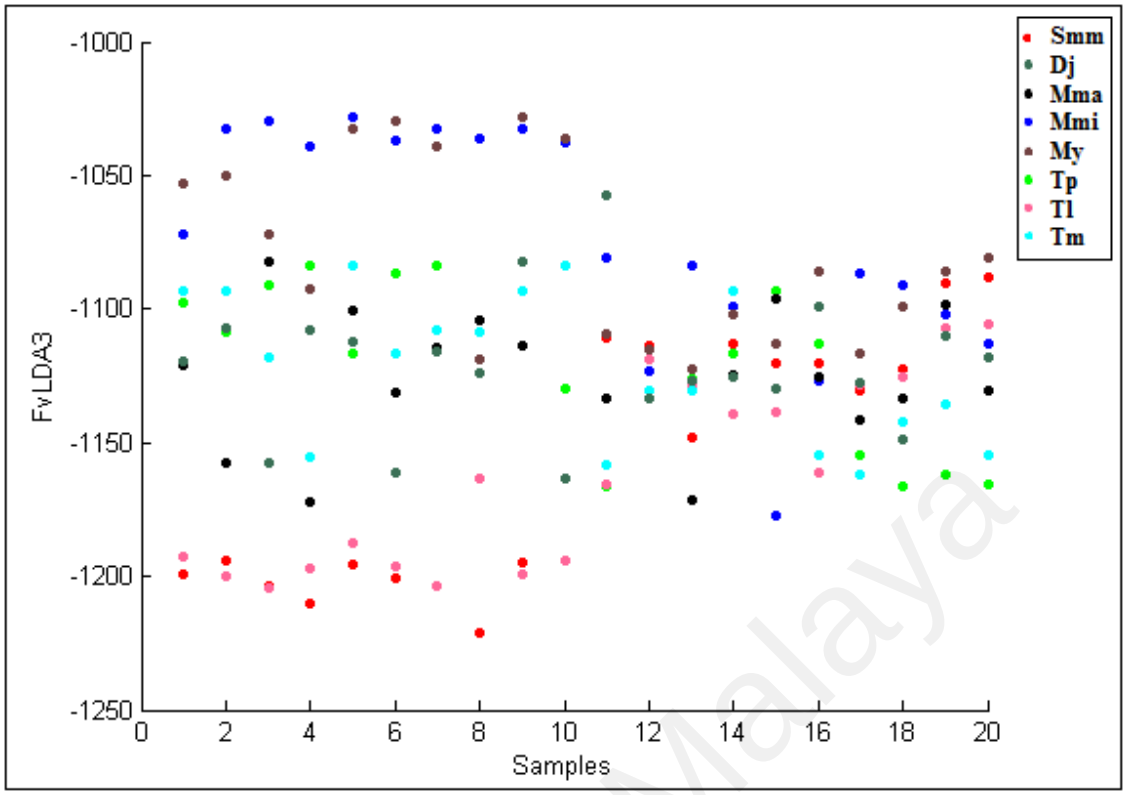


**Figure 4.14: 2D scatter plot of third element of selected feature vector by LDA for samples of *Sinodiplectanotrema malayanus, Diplectanum jaculator,Trianchoratus pahangensis, Trianchoratus lonianchoratus, Trianchoratus malayensis, Metahaliotrema ypsilocleithru, Metahaliotrema mizellei* and *Metahaliotrema similis.***

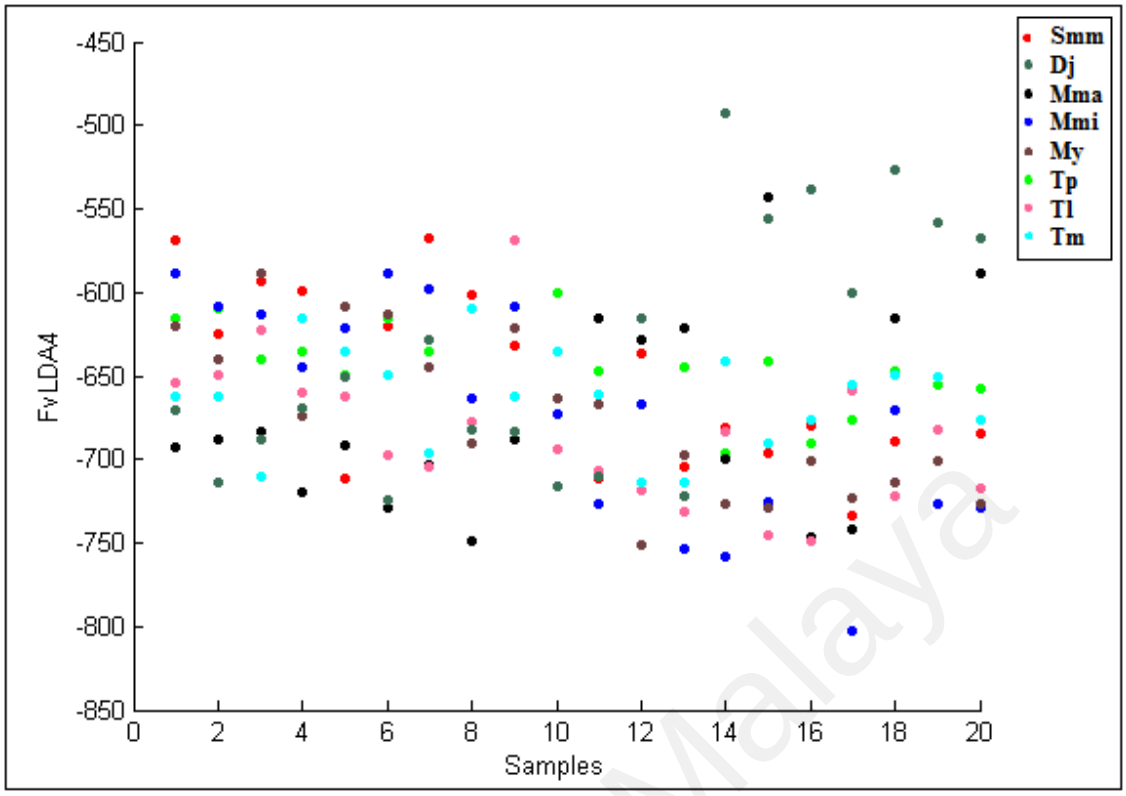


**Figure 4.15: 2D scatter plot of fourth element of selected feature vector by LDA for samples of *Sinodiplectanotrema malayanus*, *Diplectanum jaculator*,*Trianchoratus pahangensis*, *Trianchoratus lonianchoratus*, *Trianchoratus malayensis*, *Metahaliotrema ypsilocleithru*, *Metahaliotrema mizellei* and *Metahaliotrema similis*.**

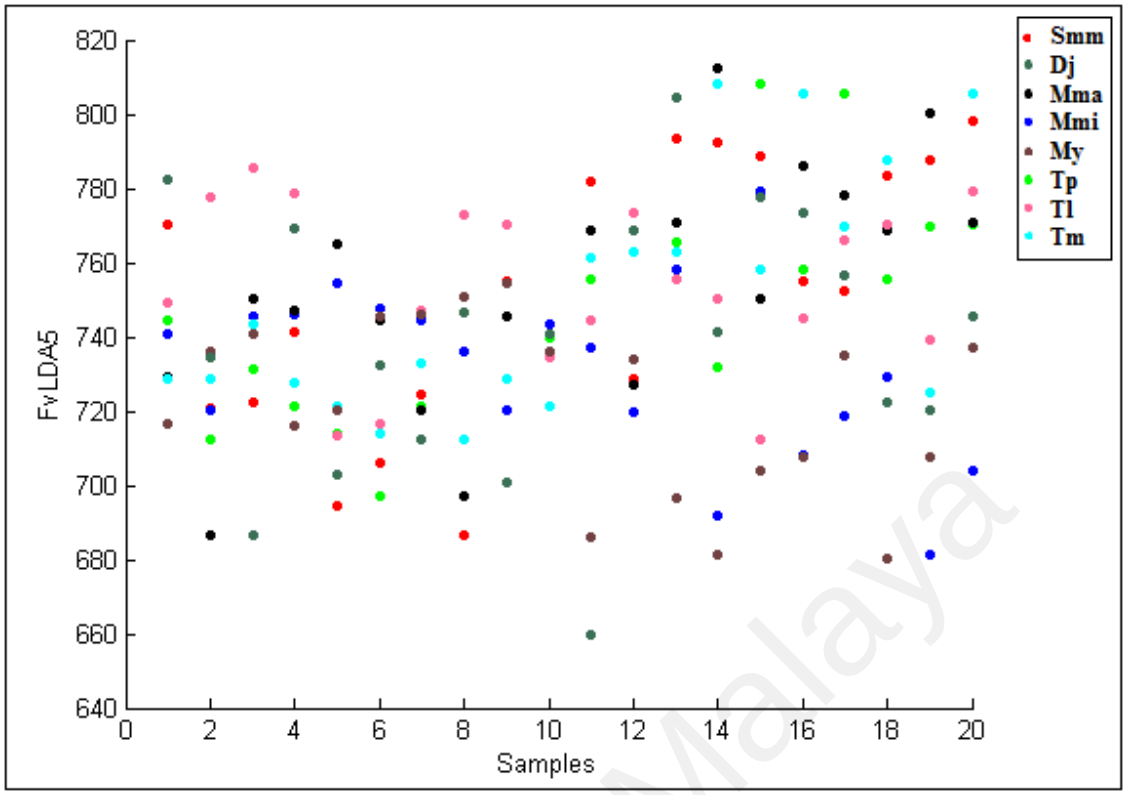


**Figure 4.16: 2D scatter plot of fifth element of selected feature vector by LDA for samples of *Sinodiplectanotrema malayanus*, *Diplectanum jaculator*,*Trianchoratus pahangensis*, *Trianchoratus lonianchoratus*, *Trianchoratus malayensis*, *Metahaliotrema ypsilocleithru*, *Metahaliotrema mizellei* and *Metahaliotrema similis*.**

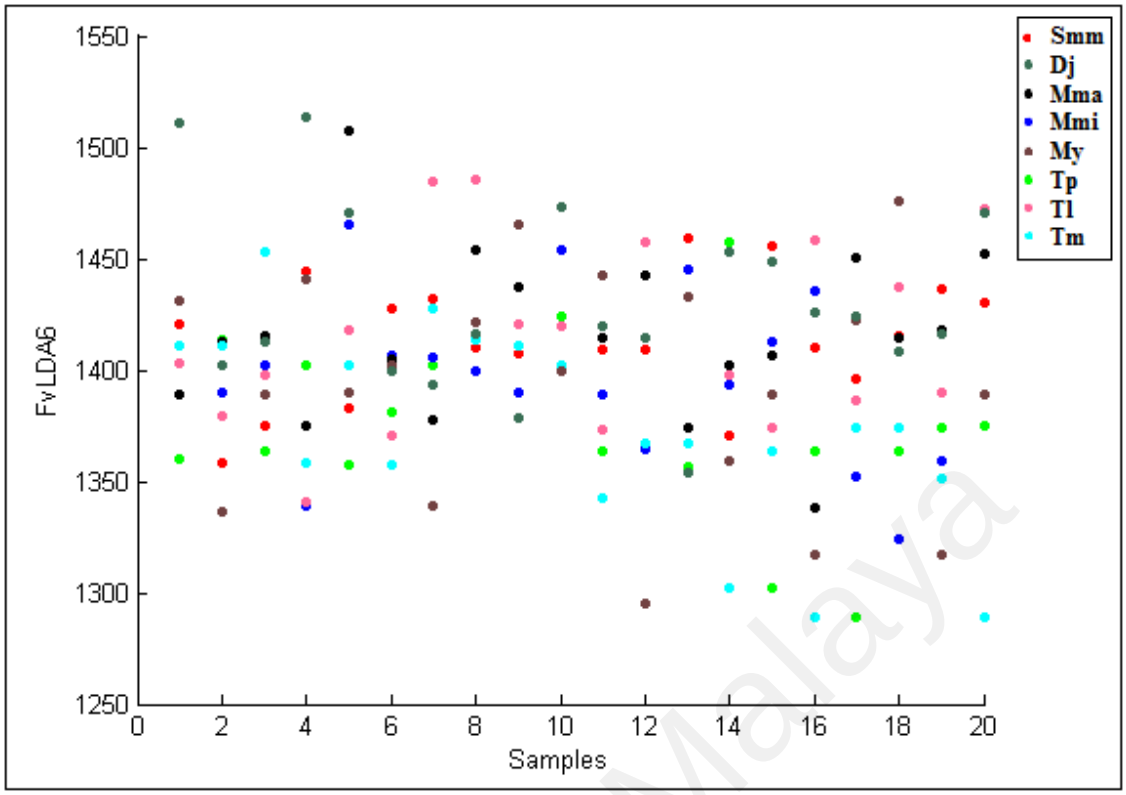


**Figure 4.17: 2D scatter plot of sixth element of selected feature vector by LDA for samples of *Sinodiplectanotrema malayanus*, *Diplectanum jaculator*,*Trianchoratus pahangensis*, *Trianchoratus lonianchoratus*, *Trianchoratus malayensis*, *Metahaliotrema ypsilocleithru*, *Metahaliotrema mizellei* and *Metahaliotrema similis*.**

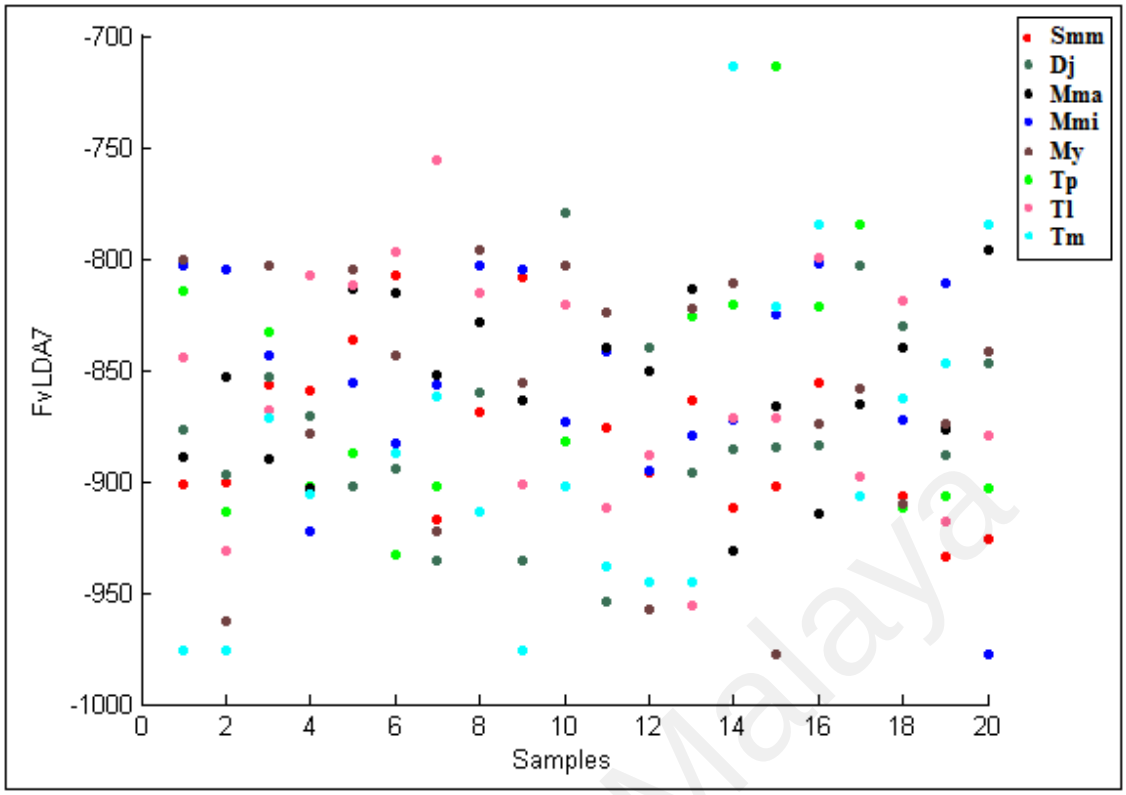


**Figure 4.18: 2D scatter plot of seventh element of selected feature vector by LDA for samples of *Sinodiplectanotrema malayanus*, *Diplectanum jaculator*,*Trianchoratus pahangensis*, *Trianchoratus lonianchoratus*, *Trianchoratus malayensis*, *Metahaliotrema ypsilocleithru*, *Metahaliotrema mizellei* and *Metahaliotrema similis*.**

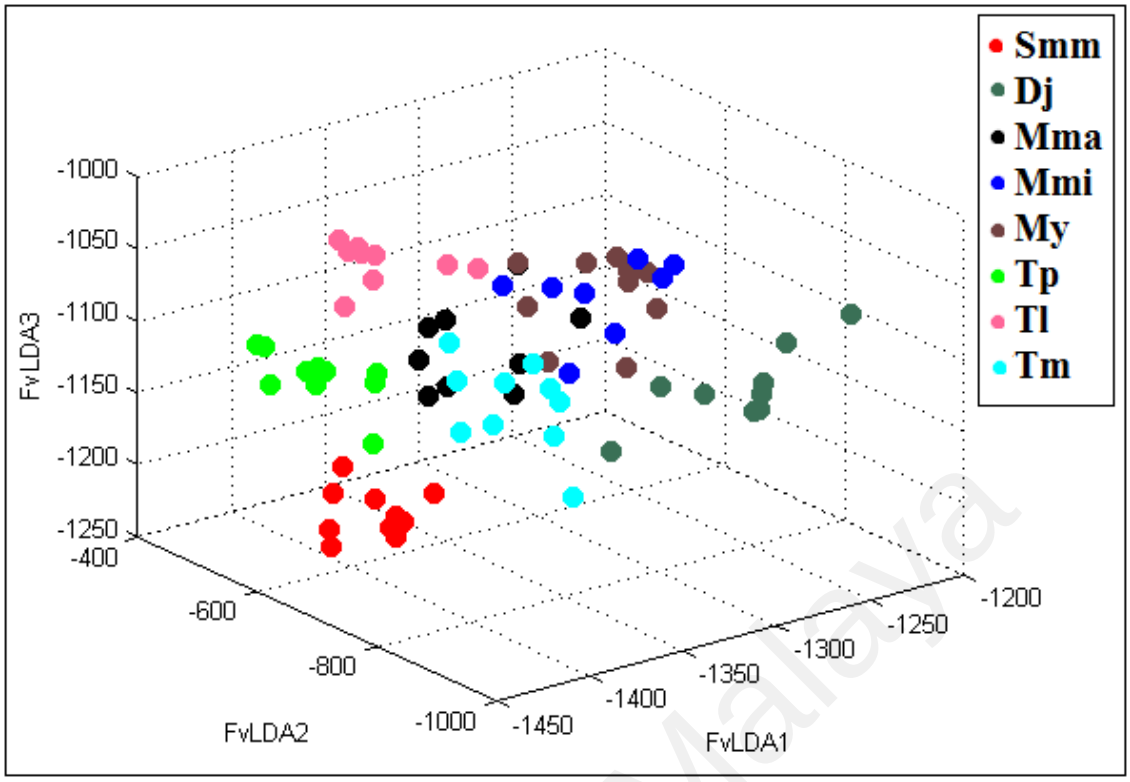


**Figure 4.19: 3D scatter plot with combination of LDA selected features: FvLDA1, FvLDA2 and FvLDA3. The samples were classified into eight classes illustrated with eight circles in different colours.**

In Figure 4.19 it is shown how transformed feature vector separates eight species; by adopting LDA feature selection method, the feature vector with 24 elements was transformed to feature space with seven distinct elements in feature space (Figure 4.20).

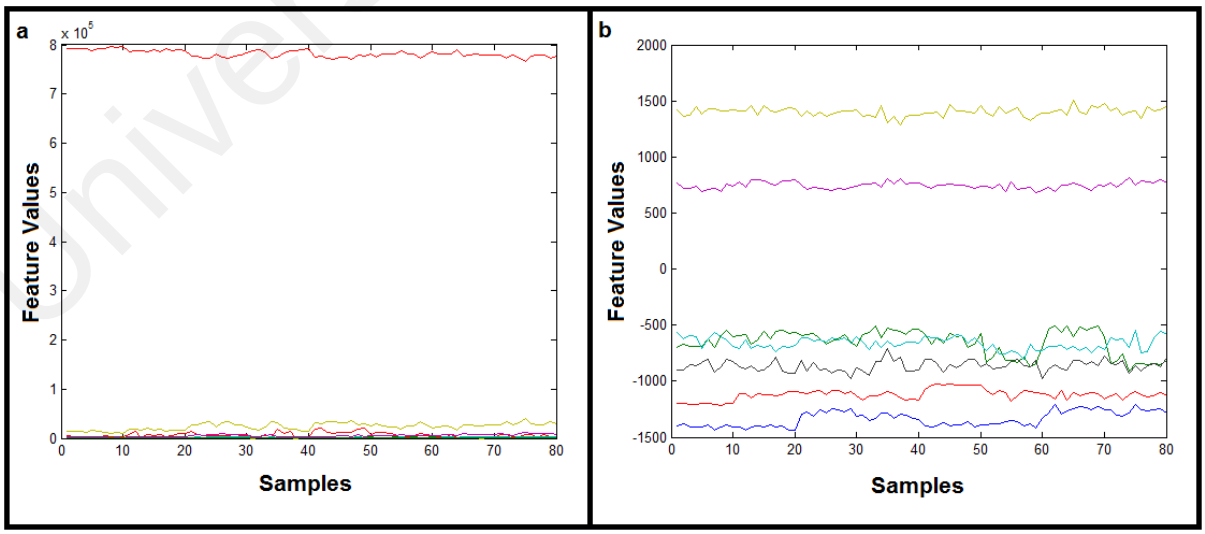


**Figure 4.20: Feature vector comparison after and before feature selection. a) Illustration of 24 dimensional extracted feature vector for 80 samples. Except one of the features, the rest contain close values. b) Illustration of seven dimensional feature vector which is the result of LDA feature selection.**

### 4.2.2 Classification

The experiment was conducted on eight species of four monogenean families which were classified by KNN and ANN. In KNN 80 images were used for training and 80 images for testing the trained model. In ANN, 112 images were used for training, 24 images for testing the network and 24 images for system validation. In achieved results, ANN with accuracy of 93.1% was outperforming KNN classifier with accuracy of 86.25%.

#### 4.2.2.1 K-Nearest Neighbour (KNN)

In KNN classification, we achieved best classification score with nine nearest neighbours (Figure 4.21). According to the confusion matrix (Table 4.4), the overall classification score for eight species was 86.25%. KNN was also employed to classify intra genus specimens of *Metahaliotrema* and *Trianchoratus*. The confusion matrix in Table 4.5 (A) shows the classification result in *Metahaliotrema* and the confusion matrix in Table 4.5 (B) shows the classification result in *Trianchoratus*. The accuracy of classification in *Metahaliotrema* genus was 76.66%. There were three misclassification of *Metahaliotrema ypsilocleithru* with *Metahaliotrema mizellei* and two with *Metahaliotrema similis*. Also it is notable that in Table 4.4 there are two misclassification of *Metahaliotrema ypsilocleithru* with *Metahaliotrema mizellei*. Mainly the misclassification between these two species is because of the shape of their anchors and the way dorsal and ventral anchors lie in front of each other.

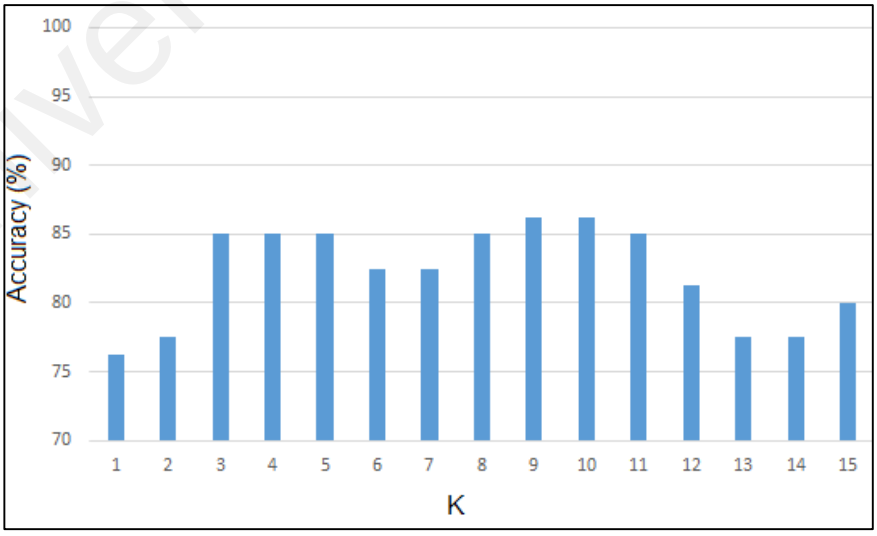


**Figure 4.21: Illustration of k value in 15 iterations of KNN classification for eight species. The best result was shown by k=9 and k=10.**

**Table 4.4: Confusion matrix of KNN classification for eight species.**

| | Results | | | | | | | | Accuracy (%) |
|---|---|---|---|---|---|---|---|---|---|
| Species | Smm | Tp | Mmi | Mma | Tl | Tm | My | Dj | |
| Smm | 10 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 100 |
| Tp | 0 | 9 | 0 | 1 | 0 | 0 | 0 | 0 | 90 |
| Mmi | 0 | 1 | 9 | 0 | 0 | 0 | 0 | 0 | 90 |
| Mma | 0 | 0 | 0 | 10 | 0 | 0 | 0 | 0 | 100 |
| Tl | 0 | 0 | 1 | 0 | 8 | 0 | 1 | 0 | 80 |
| Tm | 0 | 1 | 0 | 0 | 0 | 9 | 0 | 0 | 90 |
| My | 0 | 1 | 2 | 1 | 0 | 1 | 5 | 0 | 50 |
| Dj | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 9 | 90 |
| Overall | | | | | | | | | 86.25 |

**Table 4.5: Confusion matrix of monogenean Intra-genus KNN classification. A) *Metahaliotrema* samples B) *Trianchoratus* samples**

| **A** | **Results** | | | **Accuracy (%)** |
|---|---|---|---|---|
| Species | Mmi | Mma | My | |
| Mmi | 8 | 1 | 1 | 80 |
| Mma | 0 | 10 | 0 | 100 |
| My | 3 | 2 | 5 | 50 |
| Overall | | | | 76.66 |

| **B** | **Results** | | | **Accuracy (%)** |
|---|---|---|---|---|
| Species | Tp | Tl | Tm | |
| Tp | 10 | 0 | 0 | 100 |
| Tl | 0 | 8 | 2 | 80 |
| Tm | 0 | 2 | 8 | 80 |
| Overall | | | | 86.66 |

#### 4.2.2.2 Artificial Neural Network (ANN)

The ANN classification structure was a two layer feed-forward network which was trained with back propagation and with respect to ten hidden neurons in hidden layer and eight neurons in output layer. After 46 iterations, best trained network was constructed with MSE of 0.026168 at epoch 40 (Figure 4.22). In this experiment, the MSE value for training, testing and validation set is reported in Table 4.6. MSE is the average squared difference between output and targets and lower value of MSE means better performance of trained network. The percentage of error indicates the fraction of samples which are misclassified. According to confusion matrix in Figure 4.23, it is notable that the best overall accomplished classification was 93.1% of all 160 images in training, validation and testing set. The plot for error distribution of neural network is shown in Figure 4.24. The error histogram plot represents that the error of this proposed system is very close to zero. The progress of other variables such as gradient magnitude

and validation checks are illustrated in Figure 4.25. On the training state plot, the maximum validation check 6 at epoch 45 and at this point, the neural network halts the training process to give best performance. Receiver Operating Characteristic (ROC) curve of the network which illustrates true positive rate verses false positive rate at various threshold settings of the network, is shown in Figure 4.26. Area under the curve (AUC) shows a maximum perfect result for this proposed system. At the neural network train, test and validation conclusion, this network performs around 93% correct classification of eight classes of monogenean species.

**Table 4.6: Neural network training performance in terms of Mean Square Error (MSE) for training, testing and validation sets**

| | Samples | MSE | Error (%) |
|---|---|---|---|
| Training Set | 112 | 0.00920595 | 8.92857 |
| Validation Set | 24 | 0.0261682 | 8.33333 |
| Testing Set | 24 | 0.0205884 | 8.33333 |

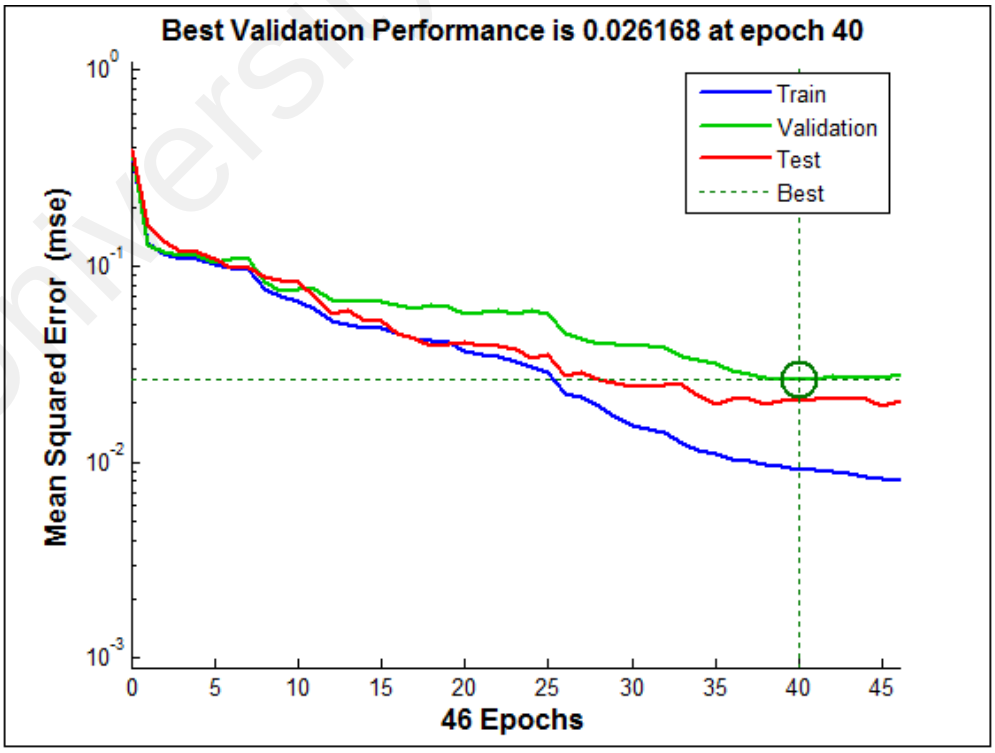


**Figure 4.22: Neural network training validation performance according to mean square error for eight species. Best validation performance achieved at epoch 40.**

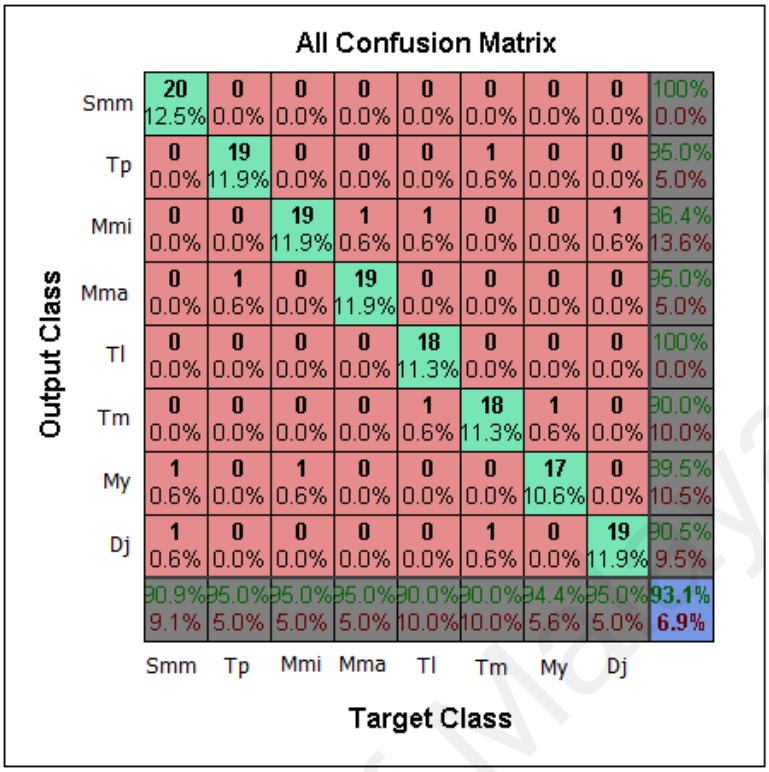


**Figure 4.23: Confusion matrix of testing dataset. The confusion matric shows the classification of eight species of monogeneans by ANN classifier.**

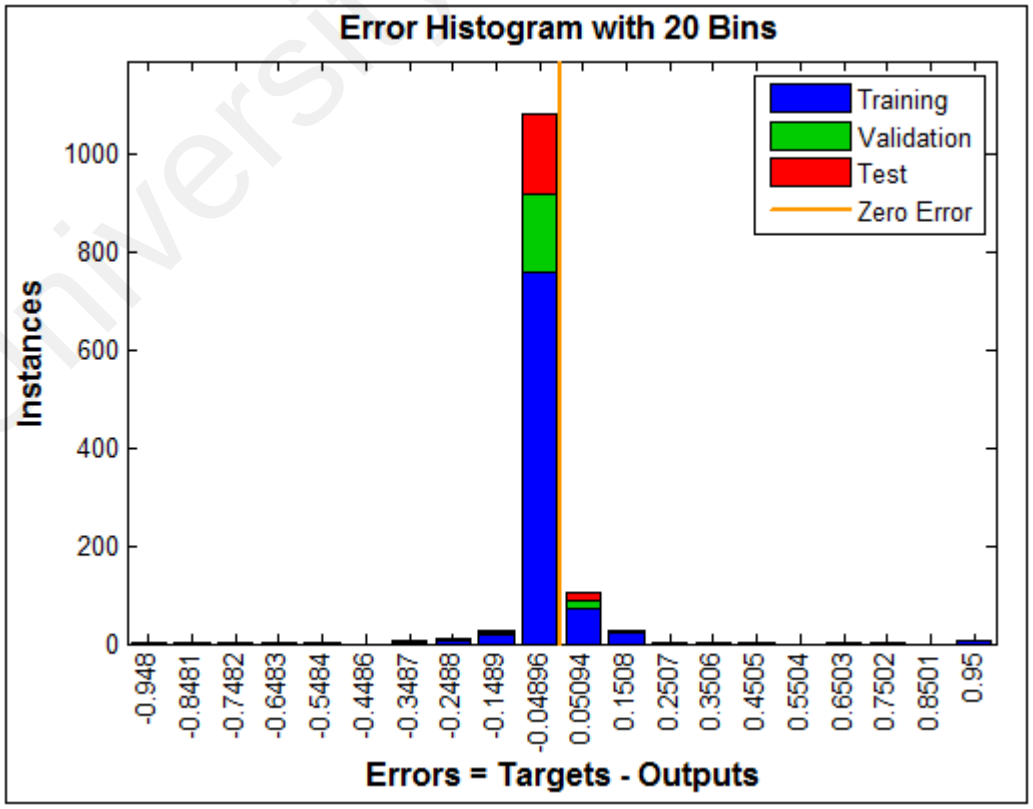


**Figure 4.24: Illustration of distribution of the neural network errors.**

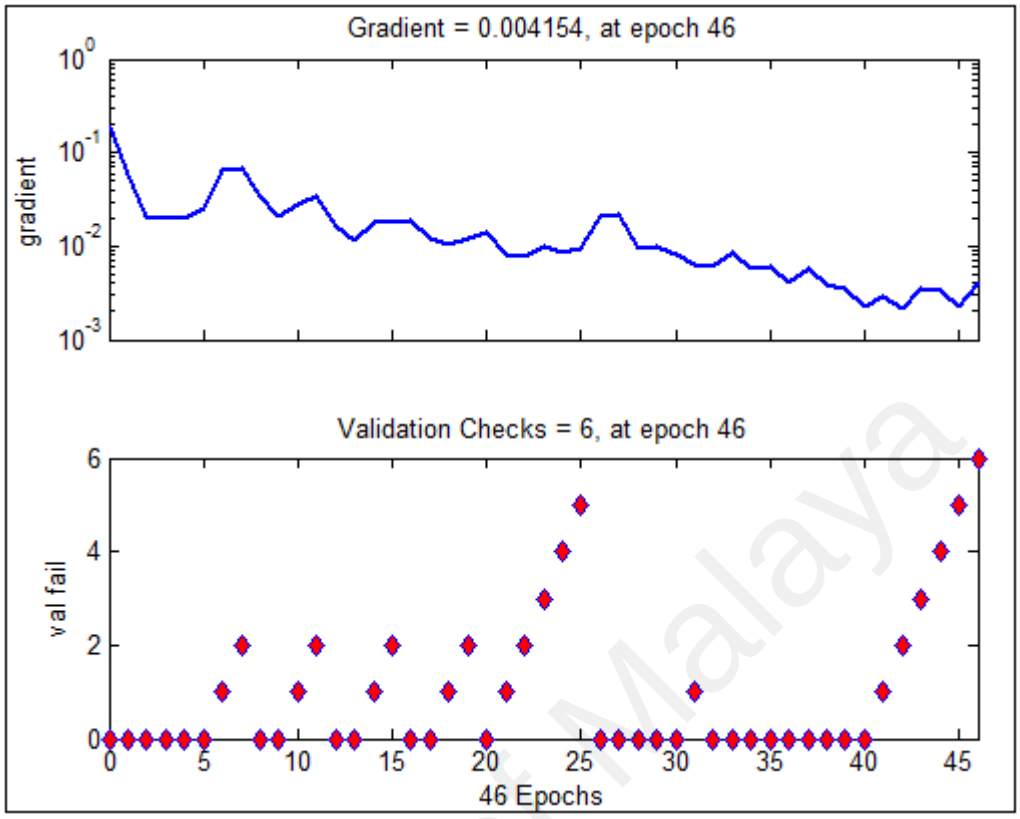


**Figure 4.25: The neural network training state showing the progress of the gradient magnitude, the number of validation checks**

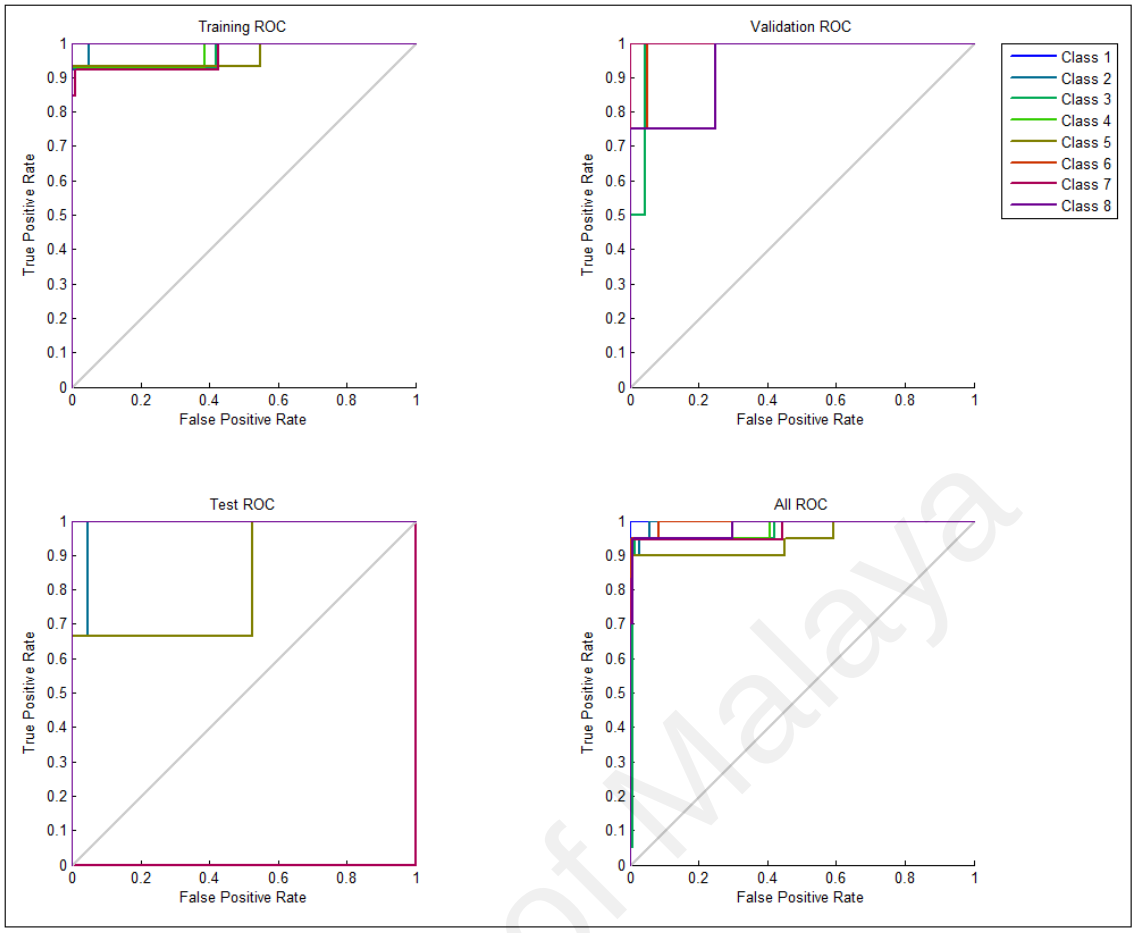


**Figure 4.26: The Receiver Operating Characteristic of training network. In the regression plot, a regression between network outputs and network targets is illustrated.**

### 4.2.3 Evaluation

The performance of the system with both classification techniques was evaluated by correct classification accuracy rate of testing data set. A total number of 80 images from image database were assigned to test the system with KNN classification and 24 images were assigned to evaluate and test the system with ANN classification. Also, since the sample size was small in this study Leave-One-Out (LOO) cross validation was applied to assess how the results of our system generalize to an independent data set. The result for KNN and ANN classificatiom reported in confusion matrices in Table 4.4 and Figure 4.23 The result of LOO cross validation is illustrated in Table 4.7 with accuracy score of 88.13%.

**Table 4.7: Confusion matrix of leave one out cross validation for eight species of *Sinodiplectanotrema malayanus (Smm)*, *Diplectanum jaculator (Dj)*,*Trianchoratus pahangensis (Tp)*, *Trianchoratus lonianchoratus (Tl)*, *Trianchoratus malayensis (Tm)*, *Metahaliotrema ypsilocleithru (My)*, *Metahaliotrema mizellei (Mmi)* and *Metahaliotrema similis (Mma)*.**

| | Results | | | | | | | | Accuracy (%) |
|---|---|---|---|---|---|---|---|---|---|
| Species | Smm | Tp | Mmi | Mma | Tl | Tm | My | Dj | |
| Smm | 19 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 95 |
| Tp | 0 | 19 | 0 | 1 | 0 | 0 | 0 | 0 | 95 |
| Mmi | 0 | 1 | 18 | 0 | 0 | 0 | 2 | 0 | 90 |
| Mma | 0 | 3 | 0 | 16 | 0 | 0 | 1 | 0 | 80 |
| Tl | 0 | 0 | 1 | 0 | 18 | 0 | 1 | 0 | 90 |
| Tm | 1 | 1 | 0 | 0 | 0 | 18 | 0 | 0 | 90 |
| My | 0 | 0 | 3 | 2 | 0 | 1 | 14 | 0 | 70 |
| Dj | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 19 | 95 |
| Overall | | | | | | | | | 88.13 |

## 4.3 Overall Results

The overall results of preliminary (first stage) and extended models (second stage) of automated identification of monogenean images, are presented in Table 4.8. According to the results, ANN outperforms KNN in both preliminary and extended models.

**Table 4.8: The performance of classification techniques` in preliminary and extended models**

| | **KNN** | **ANN** | **LOO** |
|---|---|---|---|
| **Preliminary model (first stage)** | 95% | 98.80% | 91.25% |
| **Extended model (second stage)** | 86.25% | 93.10% | 88.13% |

# CHAPTER 5: DISCUSSION AND CONCLUSIONS

The proposed automated identification method in this study is able to classify monogenean to species level with the overall accuracy of 86.25% with K-nearest neighbour (KNN) classification and 93.1% with Artificial Neural Network (ANN) classification for eight species of monogenean. In this study a model based was developed for monogenean images which can assist taxonomists and non-taxonomists or ecologists to identify monogenean according to image of their anchors and bars.

Generally, morphometric approaches are built according to distance measurements (Gussev, 1976). However, results of morphometric analyses can depend upon the acquired images (Kalafi, Tan, Town, & Dhillon, 2016) and the particular set of measurements chosen (Strauss & Bookstein, 1982; Rohlf & Marcus, 1993). According to several authors` believe, most morphological features that are extracted from haptoral hard parts are highly correlated (Shinn, des Clers, Gibson, & Sommerville, 1996; Du Preez & Maritz, 2006) and automatic classification of monogenean species require improved discriminant methods for such multicollinearity, especially for small sample sizes where several morphological measurements are used to classify a few individuals (Vignon, 2011a).

In this study, some set of measurements from shape parameters failed to capture the complete spatial arrangement of the anatomical features. Due to preserving geometric information from data collection, Linear Discriminant Analysis (LDA) was used for transforming extracted features to new feature vector. Also alternative method, based on the overall form of the haptoral hard parts, was adopted for taxonomic diagnoses of monogenean species. Combination of such method could free taxonomists from collections of landmarks and associated linear distances by directly taking into account the shape and size information of morphological features. This provided a better discrimination between individuals or species than by use of the traditional system.

## 5.1 Image Acquisition and Database

Traditionally, the morphological classification of monogenean species is based on measurements from shape of individual hard structures such as haptoral parts and copulatory organs. Therefore, images were focused on anchors and bars of specimens since these organs contain diagnostic features which are used for classification of monogenean species. Using overall form of anchors and bars for extraction of features were lead to achieve new characters in morphological classification of monogeneans which has been never used before. The need for the discovery of new characters for identification of species has been acknowledged for log by systematic parasitology (Vignon, 2011a) and because of the lack of discrimination of traditional methods, several researchers have used additional points to take into account the maximum amount of shape information (Murith & Beverley-Burton, 1985; Rehulkova ´ & Gelnar, 2005).

Although the best slides of specimens were prepared, but still because of limited number of some specimens, overlapping, broken specimens and clutters in slides were unavoidable and this caused image acquisition not to be always perfect. Since the feature extraction process is highly affected by the quality of images, therefore, one of important factors in classification is the quality and clearness of images. This could be achieved by using better specimens` slides and high quality microscope and attached camera especially in terms of lenses.

The acquired images were in two dimensional (2D) and due to loss of some information in 2D imaging, it is suggested that in future, the model can be based on three dimensional (3D) images. As the solution to loss of information in 2D imaging, in the study by Leow et al. (2015), they used built in function in imaging software, called Extended Focus Imaging (EFI) to create a single plane image with in-focus details.

The acquired images were stored in a database and based on classification method, the database was divided into testing, training and validation sets.

### 5.2 Monogenean Identification

Two classification techniques, KNN and ANN were adopted for developing automated identification model for monogeneans. Since successful experiments by using these two classifiers with small size of samples have been reported (Jin, Hou, Li, & Zhou, 2015; Ali, Hussain, Bron, & Shinn, 2012), it was reasonable to use KNN and ANN in current study. However, other classification techniques such as SVM, DA, and decision tree may improve the performance of the system if the size of database is increased as the performance of classification in some of these methods (e.g. SVM) is dependent on size of training samples (Maglogiannis, 2007).

KNN and ANN invoked features which were selected by adoption of LDA technique for transforming feature vectors to distinct feature space of seven elements. In both KNN and ANN, *Sinodiplectanotrema malayanus* was correctly classified in all cases due to distinct shape and size of anchors and bars of the species. Also the sample images of this species were clear and anchors were perfectly recognised. There was one misclassification of *Trianchoratus pahangensis* as *Metahaliotrema similis* by KNN method. Mainly, because the shape of their anchor`s tails were similar and one misclassification with *Trianchoratus malayensis* by ANN as both of them have three anchors and from same genus. There was one misclassification of *Metahaliotrema mizellei* with *Metahaliotrema similis* by KNN since both are from same genus, overall shape of all anchors and bars as an object is similar. In KNN, the classification of *Metahaliotrema similis* was 100% correct while by ANN there was one misclassification with *Trianchoratus pahangensis* as the similar shape of their anchor`s tails. The classification of *Trianchoratus lonianchoratus* by ANN was 100% correct while there were two misclassifications with *Metahaliotrema mizellei* and

*Metahaliotrema ypsilocleithru* by KNN. Since the identification of *Trianchoratus lonianchoratus* by ANN is 100% correct this means the features were distinct enough for training the network but the distance distinction by KNN was not sufficient for classification. In classification of *Trianchoratus malayensis* samples by KNN, one was misclassified as *Trianchoratus pahangensis*. The anchors of both species are similar in shape but distinct in size. In ANN classification, *Trianchoratus malayensis*, had one misclassification with *Trianchoratus lonianchoratus* and one misclassification with *Metahaliotrema ypsilocleithru*. Mainly, the images from samples of *Metahaliotrema ypsilocleithru* species were not well pre-processed. Due to overlapping of anchors and bars in images, even it is not easy for human eyes to separate them. Therefore, this is the main reason for misclassification of *Metahaliotrema ypsilocleithru* with other species.

### 5.3 Comparison with Previous Studies

The presented automated monogenean identification model in this study, used shape descriptor parameters as distinguishing features and KNN and ANN as classification techniques in pattern recognition tool to identify and classify monogeneans. Considerably, this is the first fully automated identification model for monogeneans based on monogenean diagnostic organs which are haptoral bars and anchors. In previous studies of monogenean specimes` classifications, measurements were attained from hard structure of monogeneans based on landmarks (Vignon, 2011a; Ali, Hussain, Bron, & Shinn, 2011; Khang, Soo, Tan, & Lim, 2016). But we used new morphological measurements from overall shape of all anchors and bars and successfully classified eight species according to those characters.

In 1999, an experiment was conducted by (Kay et al., 1999) in which they classified the specimens of monogeneans (*Gyrodactylus colemanensis*; *Gyrodactylus derjavini*; *Gyrodactylus caledoniensis*; *Gyrodactylus truttae*; *Gyrodactylus salaris*). They used

and compared four classification techniques in their study: Nearest Neighbours (NN), Feed-Forward Neural Network (FFNN), Projection Pursuit Regression (PPR) and Linear Discriminant Analysis (LDA). The classification by NN and LDA from total sclerotized structures acquired by light microscopy had best results among all classification methods. In the present study, the advantage of research by Kay et al (1999) was taken to choose Artificial Neural Network technique for the classification. In the previous study ( Ali, Hussain, Bron, & Shinn, 2011), the multi stage classification technique was developed for classification of nine species of *Gyrodactylus* by using LDA, KNN and Naïve Bayes (NB) techniques. They extracted 25 features from shape descriptors of anchors, ventral bar which spans the two anchors and marginal hooks. In this study, the features were extracted from ventral and dorsal anchors and bars.

In previous studies, the image processing stage was manual and features were extracted by manual pointing of landmark coordinates whereas in this study, all stages, including image processing was automated. Although the detected edges and segmented images were not perfect, but still could be used for feature extraction. In future by improvement of quality of samples and digitized images, the automatic image processing will be enhanced.

**5.4 Constraints and Limitations**

Since some of the specimens`slides were old or some were not preserved in good condition, the specimens inside them were blemished. Some specimens were broken and the background of some specimens was cluttered due to compression of monogenean`s soft parts under slides. Finding slides which contain specimens in good condition was time consuming. Still, in some cases, because of small number of available specimens, using improper slides was unavoidable.

The quality of images is one of the important factors in image analysis and it is highly affected by imaging tools and equipment. During first two months of this study,

the images were taken from three species using JVC TK-1280E colour video camera attached to Leitz Diaplan microscope. Comparing to other 23 species that have been digitized by Leica Digital Camera DFC 320 attached to Leica Leitz DMRB microscope, the quality of 23 species` images was better than the first three.

The diagnostic organs of monogenean which were used in this study are haptoral anchors and bars. Most of monogeneans have four anchors (2 dorsal and 2 ventral) and two bars (1 dorsal and 1 ventral). The geometrical structure of dorsal and ventral anchors and bars are overlapping and separating them during image processing was difficult.

With respect to conversion of three dimensional (3D) vision under the microscope lenses to two dimensional (2D) digital images, it is noticeable that some information will be lost. By use of 3D imaging equipment this weakness of automated identification system will be reduced. Another solution for this matter is focus stacking of multiple images taken at different focus distances (e.g. EFI function).

### 5.5 Future Works

As an idea to improve the automated identification model is to increase the size of datasets in future studies. By extending the size of training set, more features can be achieved and samples within a class can be identified more accurately. Also, the number of species in database could be expanded. For further application with complex models, incrementing the number of samples may yield better results. Currently, the models`s database consists of 160 images from eight different species. In addition to number of images, the number of species can be extended and the number of images will be expanded with increase in number of species used in database. By increased quantity of images, other classification techniques can also be used and a considerably more detailed, including statistical, evaluation can be performed. There are many

classification techniques such as SVM, DA, and decision tree may improve the performance of the system in future.

Besides the number of images, quality of images is an important factor in automated classification of species. Since the images processing stage is automated and the same threshold is used for all images, it is crucial that the imaging condition (e.g. light, focus, magnification) be equal during image acquisition. In this study, some of specimens were old and as a result, the quality of images acquired, was not good enough. In future works, the quality of all images should be standardized and image acquisition has to be done with better equipment such as better microscope and camera in terms of lenses and light source.

According to previous study by Khang et al. (2016) and Abu et al. (2013), monogenean classification can be based on extracted features from only anchors and bars. Therefore, in this study, the features were extracted from shape parameters of only anchors and bars, but other than these organs, monogenean can be classified by morphometric information of male and female copulatory organs and marginal hooks (Tan, 2013). In future studies, the morphological data from shape parameters of all anchors, bars, marginal hooks and copulatory organs can be used as input to classification techniques and results would be more reliable. Also, other feature extracting techniques which can extract further informative features may help to improve the future studies. One of these techniques is skeleton graph matching (Bai & Latecki, 2008) when skeleton graph is made by comparison of geodesic paths and skeleton endpoints. In this technique, the identification is made based on similarity of the each pair of endpoints and shortest paths.

This study proposes a model for automated identification of eight selected monogenean images and it works by running the commands in MATLAB workspace

which means there is no user interface. As a future work, the Graphical User Interface (GUI) can be deployed as an executable application for ease of use by taxonomists.

Finally, the adaptability and flexibility of the current work presented in this study can be explored for other species (e.g. copepods and otoliths). The integrated model of automated identification of monogenean images successfully combines the range of feature extraction, feature selection and classification techniques. In future works, the success of applying this model for other species can be evaluated.

### 5.6 Conclusions

In this study, a model for identification of monogenean based on shape of anchor and bars is proposed. The dataset consisted of 160 images, discussed in this research and has been successfully used for classification and identification of monogenean, using feature selection and pattern recognition methods. The database contained images of haptoral organs of eight species: *Sinodiplectanotrema malayanus, Diplectanum jaculator, Trianchoratus pahangensis, Trianchoratus lonianchoratus, Trianchoratus malayensis, Metahaliotrema ypsilocleithru, Metahaliotrema mizellei* and *Metahaliotrema similis*. K-Nearest Neighbour and Artificial Neural Network classification techniques were used to perform identification while Linear Discriminant Analysis was selected as a feature selection technique to select feature vector with seven elements from feature space with 24 elements. Segmentation was carried out to separate each organ of bars and anchors from the background and the challenge was overlapping of dorsal and ventral bars and anchors on each other. As a solution to this problem, whole organs were considered as an object while only one anchor was also tested in feature extraction. Two classification techniques for species identification are more reliable as this will prevent lack of confidence in the final results. The highest classification result was achieved by ANN classifier which is 93.1% whereas classification by KNN yielded 86.25% accuracy. Although KNN was less accurate than

ANN, both methods were able to identify selected eight monogenean species with accuracy more than 85%, thus the model developed in this study was successful.

## REFERENCES


Abu, A., Lim, S. L. H., Sidhu, A. S., & Dhillon, S. K. (2013). Biodiversity image retrieval framework for monogeneans. *Systematics and Biodiversity*, *11*(1), 19–33.

Ahmed, F., Kabir, M. H., Bhuyan, S., Bari, H., & Hossain, E. (2014). Automated weed classification with local pattern-based texture descriptors. *The International Arab Journal of Information Technology*, *11*(1), 87–94.

Ali, R., Hussain, A., Bron, J. E., & Shinn, A. P. (2011). Multi-stage classification of *Gyrodactylus* species using machine learning and feature selection techniques. In *2011 11th International Conference on Intelligent Systems Design and Applications*, 457–462

Ali, R., Hussain, A., Bron, J. E., & Shinn, A. P. (2012). The Use of ASM Feature Extraction and Machine Learning for the Discrimination of Members of the Fish Ectoparasite Genus *Gyrodactylus*, *Neural Information Processing, Springer Berlin Heidelberg,* 256–263.

Arbuckle, T., Schröder, S., Steinhage, V., & Wittmann, D. (2001). Biodiversity Informatics in Action: Identification and Monitoring of Bee Species Using ABIS. *Proceedings of Informatics for Environmental Protection,* 1*,* 425-430.

Arce, S. H., Wu, P.-H., & Tseng, Y. (2013). Fast and accurate automated cell boundary determination for fluorescence microscopy. *Scientific Reports*, *3*.

Athitsos, V., Alon, J., & Sclaroff, S. (2005). Efficient nearest neighbor classification using a cascade of approximate similarity measures. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, 1, 486–493.

Athitsos, V., & Sclaroff, S. (2005). Boosting nearest neighbor classifiers for multiclass recognition. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Workshops*, 45–45.

Ault, A., Zhong, X., & Coyle, E. J. (2005). K-nearest-neighbor analysis of received signal strength distance estimation across environments. In *Proceedings of the First Workshop on Wireless Network Measurements*.

Avci, D., & Varol, A. (2009). An expert diagnosis system for classification of human parasite eggs based on multi-class SVM. *Expert Systems with Applications*, *36*(1), 43–48.

Bai, X., & Latecki, L. J. (2008). Path Similarity Skeleton Graph Matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *30*(7), 1282–1292.

Benfield, M., Grosjean, P., Culverhouse, P., Irigolen, X., Sieracki, M., Lopez-Urrutia, A., Dam, H., Hu, Q., Davis, C., Hanson, A., Pilskaln, C., Riseman, E., Schulz, H., Utgoff, P., Gorsky, G. (2007). RAPID: Research on Automated Plankton Identification. *Oceanography*, *20*(2), 172–187.

Boeger, W. A., & Kritsky, D. C. (1993). Phylogeny and a revised classification of the Monogenoidea Bychowsky, 1937 (Platyhelminthes). *Systematic Parasitology*, *26*(1), 1–32.

Bovik, A. C., Huang, T. S., & Munson, D. C. (1987). The effect of median filtering on edge estimation and detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *9*(2), 181–194.

Bradbury, M. H. S. M., & Bracegirdle, D. B. (1998). *Introduction to Light Microscopy* (2 edition). Oxford: Garland Science.

Brooks, D. R., & McLennan, D. A. (1993). Comparative Study of Adaptive Radiations with an Example Using Parasitic Flatworms (Platyhelminthes: Cercomeria). *The American Naturalist*, *142*(5), 755–778.

Bykhovsky, B. E., & Nagibina, L. F. (1978). To the revision of Ancyrocephalidae Bykhovsky, 1937 (Monogenoidea). *Parazitologicheski Sbornik*, *28*, 5–15.

Cai, D., He, X., & Han, J. (2008). SRDA: An Efficient Algorithm for Large-Scale Discriminant Analysis. *IEEE Transactions on Knowledge and Data Engineering*, *20*(1), 1–12.

Canny, J. (1986). A Computational Approach to Edge Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *PAMI-8*(6), 679–698.

Castañón, C. A. B., Fraga, J. S., Fernandez, S., Gruber, A., & da F. Costa, L. (2007). Biological shape characterization for automatic image recognition and diagnosis of protozoan parasites of the genus Eimeria. *Pattern Recognition*, *40*(7), 1899–1910.

Cho, J., Choi, J., Qiao, M., Ji, C. W., & KIM, H. Y. (2008). Automatic Identification of Tobacco Whiteflies, Aphids and Thrips in Greenhouse Using Image Processing Techniques. In 4th *WSEAS International Conference on mathematical biology and ecology*, Mexico, *1*, 46–53.

Choras, R. S. (2007). Image feature extraction techniques and their applications for CBIR and biometrics systems, *1*(1), 6–16.

Coltelli, P., Barsanti, L., Evangelista, V., Frassanito, A. M., & Gualtieri, P. (2014). Water monitoring: automated and real time identification and classification of algae using digital microscopy. *Environmental Science. Processes & Impacts*, *16*(11), 2656–2665.

Culverhouse, P.F., Simpson, R.G., Ellis, R., Lindley, J.A., Williams, R., Parsini, T., Reguera, B., Bravo, I., Zoppoli, R., Earnshaw, G. and McCall, H., (1996). Automatic classification of field-collected dinoflagellates by artificial neural network. *Marine Ecology Progress Series*, *139*, 281–287.

Culverhouse, P. F., Williams, R., Reguera, B., Herry, V., & GonzlezGil, S. (2003). Do experts make mistakes? A comparison of human and machine indentification of dinoflagellates. *Marine Ecology Progress Series*, *247*, 17–25.

Cunningham, P., & Delany, S. J. (2007). K-Nearest Neighbour Classifiers. *Technical Report UCD-CSI*.

Dalal, N., & Triggs, B. (2005). Histograms of Oriented Gradients for Human Detection. In *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), 1,* 886–893.

Di Ruberto, C., Dempster, A., Khan, S., & Jarra, B. (2000). Automatic thresholding of infected blood images using granulometry and regional extrema, *IEEE Comput. Soc, 3*, 441–444.

Dietrich, C. H., & Pooley, C. D. (1994). Automated Identification of Leafhoppers (Homoptera: Cicadellidae: Draeculacephala Ball). *Annals of the Entomological Society of America*, *87*(4), 412–423.

Doncic, A., Eser, U., Atay, O., & Skotheim, J. M. (2013). An Algorithm to Automate Yeast Segmentation and Tracking. *PLoS ONE*, *8*, 57970.

Du Preez, L. H., & Maritz, M. F. (2006). Demonstrating morphometric protocols using polystome marginal hooklet measurements. *Systematic Parasitology, 63*, 1–15.

Duda, R. O., Hart, P. E., & Stork, D. G. (2012). *Pattern Classification*. John Wiley & Sons.

Feng, L., & Bhanu, B. (2013). Automated identification and retrieval of moth images with semantically related visual attributes on the wings. In 20th IEEE International Conference on Image Processing (ICIP), 2577–2581.

Feng, L., Bhanu, B., & Heraty, J. (2016). A software system for automated identification and retrieval of moth images based on wing attributes. *Pattern Recognition*, *51*, 225–241.

Flusser, J., & Suk, T. (1993). Pattern recognition by affine moment invariants. *Pattern Recognition*, *26*(1), 167–174.

Ginoris, Y. P., Amaral, A. L., Nicolau, A., Coelho, M. A. Z., & Ferreira, E. C. (2007). Recognition of protozoa and metazoa using image analysis tools, discriminant analysis, neural networks and decision trees. *Analytica Chimica Acta*, *595*(1–2), 160–169.

Gomez, A., & Salazar, A. (2016). Towards Automatic Wild Animal Monitoring: Identification of Animal Species in Camera-trap Images using Very Deep Convolutional Neural Networks. *arXiv Preprint, arXiv:1603.06169*.

Gonzalez, R. C., & Woods, R. E. (2007). *Digital Image Processing* (3 edition). Upper Saddle River, N.J: Pearson.

Grigorescu, S. E., Petkov, N., & Kruizinga, P. (2002). Comparison of texture features based on Gabor filters. *IEEE Transactions on Image Processing*, *11*(10), 1160–1167.

Gussev, A. V. (1976). Freshwater Indian Monogenoidea. Principles of systematics, analysis of the world faunas and their evolution. *Indian Journal of Helminthology*, *25*, 1–241.

Hanqing, Z., Zuorui, S. (2002). On computer-aided insect identification through math-morphology features. *Journal of China Agricultural University*, *7*, 38-42

Haralick, R. M., Shanmugam, K., & Dinstein, I. (1973). Textural Features for Image Classification. *IEEE Transactions on Systems, Man, and Cybernetics*, *3*(6), 610–621.

Haralick, R. M., & Shapiro, L. G. (1992). *Computer and Robot Vision* (1st ed.). Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc.

Hernández-Serna, A., & Jiménez-Segura, L. F. (2014). Automatic identification of species with neural networks. *PeerJ*, *2*, e563.

Holmes, C. C., & Adams, N. M. (2002). A probabilistic nearest neighbour method for statistical pattern recognition. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *64*(2), 295–306.

Huddar, S. R., Gowri, S., Keerthana, K., Vasanthi, S., & Rupanagudi, S. R. (2012). Novel algorithm for segmentation and automatic identification of pests on plants using image processing. In *third international conference on computing communication & networking technologies (ICCCNT), IEEE,* 1–5.

Image Processing Toolbox - MATLAB. (n.d.). Retrieved November 23, 2016, from https://www.mathworks.com/products/image/index.html

Islam, M. M., Dengsheng Zhang, & Guojun Lu. (2008). A geometric method to compute directionality features for texture images. In *IEEE international conference of multimedia and expo.* 1521–1524.

Jain, A. K., Duin, P. W., & Jianchang Mao. (2000). Statistical pattern recognition: a review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *22*(1), 4–37.

Jain, A., & Zongker, D. (1997). Feature selection: Evaluation, application, and small sample performance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *19*(2), 153–158.

Jalba, A. C., Wilkinson, M. H. F., Roerdink, J. B. T. M., Bayer, M. M., & Juggins, S. (2005). Automatic diatom identification using contour analysis by morphological curvature scale spaces. *Machine Vision and Applications*, *16*(4), 217–228.

Jin, T., Hou, X., Li, P., & Zhou, F. (2015). A Novel Method of Automatic Plant Species Identification Using Sparse Representation of Leaf Tooth Features. *PloS One*, *10*(10), e0139482.

Jolliffe, I. (2002). *Principal component analysis*. Wiley Online Library.

Jonker, R., Groben, R., Tarran, G., Medlin, L., Wilkins, M., García, L., Zabala, L., Boddy, L. (2000). Automated identification and characterisation of microbial populations using flow cytometry: the AIMS project. *Scientia Marina*, *64*(2), 225–234.

Kalafi, E. Y., Tan, B. W., Town, C., Dhillon, S. K. (2016). Automated identification of Monogeneans using digital image processing and Knearest neighbour approaches. *BMC Bioinformatics, 17(Suppl 19):511,* 259-266.

Kang, W.-X., Yang, Q.-Q., & Liang, R.-P. (2009). The Comparative Research on Image Segmentation Algorithms. In *first international workshop on education technology and computer science, IEEE*. 703–707

Karagiannopoulos, M., Anyfantis, D., Kotsiantis, S. B., & Pintelas, P. E. (2007). Feature selection for regression problems. In *Proceedings of the 8th Hellenic European Research on Computer Mathematics & its Applications, Athens, Greece*, *2022*.

Kay, J. W., Shinn, A. P., & Sommerville, C. (1999). Towards an automated system for the identification of notifiable pathogens: using Gyrodactylus salaris as an example. *Parasitology Today*, *15*(5), 201–206.

Kaya, Y., Kayci, L., & Uyar, M. (2015). Automatic identification of butterfly species based on local binary patterns and artificial neural network. *Applied Soft Computing*, *28*, 132–137.

Kearn, G. C. (1994). Evolutionary expansion of the Monogenea. *International Journal for Parasitology*, *24*(8), 1227–1271.

Khan, W. (2014). Image Segmentation Techniques: A Survey. *Journal of Image and Graphics*, 166–170.

Khang, T. F., Soo, O. Y. M., Tan, W. B., & Lim, L. H. S. (2016). Monogenean anchor morphometry: systematic value, phylogenetic signal, and evolution. *PeerJ*, *4*, e1668.

Kiranyaz, S., Ince, T., Pulkkinen, J., Gabbouj, M., Ärje, J., Kärkkäinen, S., Tirronen, V., Juhola, M., Turpeinen, T., Meissner, K. (2011). Classification and retrieval on macroinvertebrate image databases. *Computers in Biology and Medicine*, *41*(7), 463–472.

Kudo, M., & Sklansky, J. (2000). Comparison of algorithms that select features for pattern classifiers. *Pattern Recognition*, *33*(1), 25–41.

Larios, N., Deng, H., Zhang, W., Sarpola, M., Yuen, J., Paasch, R., Moldenke, A., Lytle, D., A., Correa, S., R., Mortensen, E., N., Shapiro, L., G., Dietterich, T. G. (2008). Automated insect identification through concatenated histograms of local appearance features: feature vector generation and region detection for deformable objects. *Machine Vision and Applications*, *19*(2), 105–123.

Latourrette, M. (2000). Toward an Explanatory Similarity Measure for Nearest-Neighbor Classification. In R. L. de Mántaras & E. Plaza (Eds.), *Machine Learning: ECML 2000*, 238–245.

Lei, Z., Liao, S., & Li, S. Z. (2012). Efficient feature selection for linear discriminant analysis and its application to face recognition. *In 21st International Conference on Pattern Recognition (ICPR), IEEE.* 1136–1139

Leow, L. K., Chew, L.-L., Chong, V. C., & Dhillon, S. K. (2015). Automated identification of copepods using digital image processing and artificial neural network. *BMC Bioinformatics*, *16*(18), 1.

Le-Qing, Z., & Zhen, Z. (2012). Automatic insect classification based on local mean colour feature and Supported Vector Machines. *Oriental Insects*, *46*(3–4), 260–269.

Li, J., Tseng, K.-K., Hsieh, Z. Y., Yang, C. W., & Huang, H.-N. (2014). Staining Pattern Classification of Antinuclear Autoantibodies Based on Block Segmentation in Indirect Immunofluorescence Images. *PloS One*, *9*(12), e113132.

Lim, L. H. S., & Gibson, D. I. (2009). A new monogenean genus from an ephippid fish off Peninsular Malaysia. *Systematic Parasitology*, *73*(1), 13–25.

Lim, L. H. S., & Gibson, D. I. (2010). Taxonomy, taxonomists & biodiversity. *Sarawak Biodiversity Centre*, 33–43.

Lim, L. H. S., Tan, W. B., & Gibson, D. I. (2010). Description of Sinodiplectanotrema malayanum n. sp.(Monogenea: Diplectanidae), with comments on the taxonomic position of the genus. *Systematic Parasitology*, *76*(2), 145–157.

Liu, F., Shen, Z. R., Zhang, J. W., & Yang, H. Z. (2008). Automatic insect identification based on color characters. *Chinese Bulletin of Entomology*, *45*(1), 150–153.

Loke, K. S., Egerton, S., Cristofaro, D., & Clementson, S. (2011). Automated real-time dynamic identification of flying and resting butterfly species in the natural environment. In *2011 International Conference on Environment Science and Engineering (ICESE 2011)*. 179–183).

Loncaric, S. (1998). A survey of shape analysis techniques. *Pattern Recognition*, *31*(8), 983–1001.

Lowe, D., & Broomhead, D. (1988). Multivariable functional interpolation and adaptive networks. *Complex Systems*, *2*, 321–355.

Lowe, D. G. (2004). Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, *60*(2), 91–110.

Lu, A., Hou, X., Lin, C., & Liu, C.-L. (2010). Insect Species Recognition using Sparse Representation. *British Machine Vision Association.* 108.1-108.10

Luo, Q., Gao, Y., Luo, J., Chen, C., Liang, J., & Yang, C. (2011). Automatic Identification of Diatoms with Circular Shape using Texture Analysis. *Journal of Software*, *6*(3).

Maglogiannis, I. G. (2007). *Emerging Artificial Intelligence Applications in Computer Engineering: Real Word AI Systems with Applications in EHealth, HCI, Information Retrieval and Pervasive Technologies*. IOS Press.

Mansoor, H., Sorayya, M., Aishah, S., Mogeeb, A., & Mosleh, A. (2011). Automatic recognition system for some cyanobacteria using image processing techniques and ANN approach. In *International Proceedings of Chemical, Biological and Environmental Engineering*, *19*, 73–78.

Martins, J., Oliveira, L. S., Nisgoski, S., & Sabourin, R. (2013). A database for automatic classification of forest species. *Machine Vision and Applications*, *24*(3), 567–578.

Mayo, M., & Watson, A. T. (2007). Automatic species identification of live moths. *Knowledge-Based Systems*, *20*(2), 195–202.

McClelland, J. L., Rumelhart, D. E., Group, P. R., & others. (1987). *Parallel distributed processing*. MIT Press Cambridge, MA.

Miller, N. A., Gregory, J. S., Aspden, R. M., Stollery, P. J., & Gilbert, F. J. (2014). Using Active Shape Modeling Based on MRI to Study Morphologic and Pitch-Related Functional Changes Affecting Vocal Structures and the Airway. *Journal of Voice*, *28*(5), 554–564.

Ming-Kuei Hu. (1962). Visual pattern recognition by moment invariants. *IEEE Transactions on Information Theory*, *8*(2), 179–187.

Moyo, T., Bangay, S., & Foster, G. (2006). The identification of mammalian species through the classification of hair patterns using image pattern recognition. In *Proceedings of the 4th international conference on Computer graphics, virtual reality, visualisation and interaction in Africa*. 177–181

Murith, D., & Beverley-Burton, M. (1985). Salsuginus Beverley-Burton, 1984 (Monogenea: Ancyrocephalidae) from Cyprinodontoidei (Atheriniformes) in North America with description of Salsuginus angularis (Mueller, 1943) Beverley-Burton, 1984 from Fundulus diaphanus and Salsuginus heterocliti n. sp. from F. heteroclitus. *Canadian Journal of Zoology, 63,* 703–714.

Mythili, C., & Kavitha, V. (2011). Efficient technique for color image noise reduction. *The Research Bulletin of Jordan, ACM*, *1*(11), 41–44.

Natchimuthu, S., Natchimuthu, S., Chinnaraj, P., Parthasarathy, S., & Senthil, K. (2013). Automatic Identification of Algal Community from Microscopic Images. *Bioinformatics and Biology Insights*, 327.

Ojala, T., Pietikainen, M., & Maenpaa, T. (2002). Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *24*(7), 971–987.

O'Neill, M. A., Gauld, I. D., Gaston, K. J., & Weeks, P. J. D. (2000). Daisy: an automated invertebrate identification system using holistic vision techniques. In *Proceedings of the Inaugural Meeting BioNET-INTERNATIONAL Group for Computer-Aided Taxonomy (BIGCAT)*, 13–22.

Pariselle, A., Boeger, W. A., Snoeks, J., Bilong Bilong, C. F., Morand, S., & Vanhove, M. P. M. (2011). The Monogenean Parasite Fauna of Cichlids: A Potential Tool for Host Biogeography. *International Journal of Evolutionary Biology*, e471480.

Parisi-Baradad, V., Manjabacas, A., Lombarte, A., Olivella, R., Chic, ò., Piera, J., & García-Ladona, E. (2010). Automated Taxon Identification of Teleost fishes using an otolith online database—AFORO. *Fisheries Research*, *105*(1), 13–20.

Peng, J., Heisterkamp, D. R., & Dai, H. K. (2001). LDA/SVM driven nearest neighbor classification. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001, 1,* I-58-I-63.

Perre, P., Faria, F. A., Jorge, L. R., Rocha, A., Torres, R. S., Souza-Filho, M. F., Lewinsohn, T. M., Zucchi, R. A. (2016). Toward an Automated Identification of Anastrepha Fruit Flies in the fraterculus group (Diptera, Tephritidae). *Neotropical Entomology*, 1–5.

Ping Tian, D. (2013). A review on image feature extraction and representation techniques. *International Journal of Multimedia and Ubiquitous Engineering*, *8*(4), 385–396.

Poisot, T., & Desdevises, Y. (2010). Putative speciation events in *Lamellodiscus* (Monogenea: Diplectanidae) assessed by a morphometric approach. *Biological Journal of the Linnean Society*, *99*(3), 559–569.

Priddy, K. L., & Keller, P. E. (2005). *Artificial Neural Networks: An Introduction*. SPIE Press.

Quinlan, J. R. (2014). *C4. 5: programs for machine learning*. Elsevier.

Quivy, C.-H., & Kumazawa, I. (2011). Normalization of Active Appearance Models for Fish Species Identification. *ISRN Signal Processing*, *2011*, 1–16.

Ranzato, M., Taylor, P. E., House, J. M., Flagan, R. C., LeCun, Y., & Perona, P. (2007). Automatic recognition of biological particles in microscopic images. *Pattern Recognition Letters*, *28*(1), 31–39.

Rehulkova, E., & Gelnar, M. (2005). A revised diagnosis of Thylacicleidus (Monogenea: Dactylogyridae) with a redescription of the type species, Thylacicleidus serendipitus, and description of two new species from southeast Asian pufferfishes (Tetraodontiformes: Tetraodontidae). *Journal of Parasitology, 91,* 794–807.

Riggs, L. A. (1973). Curvature as a Feature of Pattern Vision. *Science*, *181*(4104),

Rodríguez-González, A., Míguez-Lozano, R., Llopis-Belenguer, C., & Balbuena, J. A. (2015). Phenotypic plasticity in haptoral structures of Ligophorus cephali (Monogenea: Dactylogyridae) on the flathead mullet (Mugil cephalus): a geometric morphometric approach. *International Journal for Parasitology*, *45*(5), 295–303.

Rohlf, F. J., & Marcus, L. F. (1993). A revolution in morphometrics. Trends in Ecology and Evolution, *8*, 129–132.

Rother, C., Kolmogorov, V., & Blake, A. (2004). Grabcut: Interactive foreground extraction using iterated graph cuts. In *ACM transactions on graphics (TOG)*, *23*, 309–314.

Russell, K. N., Do, M. T., Huff, J. C., Platnick, N. I., & MacLeod, N. (2007). Introducing SPIDA-web: wavelets, neural networks and Internet accessibility in an image-based automated identification system. *Automated Taxon Identification in Systematics: Theory, Approaches and Applications*, 131–152.

Salimi, N., Loh, K. H., Dhillon, S. K., & Chong, V. C. (2016). Fully-automated identification of fish species based on otolith contour: using short-time Fourier transform and discriminant analysis (STFT-DA). *PeerJ*, *4*, e1664.

Sang-Hee, L. (2010). A Novel Approach to Shape Recognition Using Shape Outline. *Journal of the Korean Physical Society*, *56*(31), 1016.

Saraswat, M., & Arya, K. V. (2014). Automated microscopic image analysis for leukocytes identification: A survey. *Micron*, *65*, 20–33.

Savkare, S. S., & Narote, S. P. (2011). Automatic detection of malaria parasites for estimating parasitemia. *International Journal of Computer Science and Security (IJCSS)*, *5*(3), 310.

Savkare, S. S., & Narote, S. P. (2012). Automatic System for Classification of Erythrocytes Infected with Malaria and Identification of Parasite's Life Stage. *Procedia Technology*, *6*, 405–410.

Savkare, S. S., & Narote, S. P. (2015). Automated system for malaria parasite identification. In *Proceedings of the International Conference on Communication, Information & Computing Technology (ICCICT), IEEE.* 1–4

Schroder, S., Wittmann, D., Drescher, W., Roth, V., Steinhage, V., & Cremers, A. B. (2002). The new key to bees: automated identification by image analysis of wings. *Pollinating Bees–the Conservation Link Between Agriculture and Nature. Ministry of Environment: Brasilia*.

Shih, T. K., Huang, J.-Y., Wang, C.-S., Hung, J. C., & Kao, C.-H. (2001). An intelligent content-based image retrieval system based on color, shape and spatial relations. In *proceedings-national science council republic of china part a physical science and engineering, 25*, 232–243.

Shinn, A. P., des Clers, S., Gibson, D. I., & Sommerville, C. (1996). Multivariate analyses of morphometrical features from Gyrodactylus spp. (Monogenea) parasitising British salmonids: Light microscope based studies. *Systematic Parasitology, 33*, 115–125.

Shinn, A. P., Gibson, D. I., & Sommerville, C. (2001). Morphometric discrimination of *Gyrodactylus salaris Malmberg* (Monogenea) from species of *Gyrodactylus* parasitising British salmonids using novel parameters. *Journal of Fish Diseases*, *24*(2), 83–97.

Singh, H. K., Tomar, S. K., & Maurya, P. K. (2012). Thresholding Techniques applied for Segmentation of RGB and multispectral images. *Proceedings Published by International Journal of Computer Applications®(IJCA) ISSN*, 975–8887.

Soleymanpour, E., Rajae, B., & Pourreza, H. R. (2010). Offline handwritten signature identification and verification using contourlet transform and Support Vector Machine. In *2010 6th Iranian Conference on Machine Vision and Image Processing*. 1–6.

Song, F., Mei, D., & Li, H. (2010). Feature selection based on linear discriminant analysis. In *International Conference of Intelligent System Design and Engineering Application (ISDEA)*, *1*, 746–749.

Song, Y., Huang, J., Zhou, D., Zha, H., & Giles, C. L. (2007). IKNN: Informative K-Nearest Neighbor Pattern Classification. In *Proceedings of the 11th European conference on principles and practice of knowledge discovery in databases (PKDD) 2007, Springer, Berlin, Heidelberg,* 248–264.

Strauss, R. E., & Bookstein, F. L. (1982). The truss: body form reconstruction in morphometrics. *Systematic Zoology*, *31*, 113–135.

Strona, G., Montano, S., Seveso, D., Galli, P., & Fattorini, S. (2014). Identification of Monogenea made easier: a new statistical procedure for an automatic selection of diagnostic linear measurements in closely related species. *Journal of Zoological Systematics and Evolutionary Research*, *52*(2), 95–99.

Tan, W. B. (2013). *Morphological and molecular characterisation of monogeneans / Tan Wooi Boon* (PhD thesis). University of Malaya.

Thiel, S. U., Wiltshire, R. J., & Davies, L. J. (1996). Automated object recognition of blue-green algae for measuring water quality-a preliminary study. *Oceanographic Literature Review*, *1*(43), 85.

Trattner, S., Greenspan, H., Tepper, G., & Abboud, S. (2004). Automatic Identification of Bacterial Types Using Statistical Imaging Methods. *IEEE Transactions on Medical Imaging*, *23*(7), 807–820.

Ververidis, D., & Kotropoulos, C. (2008). Fast and accurate sequential floating forward feature selection with the Bayes classifier applied to speech emotion recognition. *Signal Processing*, *88*(12), 2956–2970.

Vignon, M. (2011a). Putting in shape – towards a unified approach for the taxonomic description of monogenean haptoral hard parts. *Systematic Parasitology*, *79*(3), 161–174.

Vignon, M. (2011b). Inference in morphological taxonomy using collinear data and small sample sizes: Monogenean sclerites (Platyhelminthes) as a case study. *Zoologica Scripta*, *40*(3), 306–316.

Vogt, A., Cholewinski, A., Shen, X., Nelson, S. G., Lazo, J. S., Tsang, M., & Hukriede, N. A. (2009). Automated image-based phenotypic analysis in zebrafish embryos. *Developmental Dynamics*, *238*(3), 656–663.

Wang, J., Lin, C., Ji, L., & Liang, A. (2012). A new automatic identification system of insect images at the order level. *Knowledge-Based Systems*, *33*, 102–110.

Weeks, P. J. ., O'Neill, M. ., Gaston, K. ., & Gauld, I. . (1999). Species–identification of wasps using principal component associative memories. *Image and Vision Computing*, *17*(12), 861–866.

Wen, C., Guyer, D. E., & Li, W. (2009). Local feature-based identification and classification for orchard insects. *Biosystems Engineering*, *104*(3), 299–307.

Whittington, I. D. (1998). Diversity "down under": monogeneans in the Antipodes (Australia) with a prediction of monogenean biodiversity worldwide. *International Journal for Parasitology*, *28*(10), 1481–1493.

Witten, I. H., & Frank, E. (2005). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.

Woo, P. T. K., & Leatherland, J. F. (2006). *Fish Diseases and Disorders*. CABI.

Yang, M., Kpalma, K., & Ronsin, J. (2008). A survey of shape feature extraction techniques. *Pattern Recognition*, 43–90.

Yang, Y. S., Park, D. K., Kim, H. C., Choi, M.-H., & Chai, J.-Y. (2001). Automatic identification of human helminth eggs on microscopic fecal specimens using digital image processing and an artificial neural network. *IEEE Transactions on Biomedical Engineering*, *48*(6), 718–730.

Yazdani, A., Ebrahimi, T., & Hoffmann, U. (2009). Classification of EEG signals using Dempster Shafer theory and a k-nearest neighbor classifier. In *4th International IEEE/EMBS Conference on Neural Engineering,* 327–330.

Yuan, G., Hasler, N., Klette, R., & Rosenhahn, B. (2006). *Understanding Tracks of Different Species of Rats*. CITR, The University of Auckland, New Zealand.

Zhan, M., Crane, M. M., Entchev, E. V., Caballero, A., de Abreu, D. A. F., Ch'ng, Q., & Lu, H. (2015). Automated processing of imaging data through multi-tiered classification of biological structures illustrated using *Caenorhabditis elegans*. *PLOS Computational Biology*, *11*(4), e1004194.

Zhang, D., & Lu, G. (2004). Review of shape representation and description techniques. *Pattern Recognition*, *37*(1), 1–19.

## LIST OF PUBLICATIONS AND PAPERS PRESENTED

### Published papers

**Yousef Kalafi, E.**, Tan, W. B., Town, C., & Dhillon, S. K. (2016). Automated identification of Monogeneans using digital image processing and K-nearest neighbour approaches. *BMC Bioinformatics*, *17*, 1376.

Leong, Y.M., **Yousef Kalafi, E.**, Lim, L.H.S, Dhillon, S. K. (2016). Automated Identification of Moth Species. Proc. of the International Conference on Advances in Bio-Informatics and Environmental Engineering - ICABEE 2016. ISBN: 978-1-63248-100-9.

## APPENDIX A

Leica DFC320 camera specifications

| **Digital camera** | **Leica DFC320 (R2)** |
|---|---|
| Camera type | Digital camera for microscopy with control software |
| Sensor | Interline transfer frame readout CCD – ICX252AQ |
| Sensor Grade/Size | Grade Zero / 8.10mm × 6.64mm, Diagonal 8.93mm (Type 1/1.8) |
| Color filter | RGB Bayer mosaic |
| Protective color filter | Hoya CM500S (IR cut-off 650nm) |
| Shutter control | Electronic global shutter/interlaced readout |
| Number of pixels | 3.3 Mpixel, 2088 × 1550 |
| Max scaled resolution (PC only) | 7.3 Mpixel, 3132 × 2325 |
| Sensitive area | 7.2 mm × 5.35 mm |
| Pixel size | 3.45 µm × 3.45 µm |
| Color depth | 36 Bit |
| A/D converter | 12 Bit |
| Dynamic range | > 59 dB |
| Readout noise | s < 5.0 LSB (12 Bit) typical |
| Exposure time | 230 µsec - 60 sec |
| Dark current | 1.2 LSB/sec at 12 Bit typical |
| Quantum efficiency | Relative: Blue 465nm 98%; Green 530nm 100%; Red 610nm 94% |
| Gain control/Offset control | 10× / 0.. 255 LSB (12 Bit) |
| Live image | On computer screen |
| Shading correction | Yes, stored for all formats |
| Brightness correction | On all color binning modes |

# APPENDIX B

MATLAB codes

```
close all
clear
clc
%%%%%%%%%%%%%%%%%%%%%%%
readImsTrain

for i=1 : length(A)
im=mat2gray(A{i});
mIM=imfilter(im,fspecial('average',20),'replicate');
sIM=mIM-im-0.02;
bw0=im2bw(sIM,0);
bw=bwareaopen(bw0,1000);
bw1= imfill(bw, 'holes');
cf = contour_following(bw);
img = zeros(1000,1000);
img(sub2ind(size(img), cf(:,2), cf(:,1))) = 1;
bw= imfill(img, 'holes');
CHI = imEuler1d(bw);
perim = imPerimeter(bw);
area = imArea(bw);
DENSITY = imAreaDensity(bw);
Pv = imPerimeterDensity(bw);
OBB = imOrientedBox(bw);
CHI2 = imEuler1d(bw1);
perim2 = imPerimeter(bw1);
area2 = imArea(bw1);
DENSITY2 = imAreaDensity(bw1);
Pv2 = imPerimeterDensity(bw1);
OBB2 = imOrientedBox(bw1);
ent1=entropy(bw);
ent2=entropy(bw1);
stats1 = regionprops(bw1,'MajorAxisLength');
stats = regionprops(bw,'MajorAxisLength');
CA=stats1.MajorAxisLength;
CA2=stats.MajorAxisLength;
nBlack = sum(bw1(:));
nWhite = numel(bw1) - nBlack;
a1{i}=CHI;
a2{i}=perim;
a3{i}=area;
a4{i}=DENSITY;
a5{i}=Pv;
a6{i}=OBB;
a7{i}=CHI2;
a8{i}=perim2;
a9{i}=area2;
a10{i}=DENSITY2;
a11{i}=Pv2;
a12{i}=OBB2;
a13{i}=ent1;
a14{i}=ent2;
a15{i}=CA;
a16{i}=nWhite;
Afeat{i}=[a1(i) a2(i) a3(i) a4(i) a5(i) a6(i)...
          a7(i) a8(i) a9(i) a10(i) a11(i) a12(i)...
          a13(i) a14(i) a15(i) a16(i)];
end
Afeat=padcat(Afeat{:});
Afeat=cell2mat(Afeat);


for i=1 : length(B)
im=mat2gray(B{i});
mIM=imfilter(im,fspecial('average',20),'replicate');
sIM=mIM-im-0.02;
bw0=im2bw(sIM,0);
bw=bwareaopen(bw0,1000);
bw1= imfill(bw, 'holes');
cf = contour_following(bw);
img = zeros(1000,1000);
img(sub2ind(size(img), cf(:,2), cf(:,1))) = 1;
bw= imfill(img, 'holes');
CHI = imEuler1d(bw);
perim = imPerimeter(bw);
area = imArea(bw);
DENSITY = imAreaDensity(bw);
Pv = imPerimeterDensity(bw);
OBB = imOrientedBox(bw);
CHI2 = imEuler1d(bw1);
perim2 = imPerimeter(bw1);
```

```
area2 = imArea(bw1);
DENSITY2 = imAreaDensity(bw1);
Pv2 = imPerimeterDensity(bw1);
OBB2 = imOrientedBox(bw1);
ent1=entropy(bw);
ent2=entropy(bw1);
stats1 = regionprops(bw1,'MajorAxisLength');
stats = regionprops(bw,'MajorAxisLength');
CA=stats1.MajorAxisLength;
CA2=stats.MajorAxisLength;
nBlack = sum(bw1(:));
nWhite = numel(bw1) - nBlack;
b1{i}=CHI;
b2{i}=perim;
b3{i}=area;
b4{i}=DENSITY;
b5{i}=Pv;
b6{i}=OBB;
b7{i}=CHI2;
b8{i}=perim2;
b9{i}=area2;
b10{i}=DENSITY2;
b11{i}=Pv2;
b12{i}=OBB2;
b13{i}=ent1;
b14{i}=ent2;
b15{i}=CA;
b16{i}=nWhite;
Bfeat{i}=[b1(i) b2(i) b3(i) b4(i) b5(i) b6(i)...
          b7(i) b8(i) b9(i) b10(i) b11(i) b12(i)...
          b13(i) b14(i) b15(i) b16(i)];
end
Bfeat=padcat(Bfeat{:});
Bfeat=cell2mat(Bfeat);


for i=1 : length(C)
im=mat2gray(C{i});
mIM=imfilter(im,fspecial('average',20),'replicate');
sIM=mIM-im-0.02;
bw0=im2bw(sIM,0);
bw=bwareaopen(bw0,1000);
bw1= imfill(bw, 'holes');
cf = contour_following(bw);
img = zeros(1000,1000);
img(sub2ind(size(img), cf(:,2), cf(:,1))) = 1;
bw= imfill(img, 'holes');
CHI = imEuler1d(bw);
perim = imPerimeter(bw);
area = imArea(bw);
DENSITY = imAreaDensity(bw);
Pv = imPerimeterDensity(bw);
OBB = imOrientedBox(bw);
CHI2 = imEuler1d(bw1);
perim2 = imPerimeter(bw1);
area2 = imArea(bw1);
DENSITY2 = imAreaDensity(bw1);
Pv2 = imPerimeterDensity(bw1);
OBB2 = imOrientedBox(bw1);
ent1=entropy(bw);
ent2=entropy(bw1);
stats1 = regionprops(bw1,'MajorAxisLength');
stats = regionprops(bw,'MajorAxisLength');
CA=stats1.MajorAxisLength;
CA2=stats.MajorAxisLength;
nBlack = sum(bw1(:));
nWhite = numel(bw1) - nBlack;
c1{i}=CHI;
c2{i}=perim;
c3{i}=area;
c4{i}=DENSITY;
c5{i}=Pv;
c6{i}=OBB;
c7{i}=CHI2;
c8{i}=perim2;
c9{i}=area2;
c10{i}=DENSITY2;
c11{i}=Pv2;
c12{i}=OBB2;
c13{i}=ent1;
```

```
c14{i}=ent2;
c15{i}=CA;
c16{i}=nWhite;
Cfeat{i}=[c1(i) c2(i) c3(i) c4(i) c5(i) c6(i)...
          c7(i) c8(i) c9(i) c10(i) c11(i) c12(i)...
          c13(i) c14(i) c15(i) c16(i)];
end
Cfeat=padcat(Cfeat{:});
Cfeat=cell2mat(Cfeat);

for i=1 : length(D)
im=mat2gray(D{i});
mIM=imfilter(im,fspecial('average',20),'replicate');
sIM=mIM-im-0.02;
bw0=im2bw(sIM,0);
bw=bwareaopen(bw0,1000);
bw1= imfill(bw, 'holes');
cf = contour_following(bw);
img = zeros(1000,1000);
img(sub2ind(size(img), cf(:,2), cf(:,1))) = 1;
bw= imfill(img, 'holes');
CHI = imEuler1d(bw);
perim = imPerimeter(bw);
area = imArea(bw);
DENSITY = imAreaDensity(bw);
Pv = imPerimeterDensity(bw);
OBB = imOrientedBox(bw);
CHI2 = imEuler1d(bw1);
perim2 = imPerimeter(bw1);
area2 = imArea(bw1);
DENSITY2 = imAreaDensity(bw1);
Pv2 = imPerimeterDensity(bw1);
OBB2 = imOrientedBox(bw1);
ent1=entropy(bw);
ent2=entropy(bw1);
stats1 = regionprops(bw1,'MajorAxisLength');
stats = regionprops(bw,'MajorAxisLength');
CA=stats1.MajorAxisLength;
CA2=stats.MajorAxisLength;
nBlack = sum(bw1(:));
nWhite = numel(bw1) - nBlack;
d1{i}=CHI;
d2{i}=perim;
d3{i}=area;
d4{i}=DENSITY;
d5{i}=Pv;
d6{i}=OBB;
d7{i}=CHI2;
d8{i}=perim2;
d9{i}=area2;
d10{i}=DENSITY2;
d11{i}=Pv2;
d12{i}=OBB2;
d13{i}=ent1;
d14{i}=ent2;
d15{i}=CA;
d16{i}=nWhite;
Dfeat{i}=[d1(i) d2(i) d3(i) d4(i) d5(i) d6(i)...
          d7(i) d8(i) d9(i) d10(i) d11(i) d12(i)...
          d13(i) d14(i) d15(i) d16(i)];
end
Dfeat=padcat(Dfeat{:});
Dfeat=cell2mat(Dfeat);


for i=1 : length(E)
im=mat2gray(E{i});
mIM=imfilter(im,fspecial('average',20),'replicate');
sIM=mIM-im-0.02;
bw0=im2bw(sIM,0);
bw=bwareaopen(bw0,1000);
bw1= imfill(bw, 'holes');
cf = contour_following(bw);
img = zeros(1000,1000);
img(sub2ind(size(img), cf(:,2), cf(:,1))) = 1;
bw= imfill(img, 'holes');
CHI = imEuler1d(bw);
perim = imPerimeter(bw);
area = imArea(bw);
DENSITY = imAreaDensity(bw);
```

```
Pv = imPerimeterDensity(bw);
OBB = imOrientedBox(bw);
CHI2 = imEuler1d(bw1);
perim2 = imPerimeter(bw1);
area2 = imArea(bw1);
DENSITY2 = imAreaDensity(bw1);
Pv2 = imPerimeterDensity(bw1);
OBB2 = imOrientedBox(bw1);
ent1=entropy(bw);
ent2=entropy(bw1);
stats1 = regionprops(bw1,'MajorAxisLength');
stats = regionprops(bw,'MajorAxisLength');
CA=stats1.MajorAxisLength;
CA2=stats.MajorAxisLength;
nBlack = sum(bw1(:));
nWhite = numel(bw1) - nBlack;
e1{i}=CHI;
e2{i}=perim;
e3{i}=area;
e4{i}=DENSITY;
e5{i}=Pv;
e6{i}=OBB;
e7{i}=CHI2;
e8{i}=perim2;
e9{i}=area2;
e10{i}=DENSITY2;
e11{i}=Pv2;
e12{i}=OBB2;
e13{i}=ent1;
e14{i}=ent2;
e15{i}=CA;
e16{i}=nWhite;
Efeat{i}=[e1(i) e2(i) e3(i) e4(i) e5(i) e6(i)...
          e7(i) e8(i) e9(i) e10(i) e11(i) e12(i)...
          e13(i) e14(i) e15(i) e16(i)];
end
Efeat=padcat(Efeat{:});
Efeat=cell2mat(Efeat);


for i=1 : length(F)
im=mat2gray(F{i});
mIM=imfilter(im,fspecial('average',20),'replicate');
sIM=mIM-im-0.02;
bw0=im2bw(sIM,0);
bw=bwareaopen(bw0,1000);
bw1= imfill(bw, 'holes');
cf = contour_following(bw);
img = zeros(1000,1000);
img(sub2ind(size(img), cf(:,2), cf(:,1))) = 1;
bw= imfill(img, 'holes');
CHI = imEuler1d(bw);
perim = imPerimeter(bw);
area = imArea(bw);
DENSITY = imAreaDensity(bw);
Pv = imPerimeterDensity(bw);
OBB = imOrientedBox(bw);
CHI2 = imEuler1d(bw1);
perim2 = imPerimeter(bw1);
area2 = imArea(bw1);
DENSITY2 = imAreaDensity(bw1);
Pv2 = imPerimeterDensity(bw1);
OBB2 = imOrientedBox(bw1);
ent1=entropy(bw);
ent2=entropy(bw1);
stats1 = regionprops(bw1,'MajorAxisLength');
stats = regionprops(bw,'MajorAxisLength');
CA=stats1.MajorAxisLength;
CA2=stats.MajorAxisLength;
nBlack = sum(bw1(:));
nWhite = numel(bw1) - nBlack;
f1{i}=CHI;
f2{i}=perim;
f3{i}=area;
f4{i}=DENSITY;
f5{i}=Pv;
f6{i}=OBB;
f7{i}=CHI2;
f8{i}=perim2;
f9{i}=area2;
```

```
f10{i}=DENSITY2;
f11{i}=Pv2;
f12{i}=OBB2;
f13{i}=ent1;
f14{i}=ent2;
f15{i}=CA;
f16{i}=nWhite;
Ffeat{i}=[f1(i) f2(i) f3(i) f4(i) f5(i) f6(i)...
          f7(i) f8(i) f9(i) f10(i) f11(i) f12(i)...
          f13(i) f14(i) f15(i) f16(i)];
end
Ffeat=padcat(Ffeat{:});
Ffeat=cell2mat(Ffeat);

for i=1 : length(G)
im=mat2gray(G{i});
mIM=imfilter(im,fspecial('average',20),'replicate');
sIM=mIM-im-0.02;
bw0=im2bw(sIM,0);
bw=bwareaopen(bw0,1000);
bw1= imfill(bw, 'holes');
cf = contour_following(bw);
img = zeros(1000,1000);
img(sub2ind(size(img), cf(:,2), cf(:,1))) = 1;
bw= imfill(img, 'holes');
CHI = imEuler1d(bw);
perim = imPerimeter(bw);
area = imArea(bw);
DENSITY = imAreaDensity(bw);
Pv = imPerimeterDensity(bw);
OBB = imOrientedBox(bw);
CHI2 = imEuler1d(bw1);
perim2 = imPerimeter(bw1);
area2 = imArea(bw1);
DENSITY2 = imAreaDensity(bw1);
Pv2 = imPerimeterDensity(bw1);
OBB2 = imOrientedBox(bw1);
ent1=entropy(bw);
ent2=entropy(bw1);
stats1 = regionprops(bw1,'MajorAxisLength');
stats = regionprops(bw,'MajorAxisLength');
CA=stats1.MajorAxisLength;
CA2=stats.MajorAxisLength;
nBlack = sum(bw1(:));
nWhite = numel(bw1) - nBlack;
g1{i}=CHI;
g2{i}=perim;
g3{i}=area;
g4{i}=DENSITY;
g5{i}=Pv;
g6{i}=OBB;
g7{i}=CHI2;
g8{i}=perim2;
g9{i}=area2;
g10{i}=DENSITY2;
g11{i}=Pv2;
g12{i}=OBB2;
g13{i}=ent1;
g14{i}=ent2;
g15{i}=CA;
g16{i}=nWhite;
Gfeat{i}=[g1(i) g2(i) g3(i) g4(i) g5(i) g6(i)...
          g7(i) g8(i) g9(i) g10(i) g11(i) g12(i)...
          g13(i) g14(i) g15(i) g16(i)];
end
Gfeat=padcat(Gfeat{:});
Gfeat=cell2mat(Gfeat);

for i=1 : length(H)
im=mat2gray(H{i});
mIM=imfilter(im,fspecial('average',20),'replicate');
sIM=mIM-im-0.02;
bw0=im2bw(sIM,0);
bw=bwareaopen(bw0,1000);
bw1= imfill(bw, 'holes');
cf = contour_following(bw);
img = zeros(1000,1000);
img(sub2ind(size(img), cf(:,2), cf(:,1))) = 1;
bw= imfill(img, 'holes');
CHI = imEuler1d(bw);
```

```
perim = imPerimeter(bw);
area = imArea(bw);
DENSITY = imAreaDensity(bw);
Pv = imPerimeterDensity(bw);
OBB = imOrientedBox(bw);
CHI2 = imEuler1d(bw1);
perim2 = imPerimeter(bw1);
area2 = imArea(bw1);
DENSITY2 = imAreaDensity(bw1);
Pv2 = imPerimeterDensity(bw1);
OBB2 = imOrientedBox(bw1);
ent1=entropy(bw);
ent2=entropy(bw1);
stats1 = regionprops(bw1,'MajorAxisLength');
stats = regionprops(bw,'MajorAxisLength');
CA=stats1.MajorAxisLength;
CA2=stats.MajorAxisLength;
nBlack = sum(bw1(:));
nWhite = numel(bw1) - nBlack;
h1{i}=CHI;
h2{i}=perim;
h3{i}=area;
h4{i}=DENSITY;
h5{i}=Pv;
h6{i}=OBB;
h7{i}=CHI2;
h8{i}=perim2;
h9{i}=area2;
h10{i}=DENSITY2;
h11{i}=Pv2;
h12{i}=OBB2;
h13{i}=ent1;
h14{i}=ent2;
h15{i}=CA;
h16{i}=nWhite;
Hfeat{i}=[h1(i) h2(i) h3(i) h4(i) h5(i) h6(i)...
          h7(i) h8(i) h9(i) h10(i) h11(i) h12(i)...
          h13(i) h14(i) h15(i) h16(i)];
end
Hfeat=padcat(Hfeat{:});
Hfeat=cell2mat(Hfeat);


featureVector=[Afeat;Bfeat;Cfeat;Dfeat;Efeat;Ffeat;Gfeat;Hfeat];

a='a';
b='b';
c='c';
d='d';
e='e';
f='f';
g='g';
h='h';

spp1=length(A);
spp2=length(B);
spp3=length(C);
spp4=length(D);
spp5=length(E);
spp6=length(F);
spp7=length(G);
spp8=length(H);

a2=repmat(a,spp1,1);
b2=repmat(b,spp2,1);
c2=repmat(c,spp3,1);
d2=repmat(d,spp4,1);
e2=repmat(e,spp5,1);
f2=repmat(f,spp6,1);
g2=repmat(g,spp7,1);
h2=repmat(h,spp8,1);

valname=[a2;b2;c2;d2;e2;f2;g2;h2];
%%%%%%%%%%%%%%%TEST%%%%%%%%%%%%%%%%%

readImsTest

for i=1 : length(A2)
im=mat2gray(A2{i});
mIM=imfilter(im,fspecial('average',20),'replicate');
```

```
sIM=mIM-im-0.02;
bw0=im2bw(sIM,0);
bw=bwareaopen(bw0,1000);
bw1= imfill(bw, 'holes');
cf = contour_following(bw);
img = zeros(1000,1000);
img(sub2ind(size(img), cf(:,2), cf(:,1))) = 1;
bw= imfill(img, 'holes');
CHI = imEuler1d(bw);
perim = imPerimeter(bw);
area = imArea(bw);
DENSITY = imAreaDensity(bw);
Pv = imPerimeterDensity(bw);
OBB = imOrientedBox(bw);
CHI2 = imEuler1d(bw1);
perim2 = imPerimeter(bw1);
area2 = imArea(bw1);
DENSITY2 = imAreaDensity(bw1);
Pv2 = imPerimeterDensity(bw1);
OBB2 = imOrientedBox(bw1);
ent1=entropy(bw);
ent2=entropy(bw1);
stats1 = regionprops(bw1,'MajorAxisLength');
stats = regionprops(bw,'MajorAxisLength');
CA=stats1.MajorAxisLength;
CA2=stats.MajorAxisLength;
nBlack = sum(bw1(:));
nWhite = numel(bw1) - nBlack;
ta1{i}=CHI;
ta2{i}=perim;
ta3{i}=area;
ta4{i}=DENSITY;
ta5{i}=Pv;
ta6{i}=OBB;
ta7{i}=CHI2;
ta8{i}=perim2;
ta9{i}=area2;
ta10{i}=DENSITY2;
ta11{i}=Pv2;
ta12{i}=OBB2;
ta13{i}=ent1;
ta14{i}=ent2;
ta15{i}=CA;
ta16{i}=nWhite;
tAfeat{i}=[ta1(i) ta2(i) ta3(i) ta4(i) ta5(i) ta6(i)...
          ta7(i) ta8(i) ta9(i) ta10(i) ta11(i) ta12(i)...
          ta13(i) ta14(i) ta15(i) ta16(i)];
end
tAfeat=padcat(tAfeat{:});
tAfeat=cell2mat(tAfeat);


for i=1 : length(B2)
im=mat2gray(B2{i});
mIM=imfilter(im,fspecial('average',20),'replicate');
sIM=mIM-im-0.02;
bw0=im2bw(sIM,0);
bw=bwareaopen(bw0,1000);
bw1= imfill(bw, 'holes');
cf = contour_following(bw);
img = zeros(1000,1000);
img(sub2ind(size(img), cf(:,2), cf(:,1))) = 1;
bw= imfill(img, 'holes');
CHI = imEuler1d(bw);
perim = imPerimeter(bw);
area = imArea(bw);
DENSITY = imAreaDensity(bw);
Pv = imPerimeterDensity(bw);
OBB = imOrientedBox(bw);
CHI2 = imEuler1d(bw1);
perim2 = imPerimeter(bw1);
area2 = imArea(bw1);
DENSITY2 = imAreaDensity(bw1);
Pv2 = imPerimeterDensity(bw1);
OBB2 = imOrientedBox(bw1);
ent1=entropy(bw);
ent2=entropy(bw1);
stats1 = regionprops(bw1,'MajorAxisLength');
stats = regionprops(bw,'MajorAxisLength');
CA=stats1.MajorAxisLength;
```

```
CA2=stats.MajorAxisLength;
nBlack = sum(bw1(:));
nWhite = numel(bw1) - nBlack;
tb1{i}=CHI;
tb2{i}=perim;
tb3{i}=area;
tb4{i}=DENSITY;
tb5{i}=Pv;
tb6{i}=OBB;
tb7{i}=CHI2;
tb8{i}=perim2;
tb9{i}=area2;
tb10{i}=DENSITY2;
tb11{i}=Pv2;
tb12{i}=OBB2;
tb13{i}=ent1;
tb14{i}=ent2;
tb15{i}=CA;
tb16{i}=nWhite;
tBfeat{i}=[tb1(i) tb2(i) tb3(i) tb4(i) tb5(i) tb6(i)...
          tb7(i) tb8(i) tb9(i) tb10(i) tb11(i) tb12(i)...
          tb13(i) tb14(i) tb15(i) tb16(i)];
end
tBfeat=padcat(tBfeat{:});
tBfeat=cell2mat(tBfeat);


for i=1 : length(C2)
im=mat2gray(C2{i});
mIM=imfilter(im,fspecial('average',20),'replicate');
sIM=mIM-im-0.02;
bw0=im2bw(sIM,0);
bw=bwareaopen(bw0,1000);
bw1= imfill(bw, 'holes');
cf = contour_following(bw);
img = zeros(1000,1000);
img(sub2ind(size(img), cf(:,2), cf(:,1))) = 1;
bw= imfill(img, 'holes');
CHI = imEuler1d(bw);
perim = imPerimeter(bw);
area = imArea(bw);
DENSITY = imAreaDensity(bw);
Pv = imPerimeterDensity(bw);
OBB = imOrientedBox(bw);
CHI2 = imEuler1d(bw1);
perim2 = imPerimeter(bw1);
area2 = imArea(bw1);
DENSITY2 = imAreaDensity(bw1);
Pv2 = imPerimeterDensity(bw1);
OBB2 = imOrientedBox(bw1);
ent1=entropy(bw);
ent2=entropy(bw1);
stats1 = regionprops(bw1,'MajorAxisLength');
stats = regionprops(bw,'MajorAxisLength');
CA=stats1.MajorAxisLength;
CA2=stats.MajorAxisLength;
nBlack = sum(bw1(:));
nWhite = numel(bw1) - nBlack;
tc1{i}=CHI;
tc2{i}=perim;
tc3{i}=area;
tc4{i}=DENSITY;
tc5{i}=Pv;
tc6{i}=OBB;
tc7{i}=CHI2;
tc8{i}=perim2;
tc9{i}=area2;
tc10{i}=DENSITY2;
tc11{i}=Pv2;
tc12{i}=OBB2;
tc13{i}=ent1;
tc14{i}=ent2;
tc15{i}=CA;
tc16{i}=nWhite;
tCfeat{i}=[tc1(i) tc2(i) tc3(i) tc4(i) tc5(i) tc6(i)...
          tc7(i) tc8(i) tc9(i) tc10(i) tc11(i) tc12(i)...
          tc13(i) tc14(i) tc15(i) tc16(i)];
end
tCfeat=padcat(tCfeat{:});
tCfeat=cell2mat(tCfeat);
```

```
641
642 - for i=1 : length(D2)
643 -  im=mat2gray(D2{i});
644 -  mIM=imfilter(im,fspecial('average',20),'replicate');
645 -  sIM=mIM-im-0.02;
646 -  bw0=im2bw(sIM,0);
647 -  bw=bwareaopen(bw0,1000);
648 -  bw1= imfill(bw, 'holes');
649 -  cf = contour_following(bw);
650 -  img = zeros(1000,1000);
651 -  img(sub2ind(size(img), cf(:,2), cf(:,1))) = 1;
652 -  bw= imfill(img, 'holes');
653 -  CHI = imEuler1d(bw);
654 -  perim = imPerimeter(bw);
655 -  area = imArea(bw);
656 -  DENSITY = imAreaDensity(bw);
657 -  Pv = imPerimeterDensity(bw);
658 -  OBB = imOrientedBox(bw);
659 -  CHI2 = imEuler1d(bw1);
660 -  perim2 = imPerimeter(bw1);
661 -  area2 = imArea(bw1);
662 -  DENSITY2 = imAreaDensity(bw1);
663 -  Pv2 = imPerimeterDensity(bw1);
664 -  OBB2 = imOrientedBox(bw1);
665 -  ent1=entropy(bw);
666 -  ent2=entropy(bw1);
667 -  stats1 = regionprops(bw1,'MajorAxisLength');
668 -  stats = regionprops(bw,'MajorAxisLength');
669 -  CA=stats1.MajorAxisLength;
670 -  CA2=stats.MajorAxisLength;
671 -  nBlack = sum(bw1(:));
672 -  nWhite = numel(bw1) - nBlack;
673 -  td1{i}=CHI;
674 -  td2{i}=perim;
675 -  td3{i}=area;
676 -  td4{i}=DENSITY;
677 -  td5{i}=Pv;
678 -  td6{i}=OBB;
679 -  td7{i}=CHI2;
680 -  td8{i}=perim2;
681 -  td9{i}=area2;
682 -  td10{i}=DENSITY2;
683 -  td11{i}=Pv2;
684 -  td12{i}=OBB2;
685 -  td13{i}=ent1;
686 -  td14{i}=ent2;
687 -  td15{i}=CA;
688 -  td16{i}=nWhite;
689 -  tDfeat{i}=[td1(i) td2(i) td3(i) td4(i) td5(i) td6(i)...
690 -            td7(i) td8(i) td9(i) td10(i) td11(i) td12(i)...
691 -            td13(i) td14(i) td15(i) td16(i)];
692 - end
693 -  tDfeat=padcat(tDfeat{:});
694 -  tDfeat=cell2mat(tDfeat);
695
696
697 - for i=1 : length(E2)
698 -  im=mat2gray(E2{i});
699 -  mIM=imfilter(im,fspecial('average',20),'replicate');
700 -  sIM=mIM-im-0.02;
701 -  bw0=im2bw(sIM,0);
702 -  bw=bwareaopen(bw0,1000);
703 -  bw1= imfill(bw, 'holes');
704 -  cf = contour_following(bw);
705 -  img = zeros(1000,1000);
706 -  img(sub2ind(size(img), cf(:,2), cf(:,1))) = 1;
707 -  bw= imfill(img, 'holes');
708 -  CHI = imEuler1d(bw);
709 -  perim = imPerimeter(bw);
710 -  area = imArea(bw);
711 -  DENSITY = imAreaDensity(bw);
712 -  Pv = imPerimeterDensity(bw);
713 -  OBB = imOrientedBox(bw);
714 -  CHI2 = imEuler1d(bw1);
715 -  perim2 = imPerimeter(bw1);
716 -  area2 = imArea(bw1);
717 -  DENSITY2 = imAreaDensity(bw1);
718 -  Pv2 = imPerimeterDensity(bw1);
719 -  OBB2 = imOrientedBox(bw1);
720 -  ent1=entropy(bw);
```

```
ent2=entropy(bw1);
stats1 = regionprops(bw1,'MajorAxisLength');
stats = regionprops(bw,'MajorAxisLength');
CA=stats1.MajorAxisLength;
CA2=stats.MajorAxisLength;
nBlack = sum(bw1(:));
nWhite = numel(bw1) - nBlack;
te1{i}=CHI;
te2{i}=perim;
te3{i}=area;
te4{i}=DENSITY;
te5{i}=Pv;
te6{i}=OBB;
te7{i}=CHI2;
te8{i}=perim2;
te9{i}=area2;
te10{i}=DENSITY2;
te11{i}=Pv2;
te12{i}=OBB2;
te13{i}=ent1;
te14{i}=ent2;
te15{i}=CA;
te16{i}=nWhite;
tEfeat{i}=[te1(i) te2(i) te3(i) te4(i) te5(i) te6(i)...
          te7(i) te8(i) te9(i) te10(i) te11(i) te12(i)...
          te13(i) te14(i) te15(i) te16(i)];
end
tEfeat=padcat(tEfeat{:});
tEfeat=cell2mat(tEfeat);


for i=1 : length(F2)
im=mat2gray(F2{i});
mIM=imfilter(im,fspecial('average',20),'replicate');
sIM=mIM-im-0.02;
bw0=im2bw(sIM,0);
bw=bwareaopen(bw0,1000);
bw1= imfill(bw, 'holes');
cf = contour_following(bw);
img = zeros(1000,1000);
img(sub2ind(size(img), cf(:,2), cf(:,1))) = 1;
bw= imfill(img, 'holes');
CHI = imEuler1d(bw);
perim = imPerimeter(bw);
area = imArea(bw);
DENSITY = imAreaDensity(bw);
Pv = imPerimeterDensity(bw);
OBB = imOrientedBox(bw);
CHI2 = imEuler1d(bw1);
perim2 = imPerimeter(bw1);
area2 = imArea(bw1);
DENSITY2 = imAreaDensity(bw1);
Pv2 = imPerimeterDensity(bw1);
OBB2 = imOrientedBox(bw1);
ent1=entropy(bw);
ent2=entropy(bw1);
stats1 = regionprops(bw1,'MajorAxisLength');
stats = regionprops(bw,'MajorAxisLength');
CA=stats1.MajorAxisLength;
CA2=stats.MajorAxisLength;
nBlack = sum(bw1(:));
nWhite = numel(bw1) - nBlack;
tf1{i}=CHI;
tf2{i}=perim;
tf3{i}=area;
tf4{i}=DENSITY;
tf5{i}=Pv;
tf6{i}=OBB;
tf7{i}=CHI2;
tf8{i}=perim2;
tf9{i}=area2;
tf10{i}=DENSITY2;
tf11{i}=Pv2;
tf12{i}=OBB2;
tf13{i}=ent1;
tf14{i}=ent2;
tf15{i}=CA;
tf16{i}=nWhite;
tFfeat{i}=[tf1(i) tf2(i) tf3(i) tf4(i) tf5(i) tf6(i)...
          tf7(i) tf8(i) tf9(i) tf10(i) tf11(i) tf12(i)...
```

```
          tf13(i) tf14(i) tf15(i) tf16(i)];
end
tFfeat=padcat(tFfeat{:});
tFfeat=cell2mat(tFfeat);

for i=1 : length(G2)
im=mat2gray(G2{i});
mIM=imfilter(im,fspecial('average',20),'replicate');
sIM=mIM-im-0.02;
bw0=im2bw(sIM,0);
bw=bwareaopen(bw0,1000);
bw1= imfill(bw, 'holes');
cf = contour_following(bw);
img = zeros(1000,1000);
img(sub2ind(size(img), cf(:,2), cf(:,1))) = 1;
bw= imfill(img, 'holes');
CHI = imEuler1d(bw);
perim = imPerimeter(bw);
area = imArea(bw);
DENSITY = imAreaDensity(bw);
Pv = imPerimeterDensity(bw);
OBB = imOrientedBox(bw);
CHI2 = imEuler1d(bw1);
perim2 = imPerimeter(bw1);
area2 = imArea(bw1);
DENSITY2 = imAreaDensity(bw1);
Pv2 = imPerimeterDensity(bw1);
OBB2 = imOrientedBox(bw1);
ent1=entropy(bw);
ent2=entropy(bw1);
stats1 = regionprops(bw1,'MajorAxisLength');
stats = regionprops(bw,'MajorAxisLength');
CA=stats1.MajorAxisLength;
CA2=stats.MajorAxisLength;
nBlack = sum(bw1(:));
nWhite = numel(bw1) - nBlack;
tg1{i}=CHI;
tg2{i}=perim;
tg3{i}=area;
tg4{i}=DENSITY;
tg5{i}=Pv;
tg6{i}=OBB;
tg7{i}=CHI2;
tg8{i}=perim2;
tg9{i}=area2;
tg10{i}=DENSITY2;
tg11{i}=Pv2;
tg12{i}=OBB2;
tg13{i}=ent1;
tg14{i}=ent2;
tg15{i}=CA;
tg16{i}=nWhite;
tGfeat{i}=[tg1(i) tg2(i) tg3(i) tg4(i) tg5(i) tg6(i)...
          tg7(i) tg8(i) tg9(i) tg10(i) tg11(i) tg12(i)...
          tg13(i) tg14(i) tg15(i) tg16(i)];
end
tGfeat=padcat(tGfeat{:});
tGfeat=cell2mat(tGfeat);

for i=1 : length(H2)
im=mat2gray(H2{i});
mIM=imfilter(im,fspecial('average',20),'replicate');
sIM=mIM-im-0.02;
bw0=im2bw(sIM,0);
bw=bwareaopen(bw0,1000);
bw1= imfill(bw, 'holes');
cf = contour_following(bw);
img = zeros(1000,1000);
img(sub2ind(size(img), cf(:,2), cf(:,1))) = 1;
bw= imfill(img, 'holes');
CHI = imEuler1d(bw);
perim = imPerimeter(bw);
area = imArea(bw);
DENSITY = imAreaDensity(bw);
Pv = imPerimeterDensity(bw);
OBB = imOrientedBox(bw);
CHI2 = imEuler1d(bw1);
perim2 = imPerimeter(bw1);
area2 = imArea(bw1);
DENSITY2 = imAreaDensity(bw1);
```

```
Pv2 = imPerimeterDensity(bw1);
OBB2 = imOrientedBox(bw1);
stats1 = regionprops(bw1,'MajorAxisLength');
stats = regionprops(bw,'MajorAxisLength');
CA=stats1.MajorAxisLength;
CA2=stats.MajorAxisLength;
nBlack = sum(bw1(:));
nWhite = numel(bw1) - nBlack;
ent1=entropy(bw);
ent2=entropy(bw1);
th1{i}=CHI;
th2{i}=perim;
th3{i}=area;
th4{i}=DENSITY;
th5{i}=Pv;
th6{i}=OBB;
th7{i}=CHI2;
th8{i}=perim2;
th9{i}=area2;
th10{i}=DENSITY2;
th11{i}=Pv2;
th12{i}=OBB2;
th13{i}=ent1;
th14{i}=ent2;
th15{i}=CA;
th16{i}=nWhite;
tHfeat{i}=[th1(i) th2(i) th3(i) th4(i) th5(i) th6(i)...
          th7(i) th8(i) th9(i) th10(i) th11(i) th12(i)...
          th13(i) th14(i) th15(i) th16(i)];
end
tHfeat=padcat(tHfeat{:});
tHfeat=cell2mat(tHfeat);


tfeatureVector=[tAfeat;tBfeat;tCfeat;tDfeat;tEfeat;tFfeat;tGfeat;tHfeat]

a='a';
b='b';
c='c';
d='d';
e='e';
f='f';
g='g';
h='h';

spp1=length(A);
spp2=length(B);
spp3=length(C);
spp4=length(D);
spp5=length(E);
spp6=length(F);
spp7=length(G);
spp8=length(H);

a2=repmat(a,spp1,1);
b2=repmat(b,spp2,1);
c2=repmat(c,spp3,1);
d2=repmat(d,spp4,1);
e2=repmat(e,spp5,1);
f2=repmat(f,spp6,1);
g2=repmat(g,spp7,1);
h2=repmat(h,spp8,1);

Target=[a2;b2;c2;d2;e2;f2;g2;h2];



cmap=[1 1 0;1 1 0;1 1 0;1 1 0;1 1 0;1 1 0;1 1 0;1 1 0;1 1 0;1 1 0;1 1 0;1 1 0;...
0 1 0;0 1 0;0 1 0;0 1 0;0 1 0;0 1 0;0 1 0;0 1 0;0 1 0;0 1 0;0 1 0;0 1 0;...
0 0 1;0 0 1;0 0 1;0 0 1;0 0 1;0 0 1;0 0 1;0 0 1;0 0 1;0 0 1;0 0 1;0 0 1;...
0 1 1;0 1 1;0 1 1;0 1 1;0 1 1;0 1 1;0 1 1;0 1 1;0 1 1;0 1 1;0 1 1;0 1 1];

%featureVector=[featureVector(:,1:22)];
%tfeatureVector=[tfeatureVector(:,1:22)];


[classhypo,L] = KNNI(featureVector,tfeatureVector,valname,10);
classhypo=char(classhypo);
confusion_matrix(classhypo,Target,{'Smm','Tp','Mmi','Mma','Tl','Tm','My'


[eigvector, eigvalue, elapse] = LDA(valname,'PCARatio',featureVector);
yyy=featureVector*eigvector;
tyyy=tfeatureVector*eigvector;

[classhypo,L] = KNNI(yyy,tyyy,valname,10);
classhypo=char(classhypo);
confusion_matrix(classhypo,Target,{'Smm','Tp','Mmi','Mma','Tl','Tm','My'

```