ENSEMBLE CLASSIFICATION AND SIGNAL IMAGE PROCESSING FOR THE GENUS *GYRODACTYLUS* (MONOGENEA)

ROZNIZA ALI

Doctor of Philosophy

Institute of Computing Science and Mathematics

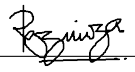
University of Stirling

May 2014

DECLARATION

I hereby declare that this thesis has been composed by myself, that the work and results have not been presented for any university degree prior to this and that the ideas that I do not attribute to others are my own.

Stirling, May 2014



Rozniza Ali

ABSTRACT

This thesis presents an investigation into *Gyrodactylus* species recognition, making use of machine learning classification and feature selection techniques, and explores image feature extraction to demonstrate proof of concept for an envisaged rapid, consistent and secure initial identification of pathogens by field workers and non-expert users. The design of the proposed cognitively inspired framework is able to provide confident discrimination recognition from its non-pathogenic congeners, which is sought in order to assist diagnostics during periods of a suspected outbreak. Accurate identification of pathogens is a key to their control in an aquaculture context and the monogenean worm genus *Gyrodactylus* provides an ideal test-bed for the selected techniques. In the proposed algorithm, the concept of classification using a single model is extended to include more than one model. In classifying multiple species of *Gyrodactylus*, experiments using 557 specimens of nine different species, two classifiers and three feature sets were performed. To combine these models, an ensemble based majority voting approach has been adopted. Experimental results with a database of *Gyrodactylus* species show the superior performance of the ensemble system. Comparison with single classification approaches indicates that the proposed framework produces a marked improvement in classification performance. The second contribution of this thesis is the exploration of image processing techniques. Active Shape Model (ASM) and Complex Network methods are applied to images of the attachment hooks of several species of *Gyrodactylus* to classify each species according to their true species type. ASM is used to provide landmark points to segment the contour of the image, while the Complex Network model is used to extract the information from the contour of an image. The current system aims to confidently classify species, which is notifiable pathogen of Atlantic salmon, to their true class with high degree of accuracy. Finally, some concluding remarks are made along with proposal for future work.

ACKNOWLEDGMENTS

First of all, I wish to thank to Allah SWT, God Almighty for giving me strength, health and determination to complete my PhD studies. This special dedication is also for my lovely husband and family who support me around the hard times and cheeky sons who always make me strong and smile during the hard times.

I am very much indebted to Prof. Amir Hussain, my first supervisor, who gave me the initial idea to develop this exciting inter-disciplinary research and provided insighful advice, guidance and support throughout. My appreciation is extended to my additional supervisors from Stirling Institute of Aquaculture, Dr. Andrew Shinn and Dr. James Bron for their valuable inputs, positive and constructive comments and guidance. To all Computer Science and Mathematics, and Aquaculture staffs, I am privileged to have learned many new things, especially to Dr. Andrew Abel who was always my timely proof reader.

Over the course of the last four years, I have made many friends from all over the world. Thus, thank you very much for all of you as I cannot mention each and every one. I really thank you for your friendship. To my PhD cluster / COSIPRA Lab colleagues: Kamran, Hicham, Thomas and Saliha, thank you very much for sharing laughter during my PhD research period.

I would not forget my employer, Universiti Malaysia Terengganu (UMT). In particular, I would express my thanks to the head of department of Computer Science, Faculty of Informatic and Applied Mathematic for giving me the recommendation and permission to continue my PhD at Stirling University.

Most importantly, I remain grateful to the Ministry of Malaysian Education for providing my scholarship to study in Scotland, United Kingdom. Without this scholarship, I would not have come to this beautiful country to gain new knowledge and many valuable lifelong experiances.

LIST OF PUBLICATIONS

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

- R. Ali, A. Hussain, A. P. Shinn and J. E. Bron. The Use of ASM Feature Extraction and Machine Learning for the Discrimination of Members of the Fish Ectoparasite Genus *Gyrodactylus*. In 19th International Conference on Neural Information Processing (ICONIP2012), volume 7666, pages 256-263, Springer Berlin Heidelberg, 2012.

- R. Ali, B. Jiang, A. Hussain, B. Luo, A. P. Shinn and J. E. Bron. Classification of fish ectoparasite genus *Gyrodactylus* SEM images using ASM and complex network model. In 21st International Conference on Neural Information Processing (ICONIP 2014), volume 8836, pages 103-110, 2014.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

1

INTRODUCTION

As wild fish stocks decline through the effects of over-fishing, anthropogenic activities and more insidious environmental changes, the importance of aquaculture in maintaining global food supply increases. Since 1970, the aquaculture sector has expanded by over 9% per annum, the highest current growth profile in any agricultural[1] sector (c.f. 1.2% for capture fisheries and 2.8% for land meat production) [91]. Accompanying this increase in aquaculture[2] production is a tendency towards intensification of aquaculture practices, with higher stocking densities and greater utilisation of available water resources. Such developments almost unavoidably lead to an increase in disease[3] problems including those associated with parasitic pathogens[4]. Amongst fish parasites, ectoparasitic [5] monogenetic platyhelminths remain a particularly intransigent economic burden for the global freshwater[6] and marine aquaculture industries. One of the most widespread groups of monogenean[7] parasites is the genus *Gyrodactylus*[8], whose members are common ectoparasitic[9] of fish within aquaculture and wild capture fisheries, with more than 409 species identified to date [61].

One of the key challenges for disease management and control in cultured and wild fish populations in the 21st Century is that of achieving secure and consistent pathogen identification. The rapid expansion of fish culture into new environments and fisheries

1 Key development in the rise of sedentary human civilization, the farming of domesticated created food surpluses that nurtured the development civilization. the study of agriculture is known as agricultural science.

2 it involves cultivating freshwater and saltwater populations under controlled conditions, and can be contrasted with commercial fishing, which is the harvesting of wild fish.

3 Disorder of structure or function in a human, animal, or plant, especially one that produces specific signs or symptoms or that affects a specific location and is not simply a direct result of physical injury.

4 A pathogen is anything that causes a disease.

5 A parasite that lives on the surface of a host organism.

6 Naturally occurring water on the Earth's surface in ice sheets, ice caps, glaciers, icebergs, bogs, ponds, lakes, rivers and streams, and underground as groundwater in aquifers and underground streams.

7 Group of largely ectoparasitic members of the flatworm phylum Platyhelminthes, class Monogenea.

8 Small monogenean ectoparasite (about 0.5 mm long) which mainly lives on the skin of freshwater fish, especially Atlantic salmon.

9 A parasite that lives on the surface of a host organism.

management operations more broadly, has overtaken our ability to recognise certain individual species of parasite pathogens. Recent evidence has also demonstrated that the international translocation of fish has increased the rate of introduction of exotic parasite species into indigenous fish stocks, with serious economic consequences. In the U.K., which is one of the best studied regions for exotic fish pathogen introduction, the documentation of 14 relatively recently introduced metazoan parasites has been of major concern, with ten of these parasite species already being well established [56]. For this reason and because of the widely varying pathogenicity seen between closely related species, accurate pathogen identification is of paramount importance. Whilst molecular techniques have, in recent years, offered much in the way of species discrimination, species definitions often continue to rely on morphological characteristics (attachment hook morphology[10]), which may correlate to pathogenicity but not necessarily to recognised discriminatory molecular markers [33]. Thus some pathogenic species can only be classified from non-pathogenic relatives by morphological characterisation. This is particularly true for many monogeneans, whose discrimination from congeners is compounded by a limited number of morphological discrete characteristics, which makes identification difficult. These characteristics include often small size, slight morphological differences in key taxonomically important features, lack of patterning or colouration and fragility, which requires specimens to be live or immediately fixed or preserved.

Accurate classification of parasitic pathogens to the correct specimen group can be a difficult and time consuming task. In the event of an outbreak of serious disease, the demand for identification may significant exceed the available supply of suitable expertise and facilities. In order to reduce the death rate of fish infected with *G. salaris* and certain *Gyrodactylus* species, it is vital to correctly diagnose infections and the species present as early as possible. If suitable measures of parasite control are not developed, then this can have a negative impact on fish production. In recent years, there has been a research focus in determining solutions to tackle these issues and various experiments have been conducted by domain experts to find the best solutions to contain disease outbreak without any detrimental impact on the environment. There are several reasons for potential difficulty in defining the parasite category among

10 Branch of biology dealing with the study of the form and structure of organisms and their specific structural features. This includes aspects of the outward appearance (shape, structure, colour, pattern) as well as the form and structure of the internal parts like bones and organs.

Gyrodactylus specimen group. This task is heavily reliant on a limited number of domain experts available to analyse and determine specimen groups. Additionally, Kay *et al.* [76] describe in detail the challenge associated with parasite specimen classification by shape due to the small nature of the difference between each one.

This study focuses on the discrimination of *Gyrodactylus* species commonly infecting salmonids. These include *G. derjavinoides* Malmberg, Collins, Cunningham et Jalali, 2007, *G. kherulensis* Ergens, 1974, *G. salaris* Malmberg, 1957, *G. thymalli* Žitňan, 1960 and *G. truttae* Gläser, 1974. In addition, species that could accidentally parasitise salmonids such as *G. arcuatus* Bychowsky, 1933 and *G. gasterostei* Gläser, 1974 were included, as well as species that may be confused with *G. salaris*, species such as *G. lucii* Kulakovskaya, 1952. Other potentially problematic species including *G. sommervillae* Turgut, Shinn, Yeomans et Wootten, 1999 were also considered. Finally *G.cichlidarum* Paperna, 1968, a parasite of Nile tilapia, Oreochromis n. niloticus (L.) was included as an outlying group. *Gyrodactylus salaris* and *G. thymalli* were identified as the most difficult to classify since they are closely related and morphologically similar. Shinn et al. [130] explains that the morphometric discrimination of *Gyrodactylus* species can be difficult due to the small size of taxonomically important structures i.e. the haptoral attachment hooks. A range of alternative techniques have been explored to assess their utility in discriminating and identifying species, these techniques include chaetotaxy, probe hybridization, and multivariate analysis [88], [130].

To assist and to provide accurate classification for *Gyrodactylus* species, a number of computer techniques are explored in this thesis for their potential usefulness. There are several methods that can make the computer more intelligent and to give it enough intelligence to recognise and to understand the images that the user gives to it. One of this ways is using the Artificial Intelligence (AI) approach. Using AI techniques such as machine learning will help us to recognise and classify the entered image, which will provide a big contribution in the aquaculture domain, especially in parasite recognition and classification.

Thus computer-assisted analysis becomes quite necessary in practice. To date, many automated detection algorithms have been developed, such as detection of the IHHN virus in shrimp tissue by digital color correlation [12], parasite detection in fish [25], identification

of mammalian species [110], and leaf species recognition [44]. These reference approaches provide the inspiration to develop such system for *Gyrodactylus* species recognition.

1.1 PROBLEM STATEMENTS

Classification of the *Gyrodactylus* species group poses a range of substantial interdisciplinary challenges: Firstly, the accumulation of enough expert knowledge to reliably distinguish between similar species, say, *G. arcuatus*, *G. salaris* and so on; Secondly, manual classification is highly labour intensive and time consuming; and thirdly, the most formidable challenge occurs when the required point-to-point measurements are not accurately taken, which lead to inaccurate species classifications. Finally, at the last stage of classification, only specialist domain experts can determine the correct species based on their own vision and experience.

To improve the correct identification of *G. salaris*, a number of morphometric techniques based on statistical classification techniques [76], [104], [126] and molecular techniques [32],[33], [105], [60] have been developed to classify this pathogen from its close relatives on salmonid hosts. Whilst expert taxonomists may be able to classify *G. salaris* from other closely related species, morphometric speciation and molecular characterisation of the *Gyrodactylus* species is frustrating and difficult for the reasons described earlier. For this reason, this project aims to develop novel computational intelligence techniques involving intelligent signal image capture, processing, and analysis and employing cutting edge Artificial Intelligence (AI) algorithms and technologies to provide automated or semi-automated species identification, initially to classify *G. salaris* from other European gyrodactylids but with the eventual objective of developing tools to classify *Gyrodactylus* generally. It is hoped that the techniques developed as a result of this work will be of practical relevance and effective for identification or classification of a range of aquatic and indeed terrestrial pathogens, and for taxonomic identification more widely.

1.2 THESIS MOTIVATIONS

The following justifications are provided for the study:

1. Aquaculture continues to expand worldwide but this expansion has been accompanied by increased disease problems including those associated with ectoparasitic monogenean worms.

2. Discrimination of pathogenic from non-pathogenic species is a key current requirement to allow expertise, industry and government level control and management of pathogens of wild and cultured fish. However, shortage of taxonomic experts and the shortcomings of molecular methods often make this difficult to achieve in practice.

3. Use of a combination of intelligent signal image processing including those nominally described as AI, provide the opportunity to develop state-of-the-art automated or semi-automated intelligent systems for pathogen recognition. These will allow rapid, consistent and secure initial identification of pathogens by field workers and non-expert users.

1.3 AIMS AND OBJECTIVES OF RESEARCH

The following are the main objectives for this research project:

1. To investigate the application of novel signal image processing based intelligent systems approaches to aquatic parasite recognition.

2. To provide a quantitative and qualitative assessment of the best current intelligent signal image processing techniques and technologies that can be applied to the classification of *Gyrodactylus* species in general and *Gyrodactylus salaris* in particular, specifically those based on hook morphology.

3. To develop an intelligent automated or semi-automated system, employing advanced intelligent signal image processing and image recognition technologies that can be utilised by non-expert users for local and global *Gyrodactylus* species recognition.

1.4 SIGNIFICANCE AND BENEFITS OF RESEARCH

The outcome of this study will assist pathogen management in wild and cultured fish stocks with attendant improvements in fish health and welfare and accompanying economic benefits. The intelligent systems and intelligent signal image processing techniques to be developed in the course of this project are hoped to have wide utility for species recognition within and outside the field of aquatic sciences. Additionally, there may be attendant intellectual property benefits within these fields and potentially in other fields where real-time image recognition or classification is important.

1.5 CONTRIBUTIONS OF THESIS

The following are the original contributions of this interdisciplinary research project:

1. In the aquaculture domain, the main contribution is in development of an intelligent system that will assist technical workers and research scientists to perform rapid, secure, and accurate prediction of multiple species of *Gyrodactylus*. The proposed feature extraction and classification methods will reduce the required staff and management resource requirements for species analysis and identification.

2. With regards to the ensemble methodology, this work has provided a novel contribution by proposing and applying this approach to a completely new research domain, *i.e.* the application of the ensemble technique in the context of aquaculture. Specifically, in the proposed ensemble based majority voting approach, two classifiers (*e.g.* Linear Discriminant Analysis (LDA) and K-nearest Neighbor (K-NN)), and three feature sets (*e.g.* 25 features, 21 features and 20 features) have been utilised and evaluated and found to accurately classify nine different species of *Gyrodactylus*.

3. Intelligent signal image processing techniques have been explored in order to identify the potential of different methods for implementation, with regard to the extraction of valuable and suitable features for the purpose of classification. Active Shape Models (ASM) were identified as being capable of delivering good results, when applied previ-

ously by other researchers to the domain of medical image processing. Therefore, we developed ASMs that could be applied to the domain of aquaculture (*Gyrodactylus* body shapes). Specifically, in this novel aquaculture research domain, our results demonstrate that ASM has significant potential for extracting Scanning Electron Microscope (SEM) images. Further, instead of applying ASM alone, the potential of integrating a Complex Network model is also explored, for the first time. A Complex Network model can be described as the intersection between graph theory and statistical mechanics, which confers a truly multidisciplinary nature upon this research, since it integrates computer sciences, mathematics and physics. Here, ASM and a Complex Network model have been combined to develop an innovative approach for segmentation and extraction of features. This new contribution has resulted in an improvement over using a Complex Network model alone, since in this proposed method, the automated process of contour segmentation is performed by ASM, compared to previous work, where contour segmentation is carried out manually.

1.6 STRUCTURE OF THE THESIS

The structure of the remainder of this thesis is as follows. Chapter 2 describes background research concerning the *Gyrodactylus* species. Rather than focusing on the entire monogenean group, this thesis focuses on the *Gyrodactlus* species. In this chapter, the history, morphology, mortality, species distinction, and identification are discussed in depth.

Chapter 3 discusses and reviews a number of machine learning classifiers and feature selection techniques. In addition to a detailed review of these techniques and classifiers, these research components are then used as part of further classification research. This chapter presents experimental results of using a number of techniques to classify 557 specimens from nine different species of *Gyrodactylus*. The positive findings of this research have been published as part of the proceedings of Intelligent Systems Design and Applications in 2011 [8]. Also in this chapter, the initial results presented in section of morphometric classification have been enhanced and improved by proposing and implementing an ensemble classification technique based on majority voting.

Chapter 4 focuses on image processing. The chapter introduces and discusses the approach used to extract the most suitable features from Scanning Electron Microscope (SEM) images of three different species of *Gyrodactylus*. In this chapter, the Active Shape Model (ASM) technique has been introduced and described in depth. This approach is used in this thesis to extract features that can be used in order to improve species classification. This feature extraction methodology and the subsequent classification results presented in the thesis chapter are published in Neural Information Processing [9]. Instead of applying ASM alone, the potential of a Complex Network model is also explored. Here, ASM and a Complex Network model have been combined for segmentation and extraction of features. This new contribution has resulted in the improvement over using a Complex Network model alone, as in this method, the automated process of contour segmentation was performed by ASM, compared to previous work, where contour segmentation was carried out manually. This research was published at the Neural Processing Conference in Kuching, Malaysia [10].

Finally, chapter 5 summarises this thesis, provides some concluding remarks, and recommends a number of directions for future work.

2

BACKGROUND OF GYRODACTYLUS

2.1 INTRODUCTION

Before the overall aim of this thesis, the development of an intelligent signal image processing ensemble classification system for *Gyrodactylus* species identification, can be presented, it is important that the background to this work is discussed. This chapter discusses the species which is the focus of this research, Monogenea of *Gyrodactylus*. Monogenea of the genus *Gyrodactylus* may occur in fresh water, and brackish and marine environments. To be precise, *G. salaris* lives and reproduces in fresh water, and can tolerate brackish water for short or longer periods depending on salinity levels [1]. *Gyrodactylus salaris* is known as the salmon killer in several countries due to the impact on mortality rates. According to Bakke *et al*. [18], many researchers have ignored or avoided studying this area due to the complexity of their taxonomy. The mortality effect on the production of Atlantic salmon, has created a *Gyrodactylus salaris* epidemic, which has stimulated research to such an extent that gyrodactylids are now the best studied of all monogeans.

After a summary of the history of this species, the remainder of this chapter presents the morphology of *Gyrodactylus* parasite species. In this section, the characteristics of the species are described. This information will be used for measurement and prediction of the class label of multiple species of Gyrodactylus. These are: *G. arcuatus*, *G. cichilidarum*, *G. derjavinoides*, *G. kherulensis*, *G. salaris*, *G. sommervillae*, *G. thymalli* and *G. truttae*. In this research, three parts of the haptoral hooks are used to provide informative features for species identification; hamuli, marginal hook and ventral bar.

The reproduction of this species has caused a huge impact on the ecological system and world food supply. This species has been demonstrated to be responsible for the decline in salmon stocks in Norway, where it has caused damage to salmon stocks in more than 40 rivers

resulting in the near extermination of the salmon population in five of these rivers [18]. This will be discussed in Section 2.4. To solve the problem of damage in salmon stocks, various techniques have been introduced, and are used for controlling and preventing the spread of infected rivers. These are presented in Section 2.5. Finally, the chapter will be summarised in the last section.

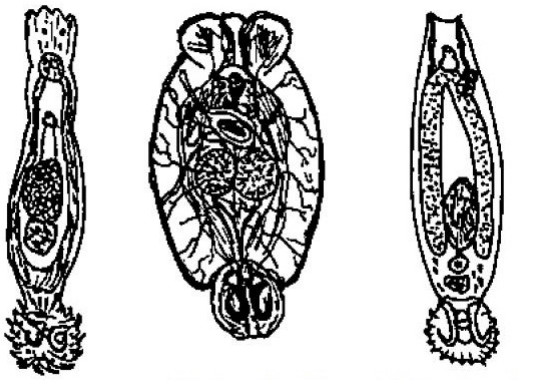
2.2 HISTORY

Gyrodactylids are ubiquitous monogenean ectoparasites on the skin and gills of teleost fish both in marine and fresh water ecosystems. The most recent species compilation lists some 400 gyrodactyids species [61]. *Gyrodactylus* species Malmberg 1957, is a species of genus found on fins and skin of Atlantic and Baltic salmon in its freshwater phase. The small (0.5-1mm) parasite was first described by Malmberg (1957) from salmon parr (scientific name of Salmon) in a hatchery situated at the river Indalsalven in Sweden. Since then, there have been a growing number of observations of *G. salaris* from several countries both on wild fish and on fish in hatcheries and freshwater fish farms [1].

There are three species that are categorised as being part of the Monogeneans group [2]. These are *Dactylogyrus*, *Benedeniella* and *Gyrodactylus*, as shown in Figure 2.1. *Dactylogyrus* is usually attached to the gills of freshwater fish. It reproduces by laying eggs, which are often resistant to chemical treatment, therefore weekly treatment over a period of 34 weeks is recommended. The second type is *Benedenielle*, which is a large monogenea that can cause chronic problems in marine systems and is difficult to eliminate from a system once established. The last type, which is the species focused on in this research, is *Gyrodactylus*. It is usually found on the skin and fins of freshwater fish and produces live young, so one treatment may be adequate to control an infestation.

Gyrodactylus has become a common infected species on fish farms and wild fish populations. 200 species have been identified as part of the *Gyrodactylus* species group and mostly it comes from North America and Eurasia [18].

The systematics of *G. salaris* and its closest relatives is complex. For example, there is no support for the morphology of either *G. salaris* or its closest relative *G. thymalli*. *G. salaris* has



(a) *Dactylogyrus* (b) *Benedenielle* (c) *Gyrodactylus*

Figure 2.1: The three type of monogeneans: (a) *Dactylogyrus*; (b) *Benedenielle*; and (c) ***Gyrodactylus***.

been introduced in recent years to rivers in Norway, to rivers on the Swedish west coast, and to a Russian river draining into the White Sea. Until 2007, *G. salaris* was not found in Poland [1].

Monogenea of the genus *Gyrodactylus* have been known for almost 180 years for their retention of fully grown offspring in utero [1] until they themselves contain developing embryos. *Gyrodactylus* was first described as being found in bream (*Abramis brama*) by von Nordmann (1832). Gyrodactylids were particularly useful to early microscopists as they are flatworms without an impervious egg shell. The enclosure of several embryos inside each other also represented an attractive model for the study of germ cell lineages, a paradigm which was just becoming established at this time [18].

In Norway in 1975, an infestation of *G. salaris* [18] caused extensive damage to salmon stocks at the Akvaforsk fish hatchery in Sunndalϕra, Mϕre and Romsdal Country. Meanwhile, in the UK, there have been no reported infections from *G. salaris*. According to Shinn *et al.* [124], fish health authorities operate and execute extensive screening of several species of salmonid to avoid mortality damage from *G. salaris*.

Monogeneans are a class of parasitic flatworms that are commonly found on fish and lower aquatic invertebrates. Most monogeneans are browsers that move about freely on the fish body surface, feeding on mucus and epithelial cells of the skin and gills; however a few adult monogeneans will remain permanently attached to a single site on the host. Some monogenean species invade the rectal cavity, ureter, body cavity and even the blood vascular

1 The state of an embryo or fetus - embryo.

system. Between 4,000 and 5,000 species of monogeneans are found on fish in fresh and salt water and in a wide range of water temperatures [2].

Until the mid-1990s, most *Gyrodactylus* species were identified by the comparing the morphology of the hard parts in the attachment organ, the opisthaptor [2] [1]. Over recent years, the application of molecular markers in the taxonomy and systematics of *Gyrodactylus* species has increased. A description of *G. salaris* identification routines is given by the World Organisation for Animal Health (OIE, Manual of Diagnostic Tests for Aquatic Animals). Today, identification of *G. salaris* is predominantly based on the sequence of the mitochondrial gene Cytochrome Oxidase 1 (CO_1) [3] [1]. From the computer scientist point of view, the aim is to provide a tool or system that will automatically and systematically predict the class species. Achieving the correct, rapid and secure prediction will help to save infected species or rivers from damage.

2.3 MORPHOLOGY

These types of ectoparasite have only been found on the skin but they are often present in significant numbers. They have viviparous or live offspring. *Gyrodactylus* are up to 2mm in length and can readily be distinguished from other monogeneans in skin smears under the microscope. Bakke *et al*. [18] stated that *Gyrodactylus* species are the smallest species among monogeneans. The *Gyrodactylus* species are widely known as a unique species as they carry babies during their development; they are known as viviparous fish parasites. This type of species does not have any specific transmission stage; the viviparous worms give birth to fully grown adults, which during the birth process to the same host as the parent and only subsequently may transfer to a new host [1].

Affected fish, of any age, but generally young fast growing stock, have dark patches over the body surface, with sloughing areas of skin. They generally do not feed, and the parasites can be readily seen in skin scrapings, often accompanied by trichodinids (Figure 2.2).

These small (<1mm) viviparous flatworms are ubiquitous, as shown on the fish in Figure 2.3, can infect cephalopods and aquatic tetrapods, and also may be pathogenic and kill their hosts. Some are economically important pathogens of wild and cultured finfish. Gyrodactylids

2 The posterior and usually complex adhesive organ of a monogenetic.

3 Protein that in humans is encoded by the MT-CO_1 gene.

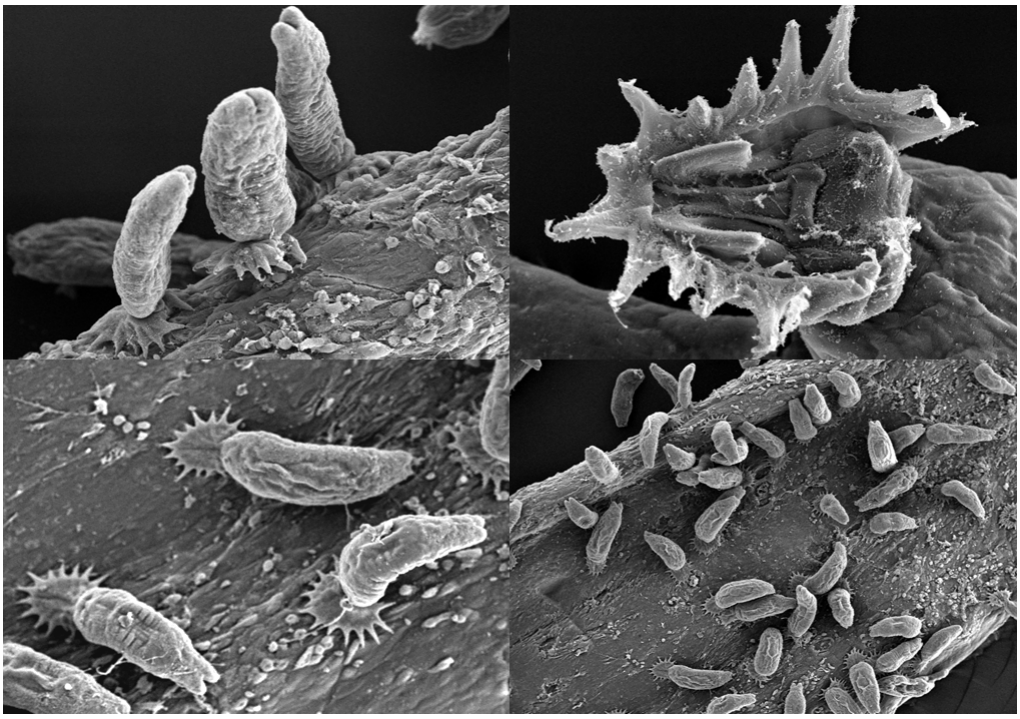


Figure 2.2: *Gyrodactylus salaris* clustered on the fin of a small salmon.

are maintained easily in the laboratory and display short, direct life cycles, making them ideal models to provide insights into a variety of key parasitology questions [61]. Because of their small size, researchers have investigated making use of the hamuli, marginal hooks and ventral bars for species recognition and identification. These features will be described in more depth in Chapter 3, and are shown in Figure 3.1. A detailed explanation of the extraction features from haptoral hooks will be provided in Chapter 3.

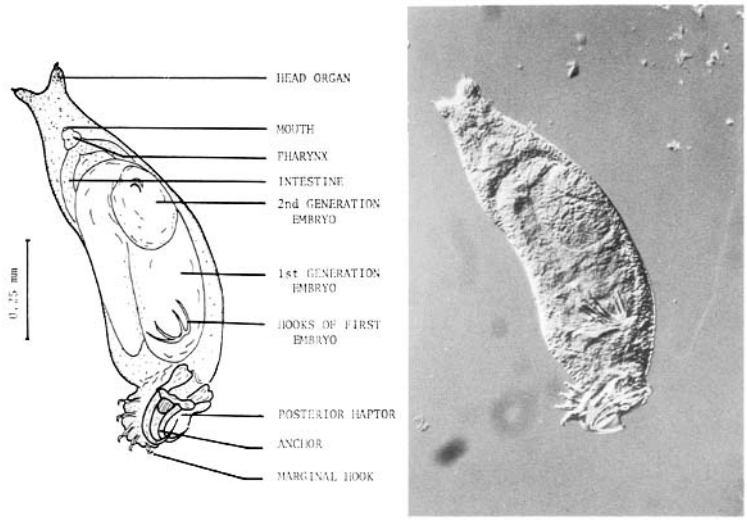


Figure 2.3: *Gyrodactylus* (Monogenea) from the skin of Clarias batrachus fry.

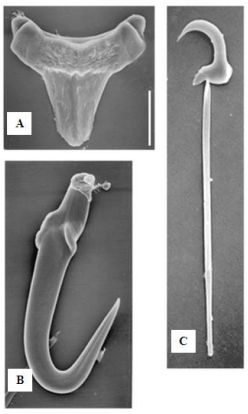


Figure 2.4: *Gyrodactylus* (Monogenea) skeleton hooks morphology; (a) ventral bar, (b) hamulus, and (c) marginal hook.

G. salaris is spread through anthropogenic movement of infected fish between hatcheries or fish farms, between hatcheries or fish farms and rivers, and by migration of infected fish in rivers and in brackish water in fjords [4] [1]. *G. salaris* can survive and reproduce on rainbow trout. Rainbow trout have been capable of sustaining *G. salaris* infections over time (in fish farms and in wild populations, respectively). Introduction of the parasite to habitats in which one or more of these species are found is thus a potential range extension. The remaining species can act as transport hosts within their respective maximum capacity for sustaining an infection.

2.4 MORTALITY

A key example of the problems described earlier is that the quick reproduction of this species has caused a huge impact on the ecological system and world food supply. The classification of a highly pathogenic monogenean species associated with wild and cultured fish mortalities, from other closely related non-pathogenic species, *Gyrodactylus salaris* Malmberg, 1957 [18], which is considered to be very highly pathogenic to Atlantic salmon, (Salmo salar L) is essential. This species was responsible for the catastrophic decline in salmon stocks in Norway, where it has caused damage to salmon stocks in more than 40 rivers [18]. Dolmen [42] and Mo [108] have estimated the impact of *G. salaris* on juvenile salmon in Norway and found *G.*

4 Deep, narrow and elongated sea or lakedrain, with steep land on three sides.

salaris to be responsible for a decline of up to 520 tonnes (20% of the total catch) per year in catches of adult salmon returning to the rivers to spawn. *Gyrodactylus salaris* is known to be present in 11 European countries, and was most recently identified in Italy, where it has been found in multiple rainbow trout sites. Analysis of archive material from Italy would suggest that it has been there for many years but has been overlooked [111].

Gyrodactylids monogeneans are widespread parasites of freshwater and marine fishes. Olstad *et al.* [1] reported that the disease resulting from *Gyrodactylus* infections, gyrodactylosis, has responsible for losses in a wide variety of captive fish species. Little is known about the disease Gyrodactylosis and the cause of death in infected individuals. In Norway, *G. salaris* has caused epidemics that have devastated stocks of Atlantic salmon in many rivers. The density of salmon in infected rivers has been reduced by an average of 86% and the catch of salmon in infected rivers is reduced on by an average of 87% [1], [68].

An indirect effect of *G. salaris* may be the negative effect upon the freshwater pearl mussel Margaritifera margaritifera caused by reductions in salmon populations. This may cause reductions in the populations of the fresh water pearl mussel because the larvae (glochidiae) of the pearl mussel are dependent on Atlantic salmon in certain aspects of their life [1]. Studies have revealed that freshwater pearl mussel larvae in many water courses have an obligatory period either in the gills of salmon or trout [84].

In Norway the catch of salmon in infected rivers is reduced by an average of 87%. Total yearly loss in a river fishery caused by *G. salaris* is estimated to be about 45 tons [1]. Social effects occur specially in the area of large salmon rivers with *G. salaris* due to loss of income and lost recreational fishery opportunities as the salmon populations is reduced to a very low level [109].

With these problems, it is important to have a mechanism to predict the identification of this species. In this study, the focus is on *Gyrodatylus* species that infect Atlantic salmon. Many studies have been carried out with regard to species identification. A number of existing applications of species classification and identification are discussed in the following section.

Many methods and strategies have been explored for classification and identification of true species of *Gyrodactylus*. The monogenean *Gyrodactylus* contains many individual species including *G. salaris* and *G. thymalli*. These two species are the most difficult species to distinguish between, due to having a very close physical resemblance. Shinn *et al*. [130] explain that the morphometric discrimination of *Gyrodactylus* species can be difficult due to the small size of taxonomically important structures of key features of the haptoral attachment hooks.

The main taxonomy for the recognition of the *Gyrodactylus* species is based on morphology, predominantly upon the morphometric features of the attachment hooks, as will be described in more depth in Chapter 3 and shown in Figure 3.1. Referring to Shinn *et al.* [128], the opisthaptor of the attachment hook is the main organ of attachment to the host that is used for identification of the species of the monogenean of *Gyrodactylus* Malmberg. Additionally, Bakke *et al*. [18] have stated that hamuli and ventral bars represent a remarkable taxonomic resource for carrying out species recognition and identification.

Taxonomic identification is frequently impeded by a multiplicity of problems because of the need for over-simplification and the use of morphometric descriptors. Boundaries between species are often poorly defined because of: (1) the simplicity of the forms under study; (2) the existence of few distinguishable morphometric features; (3) the natural, often large, variation of these features within the species; and (4) the overlaps between the species in morphometric space. Some pathogens cause particularly serious damage in host populations, and to minimise their effects, it is essential to find a method that unequivocally detects their presence in any given populations. It is desirable that such a method is simple, rapid and widely available for use in non-specialist laboratories [76].

In addition to this, Shinn *et al.* [125] have mentioned that to provide taxonomically useful information, the measurement of the attachment hook must be carried out precisely, despite the difficulties presented by these hooks being small in size and complex. The predominant feature for recognition of the *Gyrodactylus* species is through the ventral bar shape. In critical

species, like, *G. salaris* and *G. thymalli*, the examination of marginal hook and hamuli hooks is necessary due to the finer points of relationship of a *Gyrodactylus* [18].

In Norway, measures in hatcheries and fish farms have proved to be effective in exterminating *G. salaris*. This includes measures like construction of migration barriers and rotenone treatment, which has been proven to be effective in rivers [67]. Olstad [1] reported that by 2013, 33 of 48 infected rivers had been treated with rotenone. Among these, 20 have been declared free of the parasite. The remaining 13 were tested within the last 5 years and have not to date been officially declared parasite free. Olstad [1] also added that in recent years, treatments with acid aluminium to kill the parasite but not the host have also proven to be successful. This was tested in the river of Rover Laerdalselva in 2011 and 2012 in an attempt to get rid of the parasite without killing the host.

Routine screening for the presence of *G. salaris* as part of national surveillance monitoring programmes has the potential to generate huge volumes of samples, particularly during periods of a suspected outbreak, which must be identified rapidly and correctly. Amongst fish parasites, certain monogeneans such as species belonging to the genera *Dactylogyrus*, *Gyrodactylus* and *Benedenia*, remain a particularly intransigent economic burden for the global freshwater and marine aquaculture industries [116]. *Gyrodactylus salaris* is notoriously difficult to classify from closely related and morphologically similar species present on European salmonids. If government policy worldwide is to maintain high standards of fish health and welfare in cultured and wild stocks, it is vital to have in place techniques to facilitate the inspection and diagnosis of serious fish parasites such as *G. salaris*.

Classification of the *Gyrodactylus* species group poses several difficulties: Firstly, the accumulation of enough expert knowledge to reliably distinguish between species, say, *G. arcuatus*, *G. salaris* and so on; Secondly, manual classification is labour intensive and time consuming; and thirdly, the most formidable challenge occurs when the point to point measurements are not accurately taken, which will result in inaccurate species classifications. Finally, at the last stage of classification, only a domain expert can determine the correct species with their own vision and experience.

In earlier work, various classification morphometric techniques based on statistical classification techniques; such as Linear Discriminant Analysis, Nearest Neighbor, Feed-Forward

Neural Network and Projection Pursuit Regression [76], [104], [126] and molecular techniques; like a 18S rRNA sequence, Ribosomal RNA and mitochondrial cytochrome c oxidase I [32],[33], [105], [60] have been developed. These techniques have been proven to work successfully for species identification. Unfortunately, none of these methods can promise to identify many species by using a single specific method.

Whilst expert taxonomists may be able to classify *G. salaris* from other closely related species (using learned human expertise), the morphometric speciation and molecular characterisation of the *Gyrodactylus* species is frustrating and difficult for the reasons described previously. For this reason, this project aims to develop novel techniques involving intelligent signal image capture, processing and analysis, and employing state-of-the-art Artificial Intelligence (AI) algorithms and technologies to provide automated or semi-automated species identification.

The AI related techniques employed in this project will be discussed in greater depth in chapter 3 of this thesis. Initially, the work proposed in this research project will aim to successfully classify *G. salaris* apart from other European gyrodactylids, but with the eventual objective of developing tools that are capable of accurate generalised *Gyrodactylus* classification. It is hoped that the techniques developed as an outcome of this work will be of practical relevance for identification or classification of a range of aquatic and indeed terrestrial pathogens, and to taxonomic identification more widely.

2.6 CONCLUSIONS

This chapter provided an introduction to the problem this research project aims to solve, to identify the true species of the ectoparasites of monogenean of *Gyrodactylus* that infect Atlantic salmon. Firstly, the history of this species is introduced. Monogenean of the genus of *Gyrodactylus* have been known for almost 180 years [18]. In Norway, in 1975, *G. salaris* was reported [18] to heavily impact on salmon stocks at several fish hatcheries. However, in the UK, there have been at present no reported *G. salaris* infections. According to Shinn *et al.* [124], fish health authorities have operated and executed extensive screening programmes of several species of salmonid to avoid mortality damage from *G. salaris*.

This chapter also discussed the unique transformation of the species. *Gyrodactylus* species are the smallest species among monogenean, and carry babies during their development; being a viviparous fish parasite [18]. This type of species does not have any specific transmission stage; the viviparous worms give birth to fully grown adults, which during the birth process attach to the same host as the parent and only subsequently may transfer to a new host [1]. The *Gyrodactylus* species are only can be found on the skin. They have a very small body shape (up to 2mm in length) and can only be seen using a microscope. The main taxonomy for the recognition of the *Gyrodactylus* species is based on morphology, predominantly upon the morphometric features of the attachment hooks; marginal hook, ventral bar and hamuli. With regard to species identification, this is not an easy task since these hooks are very small in size and preparation is very time consuming. In addition, identification relies heavily on domain experts.

The wide range of the *Gyrodactylus* species was discussed in the mortality section. Gyrodactylids monogeneans are widespread parasites of fresh water and marine fishes. Olstad *et al.* [1] reported that the disease resulting from *Gyrodactylus* infections, gyrodactylosis, is responsible for losses in a wide variety of captive fish species. Finally, the techniques available for identification of the studied *Gyrodactylus* species were discussed. Domain experts are researching the best method for classification and identification of the correct species, which enables the most appropriate specific treatment to be provided for an infected river. Towards solving the problem described above, this research objective is to provide a mechanism to accurately identify the multiple species of *Gyrodactylus* by applying the state-of-the-art of AI algorithms and technology.

3

GYRODACTYLUS MORPHOMETRIC IDENTIFICATION

3.1 INTRODUCTION

As stated in chapter 1, the ultimate aim of this research is to develop an accurate prediction framework for ectoparasites of Monogenea genus of *Gyrodactylus* species, which infects freshwater Atlantic salmon, with a focus on reducing the reliance on the biologist / domain expert for identification of the correct species. Accurate identification of the species is a must in order for the appropriate chemical treatment to be provided to the infected river. Chapter 1 introduced the thesis and presented the motivation and goals of this research. Chapter 2 presented the background and description of the species that is the focus of this research, and also investigated the mortality effect of this species on the eco-system and food supply.

This chapter presents the application of machine learning classifiers and feature selection techniques for predicting multiple *Gyrodactylus* species. To improve the accuracy of identification of the correct species, an ensemble based majority voting approach has been proposed; with the application of multiple classifiers and feature selection methods. Firstly, the description of the dataset being utilised has been provided, including a review of the collection and preparation of the data. This dataset has been collected and prepared by the Parasitology team of the Institute of Aquaculture, University of Stirling.

Also covered in this chapter is the wrapper technique of feature selection. Here, it has been decided to apply three well-known methods. These are Sequential Forward Selection (SFS), Sequential Backward Selection (SBS) and Sequential Forward Floating Selection (SFFS). Feature selection has been considered because it is a process of transforming the existing features into a lower dimensional space. It will select a subset of the existing features without transformation. The, following section will then introduce machine learning classification. Various machine learning classification approaches are widely available and have been studied in various types

of application problems [23], [58], [94], [149]. The application of these techniques has been demonstrated and results presented.

As the objective of this research project is to classify and identify multiple species of *Gyrodactylus* accurately using machine learning classifiers and feature selection techniques, multiple classifiers and feature sets are considered for constructing an ensemble classification approach for *Gyrodactylus* species identification. In this study, the Linear Discriminant Analysis (LDA) and K-Nearest Neighbor (K-NN) approaches form the classifier base; while three different feature sets are considered in order to classify the multiple species. These are the 25 full feature set, 21 selected features using the SFS method and 20 selected features from the SBS method.

The remainder of this chapter covers a number of different areas. Firstly, section 3.2 describes the morphometric dataset. This dataset will be used in the demonstration of the results arising from this research project. This is followed by a review of feature selection techniques, where the wrapper feature selection approaches have been further discussed and applied to the morphometric features of the *Gyrodactylus* dataset. In the next section, machine learning classification techniques have been reviewed, along with experiments that have been conducted in applying the various machine learning methods to the *Gyrodactylus* specimens. This is followed in section 3.5 by the propose ensemble classification and feature selection methods being carried out. In this section, the proposed method for identifying multiple specimens has been proposed. This chapter is summarised in section 3.6.

3.2 MORPHOMETRIC DATASET

In this chapter, the morphometric dataset based on point to point feature extraction has been used for analysis and classification. All the specimens were collected and prepared by researchers within the Parasitology Laboratory at the Institute of Aquaculture, University of Stirling, United Kingdom. Those species were collected from various countries and places. In this study, nine species of *Gyrodactylus* of morphometric dataset were considered. There are *G. arcuatus*, *G. derjavinoides*, *G. kherulensis*, *G. gasterostei*, *G. salaris*, *G. sommervillae*, *G. thymalli*, *G. truttae* and *G. cichilidarum*.

3.2.1 *Data Collection*

A total of 557 species from nine different group of ectoparasites of *Gyrodactylus* were sampled for studies using light microscopy. The description of the sites sampled is given in Table 3.1. For this study, samples were collected from wild and farmed salmonids over a wide geographical range.

Referring to Shinn *et al.* [128], the opisthaptor of the attachment hook is the main organ of attachment to the host in identification of the species of the monogenean of *Gyrodactylus* Malmberg. In this study, the feature information was extracted from the tree part of opisthaptor attachment hooks; hamuli, ventral bar and marginal hook.

Data preparation is involved because once the data has been cleaned and processed, only then can the process of point-to-point measurement be performed. In data collection, only the authorised person who has the license is allowed to catch and process the infected fish. In this study, the preparation before the point to point measurements was carried out either by masters students or scientists from the parasitology lab of Stirling University.

Specimens of *Gyrodactylus* species were removed from their respective fish hosts and fixed in 80% ethanol until required. Individual specimens were subsequently rinsed in distilled water, transferred to a glass slide, had their posterior attachment organ excised with a scalpel and the attachment hooks released using a proteinase-K based digestion fluid (*i.e.* 100 µg/ml proteinase K (Cat. No. 4031-1, Clontech UK Ltd., Basingstoke, UK), 75 mM Tris-HCl, pH 8, 10 mM EDTA, 5% SDS). The digestion process was stopped through the addition of 3 µl of a 50:50 formaldehyde:glycerine solution. A coverslip was added to the preparation, which was sealed using a commercial nail varnish.

An image of the attachment hooks from each specimen was captured using an AxioCam MRC (Zeiss) 1.5 megapixel camera fitted with a MicroCam Olympus LB Neoplan D-V C mount 0.75× interfacing lens attached to an Olympus BX51 compound microscope. The specimens were viewed under 100× oil immersion objective using MRGrab v. 1.0.0.0.4 (Carl Zeiss Vision GmbH, Munchen, Germany) software.

The image for each specimen was then loaded into the Point-R software (ver. 1.0 © University of Stirling, 2003) running within the KS300 v3.0 image analysis environment (Carl Zeiss Vision

Table 3.1: Location of data sampling of the nine species of *Gyrodactylus*

True Species (*Gyrodactylus*)	**Site Location**	**Num. Of sites sampled**
arcuatus	L. Airthrey, Scotland	24
derjavinoides	Gd Denmark	135
gasterostei	L. Airthrey, Scotland	30
kherulensis	Gs Norway	30
salaris	Gs Norway	2
salaris	Gs Denmark	1
salaris	Finland RT fins Jyvaskyla	2
salaris	Gs Norway	6
salaris	Lierelva, Norway	30
salaris	Rauma, Norway	30
sommervillae	Blenheim Palace Lake	30
thymalli	Poland	44
thymalli	UK	45
thymalli	Rena, Norway	40
truttae	Czech Republic	50
truttae	Scotland	27
truttae	Denmark	3
cichlidarum	Thailand	14
cichlidarum	Stirling	16
cichlidarum	Philippines	12
cichlidarum	Ecuador	13
cichlidarum	Colombia	15

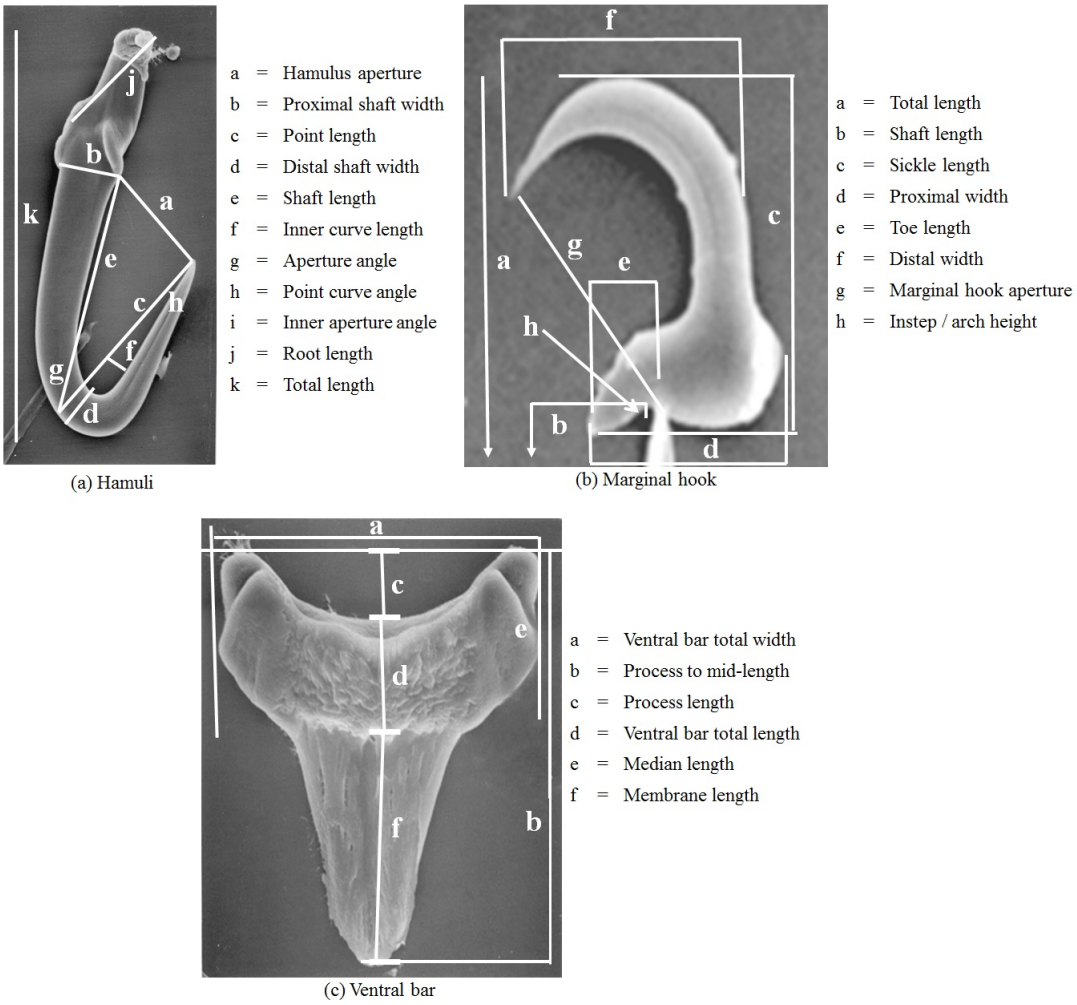


Figure 3.1: *Gyrodactylus* (Monogenea) skeleton hooks morphology; (a) hamuli, (b) marginal hook, and (c) ventral bar were measured using point-to-point measurement.

GmbH, Munchen, Germany) which permitted 25 point to point measurements to be made on the attachment hooks of each specimen. This application was used for measuring the point-to-point values of the attachment hook, such as angles etc. Only then can the classification be performed to predict the true species of *Gyrodactyllus*. Figure 3.1 illustrates the points that are measured during the point-to-point measurement procedure.

This study focuses on the discrimination of *Gyrodactylus* species commonly infecting salmonids and other key fish species which are regularly assessed as part of statutory national surveillance programmes throughout the UK. These include *G. derjavinoides* Malmberg, Collins, Cunningham *et* Jalali, 2007, *G. salaris* Malmberg, 1957, *G. thymalli* Žitňan, 1960 and *G. truttae* Gläser, 1974. In addition, species that could accidentally parasitise salmonids such as *G. arcuatus* Bychowsky, 1933 and *G. gasterostei* Gläser, 1974 were included, as well as species that

may be confused with *G. salaris*, species such as *G. lucii* Kulakovskaya, 1952. Other potentially problematic species including *G. kherulensis* Ergens, 1974 and *G. sommervillae* Turgut, Shinn, Yeomans *et* Wootten, 1999 were also considered. Finally *G. cichlidarum* Paperna, 1968, a parasite of Nile tilapia, *Oreochromis n. niloticus* (L.) was included as an outlying group. *Gyrodactylus salaris* and *G. thymalli* were identified as the most difficult to classify since they are closely related and morphologically similar. Shinn *et al*. [130] explains that the morphometric discrimination of *Gyrodactylus* species can be difficult due to the small size of taxonomically important structures *i.e.* the haptoral attachment hooks.

Morphometric data (25 point of measurements) were collected from glass slide mounted specimens prepared for light microscopy. Of the 25 features, 11 were extracted from one of the paired central hamuli (anchors), 6 from the ventral bar which spans the two hamuli, and 8 from one of the 16 peripheral marginal hooks. The 25 points are: **hamulus**: total length (HTL), point length (HPL), shaft length (HSL), root length (HRL), aperture distance (HAD), proximal shaft width (HPSW), inner angle (HIA), distal width (HDSW), inner curve length (HICL), aperture angle (HAA), point curve angle (HHPCA); **ventral bar**: total length (VBTL), total width (VBTW), process-to-mid length (VBPML), median length (VBML), process length (VBPL), membrane length (VBMBL); and **marginal hook**: total length (MHTL), shaft length (MHSHL), sickle length (MHSL), sickle proximal width (MHSPW), sickle distal width (MHSDW), sickle toe length (MHSTL), sickle aperture (MSHAD), instep height (MHIH). All 25 points are categorised as scale type data and are measured in micrometers (μm).

In this research project, only monogenea of *Gyrodactylus* species are focused on for correct identification of their true classes. In *Gyrodactylus* species, there are nine collectable species available to be used in experiments. The total number of specimens for this purpose of study is 557. The breakdown of numbers for each species is presented in Table 3.2.

3.3 FEATURE SELECTION TECHNIQUES

Feature selection is performed before classification, and is not the same as feature extraction. Feature selection is a process of transforming the existing features into a lower dimensional space. This involves selecting a subset of the existing features without transformation. A

Table 3.2: Detailed breakdown of the *Gyrodactylus* species and their number of specimens.

Species name (*Gyrodactylus*)	**Number of specimen**
G. arcuatus	24
G. derjavinoides	137
G. gasterostei	30
G. kherulensis	30
G. salaris	71
G. sommervillae	30
G. thymalli	85
G. truttae	80
G. cichilidarum	70
Total	**557**

common justification for the application of feature selection is to remove redundant features and thus improve classification performance. The problem of object classification normally involves the difficulty of extremely high dimensional feature space which sometimes makes learning algorithms intractable [43]. This section discusses the opportunities in the application of feature selection, and also defines a number of specific feature selection techniques that are used in this research project. Original experiment results of using these techniques to perform feature selection using the training database are then presented.

A standard procedure to reduce the feature dimensionality is called feature selection (FS). There are various FS methods, such as wrapper and filter techniques [4], [43], [55]. The filter method does not require the use of a classifier to select the best subset of features, while in the wrapper feature selection approach, it uses a classifier to evaluate the classification error rate as the evaluation function. These two highlighted approaches are the main categories of the available feature selection techniques, and these are divided into sub-categories that will be reviewed in further detail in the following section.

A range of FS approaches have been proposed for the task of making object classification more efficient and accurate. Given a feature set of $X = (X_p \mid p = 1...B)$, find a subset $E_C =$

$(X_{p1}, X_{p2}, ..., X_{pC})$, with $B > C$, that optimises an objective function $Z(E)$. This will maximise the ability of the system with regards to classifying object instances. Many researchers make use of FS methods because of the benefits gained from this application. FS techniques are not limited only to specific applications such as classification, but have been applied to other areas such as amongst others, bioinformatics [121], text categorisation [35] and healthcare [49]. In clinical decision support, the FS technique has also been successfully applied to automated cancer diagnosis based on histopathological images [39] and also ultrasound and mammography images [145].

There are many potential benefits gained from performing FS methods on a dataset, especially if the numbers of features present are large. According to Acuňa *et al.* [4], there are two main reasons to justify FS in order to perform classification: (1) A saving of computing time and (2) An easy interpretation of the model. Following on from these reasons, it is hoped that classification performance and accuracy will be improved. The goal of FS is to reduce dimensionality by removing redundancy and less significant features; and improving the classification rate. Certain datasets might consist partially of redundant features that will negatively affect classification performance. In general, FS methods search through subsets of features, and try to find the best one among all the competing candidate subsets according to some evaluation function.

Another proponent of the FS method is Doraisamy *et al.* [43], who found that when the best feature subset has been selected, it will result in better classification accuracy, as proven when it was implemented with regard to Traditional Malay Music (TMM). In addition, reducing the feature set before the classification stage may result in the improvement of the quality of knowledge extracted and increases the speed of computation [5]. Saeys *et al.* [121] also point out the objective of feature selection. These are: (1) to avoid overfitting and improve model performance; (2) to provide faster and more cost-effective models; and (3) to gain a deeper insight into the underlying process that generated the data.

In selecting the optimum features, the item properties may depend strongly on each other and a subset of individually bad features may prove to be rather good because of positive interaction effects. Because of this uncertainty, the only apparent way of searching for optimal subsets is simply to evaluate all the possible item combinations. However, testing

all subsets is a combinatorial problem that requires an exponential of computational time [133]. According to Dasgupta *et al.* [35], feature selection is the process of selecting a subset of features available for describing the data before applying a classification or prediction. It has been widely observed that feature selection can be a powerful tool for simplifying or speeding up computations.

FS methods depend on the way that the subsets are generated and on the evaluation function used to evaluate the subset under examination. There are many FS techniques that have been discussed and reviewed. For example, Doraisamy *et al.* [43], discuss the differences between the main categories of FS methods. The wrapper method generally provide better results than the filter method, due to the selection process being optimised for one specific classification algorithm, while the filter method works faster than the wrapper method, and this type of approach is highly recommended for high dimensional datasets that have a large number of features. For example, bioinformatic data [121].

The following section will review multiple types of feature selection techniques available, such as forward selection and correlation based feature selection. In the filter model type search, the way they evaluate the informative features is by looking only at the intrinsic properties of the data [121]. All the features are calculated and the features with the lowest score are removed. The advantages of this type of selection are that it can easily be scaled to very high-dimensional datasets, and that it is computationally fast and simple. This type of classifier is most recommended when the dataset used has a large number of features [43]. According to Gheyas & Smith [55], filter methods are fast but lack robustness against interactions among features and feature redundancy. They also added that it is not clear how to determine the cut-off point for rankings in order to select only truly important features and exclude noise.

The wrapper technique will identify good features based on the classification performance. A search procedure in the space of possible feature subsets is defined and various subsets of features are generated and evaluated. To search the space of all feature subsets, a search algorithm is then wrapped around the classification model [121]. Although it is computationally expensive to execute, it has the ability to take into account feature dependencies, unlike

the filter method. The wrapper approach is generally better than the filter technique because it uses the classification model in the evaluation [43], [55].

In addition to filter and wrapper techniques, there is also the embedded technique. This technique will search for an optimal subset of features as part of the classifier construction, and can be seen as a search in the combined space of feature subsets and hypotheses. This type of selection also relies on classification performance for evaluating the best features and is less computationally expensive to execute. However, general theoretical performance guarantees are modest and it is often difficult to claim more than a vague intuitive understanding of why a particular feature selection algorithm performs well when it does [35].

Wrapper methods can be divided into two groups based on search strategy: (1) greedy; and (2) randomised or stochastic. Greedy wrapper methods use less computer time than other wrapper approaches. Sequential Forward Selection (SFS) [26] and Sequential Backward Selection (SBS) [31] are the two most commonly used wrapper methods that use a greedy hill-climbing search strategy.

In this study, only the wrapper method has been used for experiments and implementation. The selection of wrapper method is because the accuracy of the classifier can be estimated through the selection of the feature subset. The found optimum feature subset will be used for training the classifier. As stated by Inza *et al.* [65], the wrapper approach obtains better predictive accuracy estimates than the filter approach. This fact was also supported by Hall & Smith [59], the wrappers often achieved better results due to the fact that they are tuned to the specific interaction between an induction algorithm and its training data.

There are a number of wrapper techniques, such as Forward Selection, Backward Selection, Sequential Forward Floating Selection and Sequential Backward Floating Selection. According to Somol *et al.* [134], Li *et al.* [86], this method requires one predetermined learning algorithm in feature selection, and its performance is then used to evaluate and determine which features are selected. In other words, the classifier is used to control the selection of features. Unfortunately, this method is more computationally expensive in comparison to the filter method. Figure 3.2 [78] shows the outline of how the wrapper method procedure is conducted. The feature subset selection algorithm exists as a wrapper around the induction algorithm. The feature subset algorithm conducts a search for a good result using the induction algorithm

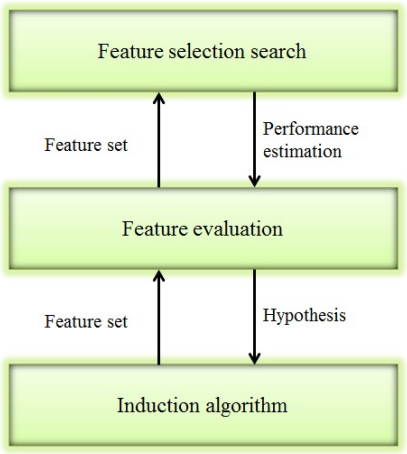


Figure 3.2: The procedure in the wrapper method for selecting informative features.

itself as part of the function evaluating feature set, while the task of the induction algorithm is to induce a classifier that will be useful for classifying future cases. In this study, the logistic regression model has been used as the induction tool for the selection of the best features for species identification. This is an iterative process, repeated until the optimum number of features has been acquired.

The selection of Logistic regression as the induction tool in feature selection is because it requires fewer restrictive assumptions [24]. The characteristic of logistic regression is that it has the ability to perform classification or selection when the distribution is not equal within the group variance covariance. As discussed previously, the number of available specimen in different species groups is unequal, therefore, the logistic regression has been considered to be more suitable. This is also supported by Komarek and Moore [80], who state that logistic regression is often faster to train than more complex models like Random Forest and SVM. In many problems it is the preferable method to deal with high dimensional data sets.

For selecting optimum features, it is not a requirement to have a unique set of feature if they result in the same accuracy using different sets of features. By this it is meant that if the set of features has been added to or removed from, and this does not cause any significant affect to the classification performance, then it will not be relevant for feature selection. The objective is to have a set of features that can boost classification performance and thus minimise the misclassification error. In the wrapper technique, a black box approach has been applied to feature selection through an induction algorithm. In the black box approach (*i.e.* no knowledge of the model is needed, just the interface), it will keep searching and evaluating the best of

features amongst the available features using a specified induction method [78]. In selection of the optimum features, a feature X is strongly relevant if removal of X alone will result in performance decrement of a classifier.

3.3.1 *Sequential Forward Selection (SFS)*

Sequential Forward Selection (SFS) starts with an empty set and greedily adds attributes one at a time. At each step, SFS adds the attribute that when added to the current set, yields the learned structure that generalises best [72]. Once a feature is added, it is never removed [55]. SFS is robust to multicollinearity problems but sensitive to feature interaction.

Kolodyazhniz *et al.* [79], Liu & Wang [89], Somol *et al.* [134] explain that SFS is an algorithm that starts with an empty set of features. The criterion in selecting a subset of features is based on the minimum value of the resulting classification errors when utilising the subset. This procedure is implemented for the classification process, where it was used for training. The selection process will be terminated once it is found that there is no more improvement in model performance that can be achieved. For example, if adding a new feature, does not increase the accuracy, then the model of selection will be stopped.

It starts from an empty set, and sequentially adds the feature A^+ that maximises $E(B_j + A^+)$ when combined with the features B_j that have already been selected.

1. Start with empty set $B_0 = \phi$.
2. Select the next best feature $A^+ = \text{argmax}_{A \notin B_j} E(B_j + A)$.
3. Update $B_j + A^+; j = j + 1$.
4. Go to 2 (repeat until the optimum set of features is acquired).

3.3.2 *Sequential Backward Selection (SBS)*

SBS starts with all attributes in the attribute sets and greedily removes them one at a time [72]. It begins with a candidate matrix and sequentially eliminates the least important row at each step until the desired number of rows remains. This explanation is also supported

by Kolodyazhniv *et al.* [79], who state that the model will start to eliminate features one by one from the full basket of features. It will only stop this selection process when the optimum features are found. SBS often finds difficulties in identifying the separate effect of each explanatory variable on the target variable. In SBS, once a feature is removed, it is removed permanently [55].

SBS starts from the full set and sequentially removes the feature A^- that least reduces the value of the objective function $F(B - A^-)$.

1. Start with full set $B_0 = A$.
2. Remove the worst feature $A^- = \text{argmax}_{A \in B_j} E(B_j - A)$.
3. Update $B_{j+1} = B_j - A^-; j = j + 1$.
4. Go to 2 (repeat until the optimum set of features is acquired).

3.3.3 *Sequential Forward Floating Selection (SFFS)*

SFFS is characterised by the changing number of features included or eliminated at different stages of the procedure [72]. SFFS works in a similar manner to the SFS, but for every new subset it enters a backtracking loop that attempts to find a better subset than that of its predecessor by removing one feature at a time. This is repeated until no better subset is found and the backtrack loop is then exited. This algorithm has been found to have near optimal results on some experiments [114]. This definition is also supported by Vervedis *et al.* [144], who describe SFFS of consisting of a forward step and a conditional backward step. This means that at every step, feature insertion and deletion will be involved throughout the process until the optimum feature subset has been achieved.

The floating search algorithm attempts to improve the feature subset after every step by means of backtracking. Consequently, the resulting dimensionality in respective intermediate stages of the algorithm is not changing monotonically but is actually floating up and down [133].

SFFS starts from an empty set. After each forward step, SFFS performs backward steps for as long as the objective function increases.

1. $B_0 = 0$.

2. Select the best feature

   $A^+ = \text{argmax}_{A \notin B_j} G(B_j + A)$.

   $B_j = B_j + A^+; j = j + 1$.

3. Select the worst feature

   $A^- = \text{argmax}_{A \in B_j} G(B_j - A)$

4. If $G(B_i - A^-) > G(B_j)$ then

   $B_{j+1} = B_j - A^-; j = j + 1$.

   Else

   Go to step 2 (repeat until the optimum set of features are obtained).

3.3.4 *Results and discussion*

Considering too many features in classification may result in difficulties in the prediction and interpretable capabilities of the model due to redundancy, non-informative features and noise [120]. Hence, it is usually necessary to apply feature selection. In general, feature selection has two components, which are the generated proposed feature subsets; and the evaluation algorithm that determines how good a proposed feature subset is.

Table 3.3 shows feature selection implemented with the original 25 features of morphometric data. As mentioned earlier, these 25 features were identified by domain experts from the School of Aquaculture in Stirling. Three feature selection techniques have been used; SFS, SBS and SFFS. Using the SFS method, instead of 25 features, SFS selects only the best 21 features to classify the *Gyrodactylus* species. On the other hand, SBS extracted 20 features. Finally, when applying the SFFS method to the original 25 features, 7 features were identified. These features are then considered for identification of the nine species of *Gyrodactylus*.

Floating search was suggested by Pudil *et al.* [114] to reduce the problems faced by SFS and SFS methods. Floating search methods such as SFFS perform greedy search, but as discussed, have additional provision for backtracking. However, a study [55] found that SFFS did not

Table 3.3: Feature selection. A wrapper method which uses Sequential Forward Selection (SFS), Sequential Backward Selection (SBS) and Sequential Forward Floating Selection (SFFS) to select new sets of morphometric features extracted from the attachment hooks of *Gyrodactylus*. The full name of each structure (abbreviation) is given in Section 3.2.

Features	**SFFS (n=7)**	**SBS (n=20)**	**SFS (n=21)**
HAD		x	
HPSW	x	x	x
HPL	x	x	x
HDSW	x	x	x
HSL			x
HICL		x	x
HAA	x	x	x
HPCA		x	x
HIA		x	x
HRL		x	x
HTL		x	x
VBTW		x	x
VBTL			x
VBPML			
VBML		x	x
VBPL		x	x
VBMBL		x	x
MHTL	x		
MHSL	x	x	x
MHSickL	x	x	x
MHSPW		x	x
MHToe		x	x
MHSDW		x	x
MHAD		x	x
MHIH			

produce superior performance to SFS because the effect of sequentially adding or removing features is that the utility of an individual feature is often not apparent on its own, but only in combinations including just the right other features, and deleting individual features without taking this into account can negatively affect results. When performing selection of 25 features of *Gyrodactylus* species with SFFS, using the 557 specimens of nine different species, only 7 features are selected as consisting of the optimum feature subset, whereas, using the SFS method, only four features are eliminated from the total number of features in the full set.

These sets of features will be used as part of experimental work with various types of machine learning for the purpose of species identification. These results will then be compared with each other in order to identify the best classifier for classifying the multiple species of *Gyrodactylus*. From there, the strength of each feature subset will be determined, by comparing results to those found using the original full set of features.

3.4 MACHINE LEARNING CLASSIFICATION

Machine learning has been utilised in this research project for Monogenea of *Gyrodactylus* species identification. Identification of the mortality of the species is a requirement in order for the specific treatment to be provided to control the spread of the species from an infected region to the entire river. Currently, there are many applications of machine learning classifier techniques available including road traffic sign detection [150], face identification [153] and bioinformatic [50]. Classification is defined as a way of grouping together objects or classes that share some properties in same class entities, while distinct classes are assigned to entities having distinct classes [32]. Classification functions by assigning an object to a specific class, where the classes or groups have already been established with the aid of a training set. Basic information is required to perform analysis using classifiers, regardless of the type of classifier to be used. The first step is to acquire the data vector, which either uses data gathered from observations or from analysis. In addition to the data vector, features are another mechanism that is required to describe the class for each data vector [147, 47].

Machine learning classification techniques, can be divided into main categories, supervised and unsupervised learning. These two categories have the same goal, which is to assign an

input feature vector $X = X_1, ..., X_n$ in D (a feature set), representing an object, to a member of the class set $Y = Y_1, ..., Y_m$. This goal can be accomplished by inducing a classifier from a given set of training examples [57].

Any classifier that has been trained first with the pattern of the problem, then processed using real data is categorised as being a supervised learning classifier [46]. Take an example of digit recognition. The objective of this classification is to assign each input vector to one of the class label; as providing using training examples. Examples of supervised learning include Linear Discriminant Analysis, K-Nearest Neighbor, and Naive Bayes.

Non-linear classifiers have non-linear, and possibly discontinous decision boundary [46], [53]. Non-linear classifier has manage to perform and provide good classification performance in many areas. According to Zhouyu *et al.* [52] the non-linear of Support Vector Machine (SVM) can handle linearly inseparable data but is not as efficient as linear classifier. This is due of their complexity with the number of support vectors. Another example of non-linear classifier is Multi-layer Perceptron (MLP).

McHugh *et al.* [104] reported on research where LDA and K-NN were used with morphometric data to discriminate the notifiable pathogen *Gyrodactylus salaris* from *G. thymalli* (Monogenea). In their analysis, when comparing between these two classifiers, LDA produced better classification performance, even though another two species (*G. derjavinoides* and *G. truttae*) were added to the system,with LDA continuing to show a higher accuracy performance in comparison to the K-NN classifier.

A statistical classification was also demonstrated by Kay *et al.* [76] for discriminating a notifiable pathogen of Atlantic salmon from its benign close relatives. In their research, they evaluated four types of classifiers: Linear Discriminant Analysis (LDA), K-Nearest Neighbor (K-NN), Feed-Forward Neural Networks (FFNN) and Projection Pursuit Regression (PPR). Among these classifiers, the K-NN was found to achieve better performance compared to the others when they were implemented using 470 specimens for discriminating between three species groups (*G. salaris*, *G. derjavinoides* and *G. truttae*).

For this project, it was decided to evaluate four different classification approaches to perform species identification using the dataset of 557 specimens from the nine different species of *Gyrodactylus* with the original feature set of 25 features. The four chosen classifier

types are LDA, K-NN, MLP, and SVM. The selection of these classifiers was according to the performance shown in previous experiments using the *Gyrodactylus* dataset. While the SVM is a relatively recent model, it was found to have good performance [19], [57]. These classifiers are then comparatively evaluated to define the best classifier for minimising the classification error with regard to species determination. Next, the four classifiers used in this study are reviewed in further detail.

3.4.1 *Linear Discrimination Analysis*

This is the standard and the oldest method used for classification. Linear Discriminant Analysis (LDA) [115] is a method used in statistical and machine learning techniques to find a linear combination of features which best characterise or separate two or more classes of objects or events. LDA is trained using continuous feature variables from different classes of items to highlight aspects that distinguish the classes, and uses these measurements to classify new items [93].

The purpose of LDA in the context of this research is to classify objects (*Gyrodactylus* specimen) into one or more classes based on a set of features (11 features extracted from hamulus, six from ventral bar, and 16 from marginal hook) that describe the objects (*G. arcuatus*, *G. derjavinoides*, *G. gasterostei*, *G. kherulensis*, *G. salaris*, *G. sommervillae*, *G. thymalli*, *G. trutte* and *G. cichilidarum*). In general, an object is assigned according to a set of features that have unique discrimination characteristics. LDA is highly recommended [147], [113] when the classes are linearly separable. Linearly separable suggests that the classes can be separated by a linear combination of features that describe the objects. The LDA score or discriminant function of an observation $\hat{N}$ is given by

$$\hat{N}_i = W_0 + W_{i1}D_1 + W_{i2}D_1 + ... + W_{in}D_n \tag{3.1}$$

The subscript i denotes the respective class, while subscript $1, 2, ..., n(p)$ denotes the n features. $\hat{N}_i$ is constant for the i^{th} class. D_p is the observed value for the respective case for the p^{th} feature. $\hat{N}_i$ is the prediction classification score.

It will compute the mean of each dataset and the mean of entire dataset. Let D_1 and D_2 be the mean of set 1 and set 2 respectively and D_3 be the mean of the entire data, which is obtained by merging sets 1 and 2. Then, within-class and between-class are used as the criteria for class separability.

The basic idea of LDA is to find the linear transformation that best discriminates between classes, and the classification is then performed in the transformed space based on some metric such as Eucidean distance. Mathematically, a typical LDA implementation is carried out via scatter matrix analysis [87].

If the number of classes is more than two, then a natural extension of Fisher Linear Discriminant exists using multiple discriminant analysis. As in the two class case, the projection is from high dimensional space to a low dimensional space and the transformation still aims to maximize the ratio of intra-class scatter to the inter-class scatter. But unlike the two-class case, the maximization should be performed among several competing classes.

For performing classification using the LDA model, three components are involved. These are known as the test set, the training dataset, and the classification labels. The procedure involved is to classify each row of the data in the test set (with each row corresponding to a single row of features, into one of the classification labels contained in the training set using the trained classifier. Both the test and training datasets must be matrices with the same number of columns (*i.e.* the same number of features). The classification label is an example of a grouping variable used for training. Groups are defined by unique classification values, with each element in the training set corresponding to a labelled group. When using the test dataset, the classifier will make use of the features to determine which classification label each row of the test data matrix is assigned to, based on one of the groups identified during training. This implementation is based on the Statistical Toolbox in MATLAB.

3.4.2 *K-Nearest Neighbours*

The K-Nearest Neighbor (K-NN) classification algorithm is the simplest method and categorised as a lazy-learning algorithm [107], where it delays the induction or generalisation process until classification is performed. Kotsiantis [81] stated that K-NN is based on the principle

that instances within a dataset will generally exist in close proximity to other instances that have similar properties. The way classes are determined is by observing the class of its nearest neighbours [135], [62].

K-NN finds the K nearest neighbour and uses a majority vote to determine the class label. The way K is determined is through prediction by testing each K value one by one. From the list of K, the K with the maximum accuracy is then selected. The training data are computed first, the similarities of one sample from the testing data to the K-NN can then be calculated according to the class of the neighbours [34]. In this study, $k = 3$ is identified as the best nearest neighbour for predicting species class, while Euclidean distance has been chosen for calculating the distance.

Suppose that two vectors X_p and X_u, $X_t = X_p^1, X_p^2, X_p^m$, $X_u = X_u^1, X_u^2, X_u^m$, the distance between X_p and X_u is

$$(X_p, X_u) = \sqrt{\sum_{mm=1}^{m} (X_p^{mm} - X_u^{mm})^2} \tag{3.2}$$

Referring to Cunningham *et al.*, [34], the advantage of K-NN classifiers is their robustness to noisy training data and that is why most many recognition systems, such as offline handwritten signature identification [131], classification for unbalanced dataset [152] and analysing received signals [13] use this type of classifier in their analysis. Yazdani *et al.* [148] mentioned that K-NN is a good tool for dealing with problems in which the number of classes are larger than two.

The procedure for performing K-NN classification is similar to the LDA approach. This type of machine learning classifier is implemented using the Statistical Toolbox in MATLAB. It also involves the same three components as in LDA; test and training datasets, and the classification labels. In this model, the classifier assigns each row of the test dataset to a particular label, based on the labels determined by the labelled training data. Again, similar to the LDA approach, the training and test datasets must be matrices with the same number of columns. The classifier will assign each row of the test data matrix to the classification label that corresponds to the closest match, based on the training data. The output of this classification will provide a classification label for each row of the test dataset.

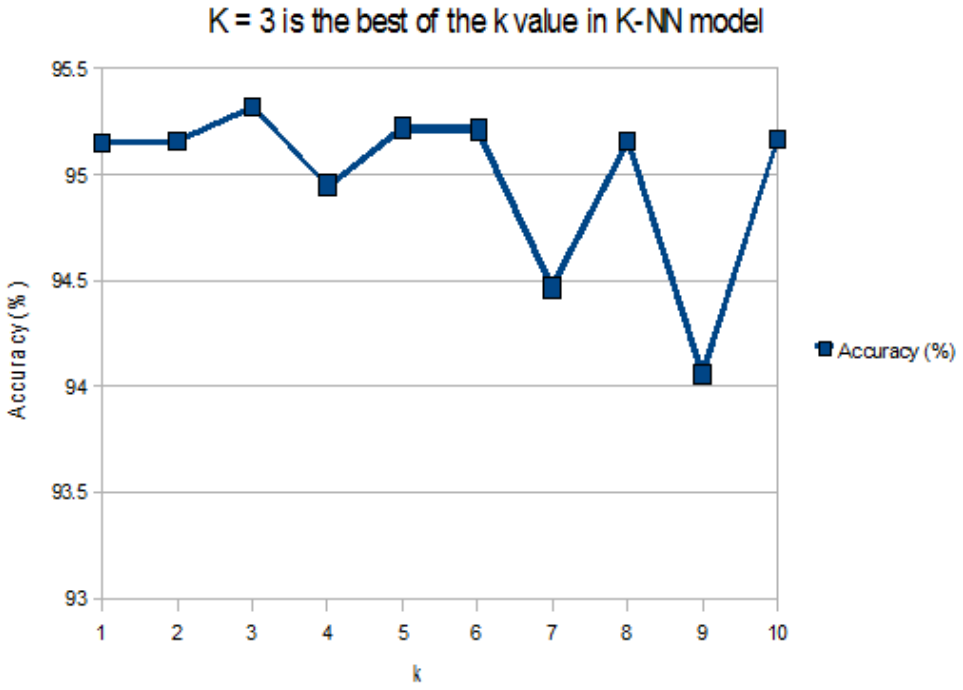


Figure 3.3: K = 3 has been identified to be the best *k* value in the K-NN model.

The selection of k = 3 is due to preliminary experimental results. As shown in Figure 3.3, a number of different values for k were experimented with for classifying nine different species of *Gyrodactylus* using the full 25 feature set. It can be seen Figure 3.3 that the highest average value was found when k = 3. It was therefore decided that in further experiments and analysis in the remainder of this research, it would be appropriate to consistently use the K-NN with the value of k set to 3.

3.4.3 *Multi-layer perceptron*

The multi-layer perceptron (MLP) is one of the most popular types of Artificial Neural Network classifier (ANN). An ANN is a set of connection models inspired by the behaviour of the human brain. It is a mathematical or computational model that tries to simulate the structure or functional aspects of biological neural networks. An ANN consists of an interconnected group of artificial neurons, and processes information using a connectionist approach to computation [140]. This type of classifier is a kind of non-linear statistical data modelling tool. It can be used with a complex model to find patterns or / of data. In a MLP, the neurons are

grouped in layers (an input layer, one or more hidden layers, and an output layer) and only forward connections are involved. A MLP provides a powerful base-learner, with advantages such as non-linear mapping and noise tolerance. This type of classifier has been increasingly used in data mining due to its good behaviour in terms of prediction of objects.

In order to classify objects, MLP uses the backpropagation function. Backpropagation is a like mapping system for organising the input/output of the objects. For a pp dimensional input vector and a qq dimensional output vector, the MLP input/output relationship defines a mapping from a pp dimensional Euclidean space to a qq dimensional Euclidean output space, which is infinitely continuously differentiable [139].

MLP are formed by $\check{Z}$ number of neurons in the hidden layer, with V_{rs} representing the weight between the neuron $\check{B}$ (hidden layer) and the neuron $\check{R}$ (output layer). V_{sc} is the weight between the neuron $\check{A}$ (input layer) and the neuron $\check{B}$, $\varphi\check{R}$ is the non-linear activation function in the output layer and φs is the non-linear activation function in the hidden layer. In sample p, the input vector is $X_p = (X_{1p}, X_{2p}, ..., X_{Bp})$, the MLP output of the neuron is $\check{R}$ (where $\check{R} = 1, 2, ..., k$); and these are expressed through [139]:

$$\hat{R}_j(X_p) = \varphi_{\check{R}} \left(\sum_{s=0}^{z} V_{rs} \varphi_s \left(\sum_{c=0}^{zz} V_{sc} D_{cn} \right) \right) \tag{3.3}$$

Albuquerque *et al.* [36] discusses the advantages and disadvantages of the MLP algorithm. This type of non-linear classifier, offers a reduction in segmentation time and promises a higher quality of results. However, the drawback with this algorithm is that it is time-consuming, especially in scenarios where the initial weights are randomly defined, which may result in considerable training time if a lot of training is needed.

This type of classifier uses back propagation to classify instances. The network can be monitored during training. The nodes in this network are all unthresholded units because the classification label is numeric. Table 3.4 shows the parameter settings used in this work for MLP based classification. The implementation and running of the MLP uses WEKA [101].

The MLP model we use involves three types of layers; an input layer for features, an output layer for classification, and a hidden layer. The number of neurons in the hidden layer for this model is based on the number of features and classes. In this case, if 25 features is given as the input layer and nine species or classifications as the output layer, then the hidden layer is

Table 3.4: Parameter settings for MLP classification model

Parameter	**Value**
Hidden layer	number of features + number of species
Learning rate	0.3
Momentum	0.2
Training time	500
Validation threshold	20

the sum of these two layers. The learning rate parameter represents the amount the weights are updated during the training process. Momentum is applied to the weights during the updating process. Another parameter shown in Table 3.4 is the validation of the threshold. It is used to terminate validation testing. The value here dictates how many times in row the validation set error can get worse before training is terminated.

3.4.4 *Support Vector Machine*

The Support Vector Machine (SVM) was introduced by Vladmir Vapnik in early 1970 [112]. This type of classifier was said to outperform many well other known classification algorithms [57], [66]. SVM are based on the Structural Risk Minimisation (SRM) principle, and their goal in the context of this research is to produce a model which predicts a species classification of data instances in the testing set when are given only the input features [57], [94]. From the definition given by Vapnik, a SVM has support vectors, which are the data points that lie closest to the decision margin (hyperplane). They have a direct bearing on the optimum location of the decision margin. Given labelled training data, the algorithm outputs an optimal hyperplane which can be used to categorise new examples. The operation of the SVM algorithm is based on finding the hyperplane that gives the minimum distance to the training examples.

A trained SVM has a scoring function which computes a score for a new input. The following equation is the scoring function that is used to compute the score for an input vector x,

$$\sum_{i=1}^{ii} Y^i \check{K}(X^i, X) + \check{b} \tag{3.4}$$

where X^i, Y^i represents the ith training example that consists of the features and class information. Here, $\check{K}$ is what is known as a kernel function, while $\check{b}$ is the scalar value. In this study, it was decided to use the polykernel or polynominal function [66], since it is a standard kernel function in SVM implementation.

In the same way as the MLP, classification with the SVM is implemented using WEKA [101]. In WEKA, the SVM model implements sequential minimal optimisation by John Platt [105] for training a SVM. The parameter setting of the SVM model is defined in Table 3.5.

Table 3.5: Parameter setting for SVM classification model

Parameter	**Value**
Complexity	1.0
Kernel	Polykernel

The complexity parameter is used to build the hyperplane between any target classes, which can be used for classification. Intuitively, a good separation is achieved by the hyperplane that has the largest distance to the nearest training data points of any class. While, as stated above, the type of kernel used in the SVM model is the polykernel function.

3.4.5 *Evaluation strategy*

Typically the data would be divided or allocated into three subsets, *i.e.* training, testing and validation data, however, given the unbalanced number of specimens per species, we abandon the traditional training strategy in favour of 10-fold cross validation which has been demonstrated to be an appropriate approach under such circumstances (see Refaeilzadeh *et al.* [117]). Here, the samples were randomly divided into $K(10)$ subsets, where $k-1$ subsets were

used for training and the remaining subset was used as the test set. This process was repeated 10 times with a different test set being used on each run, and then an average classification performance being computed. For statistical classification, 10-fold cross validation is applied by $10-\mathtt{fold} = \mathtt{accuracy}/k$, where the accuracy is the number of correctly classified samples in k experiments.

3.4.5.1 *Overall accuracy*

The overall accuracy is calculated in order to define the proportion of the total number of predictions that were correct. Given the following example:

- a is the number of true negative predictions;
- b is the number of false positive predictions;
- c is the number of false negative predictions;
- d is the number of true positive predictions;

the prediction accuracy can be obtained from this matrix as follows; $\mathtt{accuracy} = \frac{a+d}{a+b+c+d}$.

3.4.5.2 *Confusion matrix*

The confusion matrix table shows the class distribution for each truth class. It contains information about actual and predicted classifications performed by a classification system. The performance for such system is evaluated using the data in the matrix. A confusion matrix of size nxn associated with a classification shows the predicted and actual classification, where n is the number of different classes present in the data [51], [132]. Table 3.6 shows the entities involved in the confusion matrix table of two class problems.

There are two possible predicted classes: "yes" and "no". For example, in predicting the presence of disease, "yes" would mean they have the disease and "no" would mean they do not have the disease. The classifier made a total of 165 predictions. Which mean the 165 patients were being tested for the presence of that disease. Out of those 165 cases, the classifier predicted "yes" 110 times and "no" 55 times.

The basic terms of the confusion matrix can then be identified as follows. True Positive (TP) is where there are cases in which "yes" is predicted and the patient has the disease

Table 3.6: Example of a confusion matrix for a two class problem.

	Predicted NO	Predicted YES	SUM
Actual NO	TN = 50	FP = 10	60
Actual Yes	FN = 5	TP = 100	105
	55	110	165

(therefore detecting the presence of disease correctly). The True Negative value (TN) is where the classifier predicts as "no" and the patient does not have the disease (again, detecting no disease correctly). The False Positive (FP) value is where the predicted value is "yes", but the patient does not have the disease (meaning that the prediction is incorrect). Similarly, the False Negative (FN) value is where the output is predicted to be "no", but the patient has the disease (again, an incorrect prediction).

3.4.5.3 *Other performance criteria*

In this work, while the optimization of an algorithm is obviously of interest, we are not focused on finding the optimisim for one single algorithm, but looking at the classification of multiple types of species, which is slightly different from time based optimisation. In this work, we have therefore focused on this, as precision is much more important than training time in the context of this research. The criteria we have made use of as a comparison are precision and recall, as discussed:

- Precision it is also referred to as the true positive rate is a measure indicating the probability that the classifier has labelled a prediction into class A given that the ground truth is class A. The precision is the proportion of the predicted positive cases that were correct. It is defined by $\mathtt{precision} = \mathsf{TP}/(\mathsf{TP} + \mathsf{FP})$.

- Recall or user accuracy, is a measure indicating the probability that a prediction is class A given that the classifier has labelled it as being class A. Recall is the proportion of

positive cases that were correctly identified. Recall may also be referred to as sensitivity, and corresponds to the true positive rate. It is defined by $\mathtt{recall} = \mathsf{TP}/(\mathsf{TP} + \mathsf{FN})$.

3.4.6 *Results and discussions*

The morphometric point to point measurement dataset of *Gyrodactylus* specimen from nine different classes discussed in Section 3.2 has been used for all the results discussed in this section. Three parts of the skeleton hook morphology (*e.g.* hamuli, marginal hook and ventral bar) have been used as features. For extracting these features, the point-to-point measurement approach discussed previously has been applied. Then, using this set of features (25 features), classification has been performed for predicting the true class of multiple species of *Gyrodactylus* (*G. arcuatus*, *G. derjavinoides*, *G. gasterostei*, *G. kherulensis*, *G. salaris*, *G. sommervillae*, *G. thymalli*, *G. truttae* and *G. cichilidarum*).

The 557 examples of morphometric point to point measurement data from nine classes has been used in all the following feature selection and classification experiments. Three types of feature selection approaches have been chosen for selecting the best features to correctly predict the *Gyrodactylus* species. In addition, for the classification strategy, four machine learning classifiers have been used for classifying the 557 specimens with four different feature sets. These results will be compared to each other to define the best model to use for identification of the species.

To determine the classifier performance with four different feature sets, as discussed before, we use two classifiers. These are, Linear Discriminant Analysis (LDA), K-Nearest Neighbor (K-NN), Multi-Layer Perceptron (MLP) and Support Vector Machine (SVM). Each of these classifiers will use three different feature sets for classifying the multiple species of *Gyrodactylus*. The classification performance is then assessed classification accuracy and confusion matrices to determine the best model for species identification.

Table 3.7 summarises the results of the individual classifiers with the different feature sets. Classification is first performed using the original feature set of 25 features. To investigate and attempt to achieve the highest accuracy by reducing the classification error, feature selection has been applied to the original features (as discussed in Section 3.3). These results are

Table 3.7: Average of species identification between individual classification.

Feature set	Individual classifier			
	LDA (%)	K-NN (%)	MLP (%)	SVM (%)
Original feature (25F)	96.38 ±1.95	95.32 ±2.71	97.67 ±2.33	96.41 ±2.34
SFS (21F)	96.74 ±1.69	95.34 ±2.55	96.59 ±2.37	96.59 ±2.16
SBS (20F)	96.55 ±1.95	93.89 ±2.11	97.13 ±2.17	96.95 ±2.12
SFFS (7F)	94.81 ±2.57	91.71 ±4.35	95.44 ±2.41	79.13 ±2.47

compared for the best *Gyrodactylus* species identification performance. The three feature sets tested consist of the full 25 features (the original features from point-to-point measurement extraction), 21 features (selected features from the SFS method) and 20 features (selected features from the SBS method), and seven features only (selected from the SFFS method).

In addition to the accuracy in classifying the species, we are also interested in exploring the misclassification error through the confusion matrix. The following tables (Tables 3.8 to 3.23) are the confusion matrices of four feature sets for each classifier that were implemented with data from the *Gyrodactylus* fish parasite species. For LDA, the highest rate of classification was achieved using 21 features, which correctly allocated 96.41% of specimens to their true class (Table 3.12). The K-NN classifier, by comparison, also performed well using 21 features, correctly classifying 94.97% (see Table 3.13). The two non-linear methods MLP and SVM were also able achieve high rates of correct classification. Among the four different feature sets, MLP with 25 original features has achieved the highest classification at 97.67% (Table 3.10). This rate is also the highest accuracy achieved amongst all compared models. Even the SVM classifier only has 96.95% (Table 3.19), when considering with 20 features. It is not a surprise that the MLP classifier with 25 features has achieved the highest accuracy, since it is a power classifier model and this method has excellent results in many fields [28], [70], although it is time-consuming in execution as it need to train the initial weight set [36].

Although the LDA method is the oldest and the simplest classifier considered here, its performance in correctly classifying the *Gyrodactylus* specimen is impressive. LDA is, of course, only applicable when labelled training data exists and the classes are linearly separated. Table 3.8, Table 3.9, Table 3.11, Table 3.13, Table 3.14, Table 3.15, Table 3.16, Table 3.17 and Table 3.18

show the confusion matrices generated by the LDA and K-NN classifiers respectively using 25, 21 and 20 features respectively.

After consideration and comparison of the SFS, SBS, and SFFS methods, the results showed that although SFFS identified the least features (seven features were identified, compared to a much larger number using the other methods), it also produced the poorest classification results. Tables 3.20, 3.21, 3.22, and 3.23 present the confusion matrices of results achieved using the features identified using the SFFS approach with the same classifiers as used for the other feature selection approaches. It can clearly be seen in these results that although it has managed to reduce the number of key features to seven, the classifier performances have dropped, leading to poorer species identification results. This shows that SFFS is clearly not superior to the other feature selection approaches covered in this thesis.

The consideration of FS in multiple species is not easy to predict well. Certain classifier managed to well classify certain species with different feature sets. Therefore the varieties of the number of features identified by feature selection are investigated. Certain features obscure the boundaries between species and could be rejected from future analysis, therefore justifying the use of FS. Table 3.2 suggests that the features of VBPML and MHIH (see Section 3.3), contribute almost nothing to the separation of species as these two features were not used by any of the feature selection methods.

Table 3.24 presents a summary of the correct identification of the multiple species of *Gyrodactylus*. Among the nine species, three of them remain misclassified (not achieving full classification). None of the experimental models manage to provide full identification of *G. derjavinoides*, *G. thymalli* and *G. truttae*. And surprisingly, the focused species [76], [104], [130], [129], *G. salaris*, manages to achieve true classification using the MLP classifier when considering the 25, 21, 20 and 7 features. While the K-NN classifier performance is also good for classifying the focused species, where it manages to classify correctly with the 21 features. The SFFS method with 7 features is the worst feature selection method for classifying the nine species of *Gyrodactylus*. Reducing the number of features has clearly eliminated important attributes for discriminating between the true species.

Due to the remaining species that are not fully classified, an ensemble method is proposed in the following section. An ensemble is proposed motivated by the requirement to provide

	G. art	*G. cic*	*G. der*	*G. gas*	*G. khe*	*G. sal*	*G. som*	*G. thy*	*G. tru*	**Sum**	**Precision (%)**
G. art	**24**	0	0	0	0	0	0	0	0	24	100
G. cic	0	**70**	0	0	0	0	0	0	0	70	100
G. der	0	0	**129**	0	0	1	0	1	6	137	94.16
G. gas	0	0	0	**30**	0	0	0	0	0	30	100
G. khe	0	0	0	0	**30**	0	0	0	0	30	100
G. sal	0	0	0	0	0	**67**	0	3	1	71	94.37
G. som	0	0	0	0	0	0	**30**	0	0	30	100
G. thy	0	0	1	0	0	2	0	**81**	1	85	95.29
G. tru	0	0	3	0	0	1	0	0	**76**	80	95
Sum	24	70	133	30	30	71	30	85	84	557	
Recall (%)	100	100	96.99	100	100	94.37	100	95.29	90.48		

Table 3.8: LDA classifier with the 25 morphometric features.

	G. art	*G. cic*	*G. der*	*G. gas*	*G. khe*	*G. sal*	*G. som*	*G. thy*	*G. tru*	**Sum**	**Precision (%)**
G. art	**24**	0	0	0	0	0	0	0	0	24	100
G. cic	0	**70**	0	0	0	0	0	0	0	70	100
G. der	0	0	**125**	1	0	2	0	0	9	137	91.24
G. gas	0	0	1	**29**	0	0	0	0	0	30	96.67
G. khe	0	0	0	0	**30**	0	0	0	0	30	100
G. sal	0	0	0	0	0	**70**	0	1	0	71	98.59
G. som	0	0	0	0	0	0	**29**	1	0	30	96.67
G. thy	0	0	1	0	0	5	0	**78**	1	85	91.76
G. tru	0	0	2	0	1	0	0	0	**77**	80	96.25
Sum	24	70	129	30	31	77	29	80	87	557	
Recall (%)	100	100	96.90	96.67	96.77	90.91	100	97.5	88.51		

Table 3.9: K-NN classifier with the 25 morphometric features.

	G. art	*G. cic*	*G. der*	*G. gas*	*G. khe*	*G. sal*	*G. som*	*G. thy*	*G. tru*	**Sum**	**Precision (%)**
G. art	**24**	0	0	0	0	0	0	0	0	24	100
G. cic	0	**70**	0	0	0	0	0	0	0	70	100
G. der	0	0	**131**	0	0	2	1	0	3	137	95.62
G. gas	0	0	0	**30**	0	0	0	0	0	30	100
G. khe	0	0	0	0	**30**	0	0	0	0	30	100
G. sal	0	0	0	0	0	**71**	0	0	0	71	100
G. som	0	0	0	0	0	0	**29**	0	1	30	96.67
G. thy	0	0	2	0	0	2	0	**80**	1	85	94.12
G. tru	0	0	1	0	0	0	0	0	**79**	80	98.75
Sum	24	70	134	30	30	75	30	80	84	557	
Recall (%)	100	100	97.76	100	100	94.67	96.67	100	94.05		

Table 3.10: The 25 features implemented with the MLP classifier.

	G. art	*G. cic*	*G. der*	*G. gas*	*G. khe*	*G. sal*	*G. som*	*G. thy*	*G. tru*	**Sum**	*Precision (%)*
G. art	**24**	0	0	0	0	0	0	0	0	24	100
G. cic	0	**70**	0	0	0	0	0	0	0	70	100
G. der	0	0	**126**	0	0	2	0	0	9	137	91.97
G. gas	0	0	0	**30**	0	0	0	0	0	30	100
G. khe	0	0	0	0	**30**	0	0	0	0	30	100
G. sal	0	0	0	0	0	**69**	0	2	0	71	97.18
G. som	0	0	0	0	0	0	**30**	0	0	30	100
G. thy	0	0	1	0	0	2	0	**81**	1	85	95.29
G. tru	0	0	3	0	0	0	0	0	**77**	80	96.25
Sum	24	70	130	30	30	73	30	83	87	557	
Recall (%)	100	100	96.92	100	100	94.52	100	97.59	88.51		

Table 3.11: The SVM classifier with the 25 original feature set.

	G. art	*G. cic*	*G. der*	*G. gas*	*G. khe*	*G. sal*	*G. som*	*G. thy*	*G. tru*	**Sum**	**Precision (%)**
G. art	**24**	0	0	0	0	0	0	0	0	24	100
G. cic	0	**70**	0	0	0	0	0	0	0	70	100
G. der	0	0	**129**	0	0	1	0	1	6	137	94.16
G. gas	0	0	0	**30**	0	0	0	0	0	30	100
G. khe	0	0	0	0	**30**	0	0	0	0	30	100
G. sal	0	0	0	0	0	**67**	0	3	1	71	94.37
G. som	0	0	0	0	0	0	**30**	0	0	30	100
G. thy	0	0	1	0	0	0	0	**82**	2	85	96.47
G. tru	0	0	3	0	0	0	0	0	**77**	80	96.25
Sum	24	70	133	30	30	68	30	86	86	557	
Recall (%)	100	100	96.99	100	100	98.53	100	95.35	89.53		

Table 3.12: Confusion metric of classification using the LDA classifier with 21 selected features.

	G. art	*G. cic*	*G. der*	*G. gas*	*G. khe*	*G. sal*	*G. som*	*G. thy*	*G. tru*	**Sum**	**Precision (%)**
G. art	**24**	0	0	0	0	0	0	0	0	24	100
G. cic	0	**70**	0	0	0	0	0	0	0	70	100
G. der	0	0	**125**	2	0	2	0	0	8	137	91.24
G. gas	0	0	0	**29**	0	0	0	0	1	30	96.67
G. khe	0	0	0	0	**30**	0	0	0	0	30	100
G. sal	0	0	0	0	0	**71**	0	0	0	71	100
G. som	0	0	0	0	0	0	**29**	1	0	30	96.67
G. thy	0	0	1	0	0	5	0	**78**	1	85	91.76
G. tru	0	0	4	0	1	0	0	0	**75**	80	93.75
Sum	24	70	130	31	31	78	29	79	85	557	
Recall (%)	100	100	96.15	93.55	96.77	91.03	100	98.73	88.23		

Table 3.13: K-NN classifier with the 21 morphometric features, selected using the SFS method.

	G. art	*G. cic*	*G. der*	*G. gas*	*G. khe*	*G. sal*	*G. som*	*G. thy*	*G. tru*	**Sum**	Precision (%)
G. art	**24**	0	0	0	0	0	0	0	0	24	100
G. cic	0	**70**	0	0	0	0	0	0	0	70	100
G. der	0	0	**129**	0	**2**	0	0	0	6	137	94.16
G. gas	0	0	0	**30**	0	0	0	0	0	30	100
G. khe	0	0	1	0	**29**	0	0	0	0	30	96.67
G. sal	0	0	0	0	0	**71**	0	0	0	71	100
G. som	0	0	0	0	0	0	**29**	0	1	30	96.67
G. thy	0	0	1	0	0	0	0	**83**	1	85	97.65
G. tru	0	0	0	4	0	**2**	1	0	**73**	80	91.25
Sum	24	70	131	34	31	73	30	83	81	557	
Recall (%)	100	100	98.47	88.23	93.23	97.26	96.67	100	90.12		

Table 3.14: Confusion matrix of the nine species of *Gyrodactylus* implemented using the MLP classifier with 21 features.

	G. art	*G. cic*	*G. der*	*G. gas*	*G. khe*	*G. sal*	*G. som*	*G. thy*	*G. tru*	**Sum**	**Precision (%)**
G. art	**24**	0	0	0	0	0	0	0	0	24	100
G. cic	0	**70**	0	0	0	0	0	0	0	70	100
G. der	0	0	**128**	0	0	1	0	1	7	137	93.43
G. gas	0	0	0	**30**	0	0	0	0	0	30	100
G. khe	0	0	1	0	**29**	0	0	0	0	30	96.67
G. sal	0	0	0	0	0	**70**	0	1	0	71	98.59
G. som	0	0	0	0	0	0	**30**	0	0	30	100
G. thy	0	0	1	0	0	0	0	**83**	1	85	97.65
G. tru	0	0	6	0	0	0	0	0	**74**	80	92.5
Sum	24	70	136	30	29	71	30	85	82	557	
Recal (%)	100	100	94.12	100	100	98.59	100	97.65	90.24		

Table 3.15: Confusion matrix for the SVM classifier, when implemented using 21 features.

	G. art	*G. cic*	*G. der*	*G. gas*	*G. khe*	*G. sal*	*G. som*	*G. thy*	*G. tru*	**Sum**	**Precision (%)**
G. art	**24**	0	0	0	0	0	0	0	0	24	100
G. cic	0	**70**	0	0	0	0	0	0	0	70	100
G. der	0	0	**129**	0	0	2	0	0	6	137	94.16
G. gas	0	0	0	**30**	0	0	0	0	0	30	100
G. khe	0	0	0	0	**30**	0	0	0	0	30	100
G. sal	0	0	0	0	0	**68**	0	2	1	71	95.77
G. som	0	0	0	0	0	0	**30**	0	0	30	100
G. thy	0	0	1	0	0	0	0	**82**	2	85	96.47
G. tru	0	0	3	0	0	1	0	0	**76**	80	95
Sum	24	70	133	30	30	71	30	84	85	557	
Recall (%)	100	100	96.99	100	100	95.77	100	97.62	89.41		

Table 3.16: LDA classifier with the 20 morphometric features.

	G. art	*G. cic*	*G. der*	*G. gas*	*G. khe*	*G. sal*	*G. som*	*G. thy*	*G. tru*	**Sum**	**Precision (%)**
G. art	**24**	0	0	0	0	0	0	0	0	24	100
G. cic	0	**70**	0	0	0	0	0	0	0	70	100
G. der	0	1	**125**	0	0	2	0	0	9	137	91.24
G. gas	0	0	1	**28**	0	0	0	0	1	30	93.33
G. khe	0	0	0	0	**30**	0	0	0	0	30	100
G. sal	0	0	1	0	0	**67**	0	3	0	71	94.37
G. som	0	0	0	0	0	0	**29**	1	0	30	96.67
G. thy	0	0	1	0	0	6	0	**77**	1	85	90.59
G. tru	0	0	6	0	1	0	0	0	**73**	80	91.25
Sum	24	71	134	28	31	75	29	81	84	557	
Recall (%)	100	98.59	93.28	100	96.77	89.33	100	95.06	86.90		

Table 3.17: K-NN classifier with the 20 morphometric features.

	G. art	*G. cic*	*G. der*	*G. gas*	*G. khe*	*G. sal*	*G. som*	*G. thy*	*G. tru*	**Sum**	Precision (%)
G. art	**24**	0	0	0	0	0	0	0	0	24	100
G. cic	0	**70**	0	0	0	0	0	0	0	70	100
G. der	1	0	**130**	0	0	**2**	0	0	4	137	94.89
G. gas	0	0	0	**30**	0	0	0	0	0	30	100
G. khe	0	0	0	0	**30**	0	0	0	0	30	100
G. sal	0	0	0	0	0	**71**	0	0	0	71	100
G. som	0	0	0	0	0	0	**29**	0	1	30	96.67
G. thy	0	0	1	0	0	**2**	0	**81**	1	85	95.29
G. tru	0	0	3	0	1	0	0	0	**76**	80	95
Sum	25	70	134	30	31	75	29	81	82	557	
Recall (%)	96	100	97.01	100	96.77	94.67	100	100	92.68		

Table 3.18: MLP classifier with the 20 morphometric features.

	G. art	*G. cic*	*G. der*	*G. gas*	*G. khe*	*G. sal*	*G. som*	*G. thy*	*G. tru*	**Sum**	Precision (%)
G. art	**24**	0	0	0	0	0	0	0	0	24	100
G. cic	0	**70**	0	0	0	0	0	0	0	70	100
G. der	0	0	**132**	0	0	1	0	1	3	137	96.35
G. gas	0	0	0	**30**	0	0	0	0	0	30	100
G. khe	0	0	0	0	**30**	0	0	0	0	30	100
G. sal	0	0	0	0	0	**66**	0	5	0	71	92.96
G. som	0	0	0	0	0	0	**30**	0	0	30	100
G. thy	0	0	1	0	0	1	0	**82**	1	85	96.47
G. tru	0	0	4	0	0	0	0	0	**76**	80	95
Sum	24	70	137	30	30	68	30	88	80	557	
Recall (%)	100	100	96.35	100	100	97.06	100	93.18	95		

Table 3.19: SVM classifier with the 20 morphometric features.

	G. art	*G. cic*	*G. der*	*G. gas*	*G. khe*	*G. sal*	*G. som*	*G. thy*	*G. tru*	**Sum**	**Precision (%)**
G. art	**24**	0	0	0	0	0	0	0	0	24	100
G. cic	0	**70**	0	0	0	0	0	0	0	70	100
G. der	0	0	**123**	0	0	1	0	1	12	137	89.78
G. gas	0	0	0	**30**	0	0	0	0	0	30	100
G. khe	0	0	0	0	**30**	0	0	0	0	30	100
G. sal	0	0	0	0	0	**66**	1	4	0	71	92.96
G. som	0	0	0	0	0	0	**30**	0	0	30	100
G. thy	0	0	1	0	0	2	1	**80**	1	85	94.12
G. tru	0	0	4	0	0	1	0	0	**75**	80	93.75
Sum	24	70	128	30	30	70	32	85	88	557	
Recall (%)	100	100	96.09	100	100	94.29	93.75	94.12	85.23		

Table 3.20: LDA classifier with the 7 morphometric features.

	G. art	*G. cic*	*G. der*	*G. gas*	*G. khe*	*G. sal*	*G. som*	*G. thy*	*G. tru*	**Sum**	**Precision (%)**
G. art	**24**	0	0	0	0	0	0	0	0	24	100
G. cic	0	**70**	0	0	0	0	0	0	0	70	100
G. der	0	0	**124**	1	0	1	0	1	10	137	90.51
G. gas	0	0	2	**28**	0	0	0	0	0	30	93.33
G. khe	0	0	1	0	**28**	0	0	0	1	30	93.33
G. sal	0	0	0	0	0	**64**	0	5	2	71	90.14
G. som	0	0	0	0	0	0	**29**	0	1	30	96.67
G. thy	0	0	1	0	0	9	0	**74**	1	85	87.06
G. tru	0	0	8	0	0	0	0	0	**72**	80	90
Sum	24	70	136	29	28	74	29	80	88	557	
Recall (%)	100	100	91.18	96.55	100	86.49	100	92.5	82.76		

Table 3.21: K-NN classifier with the 7 morphometric features.

	G. art	*G. cic*	*G. der*	*G. gas*	*G. khe*	*G. sal*	*G. som*	*G. thy*	*G. tru*	**Sum**	**Precision (%)**
G. art	**23**	0	0	1	0	0	0	0	0	24	95.83
G. cic	0	**70**	0	0	0	0	0	0	0	70	100
G. der	0	0	**128**	0	0	2	0	1	6	137	93.43
G. gas	0	0	1	**29**	0	0	0	0	0	30	96.67
G. khe	0	0	0	0	**30**	0	0	0	0	30	100
G. sal	0	0	1	0	0	**68**	1	1	0	71	95.77
G. som	0	0	0	0	0	0	**30**	0	0	30	100
G. thy	0	0	0	0	0	2	1	**72**	0	85	96
G. tru	0	0	7	0	0	0	0	0	**73**	80	91.25
Sum	23	70	137	29	31	72	32	84	79	557	
Recall (%)	100	100	93.43	96.67	100	94.44	93.75	97.29	92.40		

Table 3.22: The 7 features implemented using the MLP classifier.

	G. art	*G. cic*	*G. der*	*G. gas*	*G. khe*	*G. sal*	*G. som*	*G. thy*	*G. tru*	**Sum**	**Precision (%)**
G. art	**24**	0	0	0	0	0	0	0	0	24	100
G. cic	1	**69**	0	0	0	0	0	0	0	70	98.57
G. der	0	0	**134**	0	0	1	0	1	1	137	97.81
G. gas	0	0	1	**29**	0	0	0	0	0	30	96.67
G. khe	0	0	5	0	**25**	0	0	0	0	30	83.33
G. sal	0	0	0	0	0	**69**	0	2	0	71	97.18
G. som	0	0	8	0	0	3	**13**	0	6	30	43.33
G. thy	0	0	0	0	0	2	0	**83**	0	85	97.65
G. tru	0	0	79	0	0	0	0	0	**1**	80	1.25
Sum	25	69	227	29	25	75	13	86	8	557	
Recall (%)	96	100	59.03	100	100	92	100	96.51	12.51		

Table 3.23: The SVM classifier with the 7 feature set.

accurate and trusted identification. By combining multiple models in one system, it aims to provide a reliable model for species prediction.

3.5 ENSEMBLE CLASSIFICATION

The *Gyrodactylus* classification methods described previously make use of a single feature set in conjunction with a single classifier for pattern recognition. The main drawback of these approaches is that a single feature or classifier only captures the true identification for certain species. To maximise the accuracy in classification, it is necessary to apply different feature sets and different classifier sets. This is the motivation for the use of multiple features and classifiers to identify the *Gyrodactylus* species. In previous research [8], a variety of single classifiers and different set of features have been experimented with. It was found that none of these approaches produced significantly better results with regard to accurately classifying different types of *Gyrodactylus* species.

Ensemble based methods have recently enjoyed great attention [22] due to their reported superiority over single method based system generalisation performance [63], [21]. The aim of classification is to combine multiple models (classifiers or features) to solve particular problems [40]. Ensemble methods can be divided into a number of categories, such as ensemble classifiers [118]; ensemble features [69]; and ensemble feature and classifiers [45]. To demonstrate the full and practical importance of using a multiple classifier system, an analogy can be made with decision making in everyday life. When making an important decision, an expert is likely to ask opinions from several other experts before the final decision is made. In such a situation, the final decision is made by combining the individual decisions of several experts. The idea behind all ensemble based systems is that if individual classifiers or features are diverse, then they can make different errors, and combining these models can reduce the error through averaging.

Ensemble learning is primarily used to improve classification or prediction performance, where a single model does not have these capabilities, especially in dealing with multiclass problems. According to Yu and Xu [149], ensemble classification is considered due to the difficulty in acquiring full classification accuracy using traditional classification approaches,

Table 3.24: Summary of the correct identification of *Gyrodactylus* (e.g: *a* = *G. arcuatus*, *c* = *G. cichilidarum*, *d* = *G. derjavinoides*, *g* = *G. gasterostei*, *k* = *G. kherulensis*, *s* = *G. salaris*, *m* = *G. sommervillae*, *t* = *G. thymalli*, *r* = *G. truttae*) species by the different models.

Feature set / Classifier	*a*	*c*	*d*	*g*	*k*	*s*	*m*	*t*	*r*
25 features									
LDA	x	x		x	x		x		
K-NN	x	x			x				
MLP	x	x		x	x	x			
SVM	x	x		x	x		x		
21 features									
LDA	x	x		x	x		x		
K-NN	x	x			x	x			
MLP	x	x		x		x			
SVM	x	x		x			x		
20 features									
LDA	x	x		x	x		x		
K-NN	x	x			x				
MLP	x	x		x	x	x			
SVM	x	x		x	x		x		
7 features									
LDA	x	x		x	x		x		
K-NN	x	x							
MLP		x			x		x		
SVM	x								

due to large datasets (such as those containing a large number of features or data points). In their paper, SREC (Simple Rule-based ensemble classifiers) was proposed. In SREC, the final classification is identified with weighting voting. Weighting voting is one example of an ensemble method. In weighting voting, a certain weight is given to a specific classifier, where this classifier is a good classifier compared to other classifier models also used in the same model. Results indicate that the proposed method is effective and feasible, and produces less classification errors than many other classifiers [149].

There is a need for methods that can learn interpretable multi-target models to predict several target classes simultaneously. A study by Aho *et al*. [7], introduced the FIRE (Fitted Rule Ensembles) method that can learn multi-target regression rule ensembles. Results show the general trend of larger models having better accuracy. In the scenario of inter-disciplinary research, one example is field line proteomic mass spectra classification [54], which proposed a systematic approach based on decision tree ensemble methods. This is used to automatically determine proteomic biomarker and predictive models. The framework [54] relies on a toolbox of generic supervised machine learning algorithms, consisting of decision tree induction and several decision tree based ensemble methods. This proposed technique improves processing time and provides promising results for predictive models and for the identification of biomarkers.

The success of an ensemble system depends on its ability to correct the errors of some of its members; classifiers and feature sets. This is dependent on the diversity of the classifiers that make up the ensemble. If all classifiers provide the same output, correcting a possible mistake is not possible. Therefore, individual classifiers in an ensemble system need to make different errors on different instances [71]. If each classifier makes different errors, then a strategic combination of these classifiers can reduce the total error. Specially, an ensemble system needs classifiers whose decision boundaries are adequately different from those of others. To classify various species of *Gyrodactylus*, a single feature format or classifier is not sufficient for correctly classifying the true species. This was shown by the results discussed in Section 3.4.6, where, for example, the LDA classifier with 25 features performed well with regard to the identification of *G. sommervillae*, but poorly with regard to *G. salaris*.

More than one method is then combined using two different ensemble combination techniques. To classify various species of *Gyrodactylus* (nine species to be specific), it is not an easy task, as mentioned in earlier sections. To produce accurate and efficient performance of species identification, more than one technique is required. Multiple classifiers which consider different sets of features have been implemented to achieve the objective of this study. The motivation of this work is due to the high misclassification rate identified in previous research [8].

3.5.1 *Majority voting*

As the objective of this research project is to classify and identify accurately multiple species of *Gyrodactylus* using machine learning classifiers and feature selection techniques, in addition to attempting to minimise misclassification errors through individual classification and reducing the number of features using feature selection techniques, an ensemble based majority voting approach has been proposed here. Ensemble learning is primarily used to improve classification prediction performance, where a single model does not have these capabilities, especially in dealing with multiclass problems [40], [149].

The main idea with ensembles of several classifiers is that several classifiers are created and then combined into one model [41]. The ensemble method is usually categorised into two categories, fixed rules and trained methods. Fixed rules combine the individual outputs in a fixed manner, such as the same rule and majority voting [119]; whereas trained methods, including the weighted combination and meta-classifier [77], combine outputs via training on validation dataset.

In emerging ensemble models, majority voting has been applied. According to Kainulainen [71], if the output consists of class labels, then majority voting is recommended. For the purpose of this study, it was decided to apply majority voting to classification and identification of multiple type of *Gyrodactylus* species. Simple majority voting is a decision rule that selects one of many alternatives, based on the predicted classes with the most votes [77]. For example, a hypothetical ensemble could consist of three classifiers, $h1$, $h2$ and $h3$. If $h1(x)$ is wrong, $h2(x)$ and $h3(x)$ may be correct, and the majority vote will correctly classify the sample x.

Majority vote counts the votes each class over the input classifiers and selects the majority class. Theoretically, if the classifier makes independent errors, the majority vote outperforms the best classifier.

The classification of an unlabelled instance by the ensemble is obtained by combining the predictions of the individual classifiers [102]. In majority voting, each classifier in the ensemble predicts the class label of the instance consider. Once all the classifiers have been queried, the class that received the greatest number of votes is returned as the final decision of the ensemble. The time needed to classify an instance increases linearly with the size of the ensemble.

In work by Bouziane [21], majority voting has been applied to combine K-Nearest Neighbor, Artificial Neural Networks and Multi-class Support Vector Machines for predicting the Secondary Structure of globular proteins. They implement three voting strategies; Simple Majority Voting (SMV), Influence Majority Voting (IMV) and Weighted Majority Voting (WMV), and these techniques are compared. IMV gives better results than SMV but the results given by WMV are best. In this study, two widely used datasets have been used, these are: RS126 and CB513.

Ensemble voting system for multiclass protein fold recognition [23] is another example of work that applies majority voting for combining more than one model. In this research, three type of homogenous ensemble classifiers are first evaluated (feature selection methods and classifiers), and then a heterogeneous ensemble voting system was introduced for multiclass protein fold classification. The results show an improvement in prediction accuracy with this proposed method of ensemble classifier with different features.

3.5.2 *Ensemble classification for Gyrodactylus species identification*

In this work, more than one classifier and feature set have been considered for constructing an ensemble classification for *Gyrodactylus* species identification. In this research, Linear Discriminant Analysis (LDA) and K-Nearest Neighbor (K-NN) form the classifier base; while three different feature sets have been considered for classifying multiple species. In this work it was decided to make use of LDA and K-NN because when comparing these approaches to

the MLP and SVM techniques, the misclassification errors are similar, meaning that there was little additional value from their integration. In addition, as these are non-linear classifiers, it was felt that the implementation of the chosen classifiers was less complex than the additional integration of non-linear approaches. The success of an ensemble system depends on its ability to correct the errors of some of its members [71]. If all classifiers provide the same output, correcting a possible mistake is not possible. Therefore, individual classifiers in an ensemble system need to make different errors on different instances.

The chosen features are the 25 full feature set, 21 selected features using the SFS method and 20 selected features from SBS. For the decision making in combining the classifiers and three different feature sets, majority voting is applied. Majority voting is often used to combine the decisions of the classifiers that make up an ensemble [77].

This section discusses the construction of an ensemble model by combining several single LDA and K-NN models to learn the same data with different subsets of morphometric *Gyrodactylus* features. In Fig. 3.4, the main structure of an ensemble classifier is depicted. The model contains N single classifiers; each single model has D inputs. Thus, the whole model can input $D * N$ features. The output strategy for the model is majority voting. The classifier of an ensemble model consists of different individual classifiers with different feature sets; 25 features, 21 features and 20 features.

The proposed ensemble voting system is composed of a feature selection system, a number of individual classifiers and a voting system. The framework is presented in Fig. 3.4 and contains:

(1) Feature selection. 25 features were extracted from SEM images using manual point-to-point measurement techniques from three parts of *Gyrodactylus* hook features. From these 25 features, a number of feature selection techniques have been applied to acquire the optimum features for correctly classifying the species.

(2) Classification. Two classifiers are defined for classifying the *Gyrodactylus* species. These are Linear Discriminant Analysis (LDA) and K-Nearest Neighbor (K-NN).

(3) Voting system. Different results will be obtained from the different classifiers by using different features sets. These results are input into the voting system. These classifiers do not

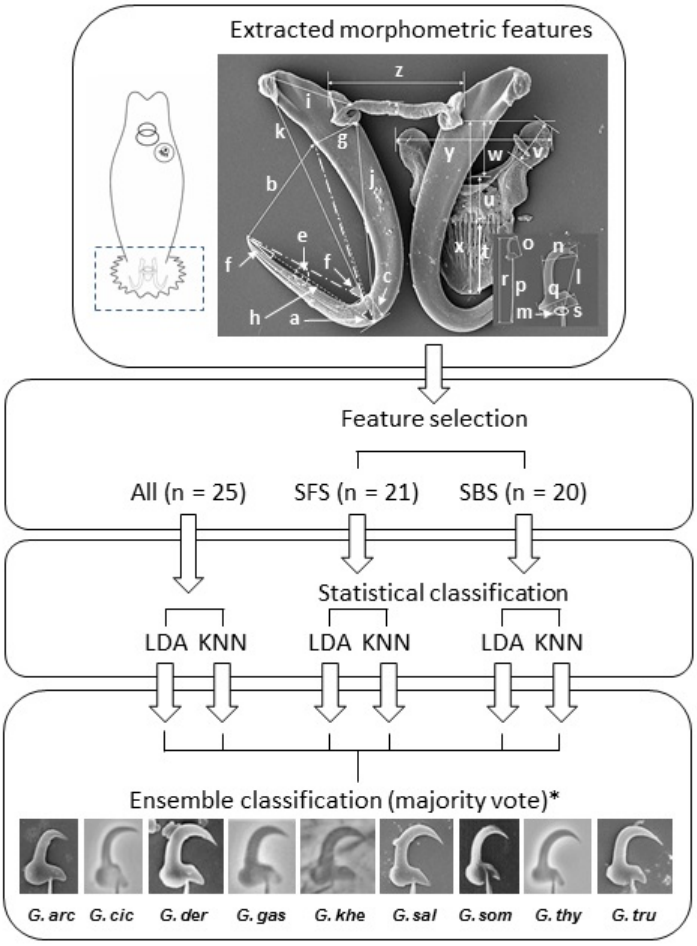


Figure 3.4: Framework of an ensemble based majority voting classifier and feature selection model in classifying multiple species of *Gyrodactylus*.

have any specific weight to adjust the contributions of the classifier to the voting system but are weighted equally.

Following the statistical step, to maximise the accuracy of classifying each specimen, an ensemble based method of majority voting was applied to combine the results from the multiple models (classifiers of features). The voting based method, as defined by Zhang and Yunqian [151], operates on the output from the statistical step and is defined by:

$$\arg\max_{y} \sum_{t=1}^{T} I(\hat{H}_t(D) = y), y \epsilon Y \tag{3.5}$$

where Y is the number of *Gyrodactylus* species outcomes defined by Y is $1, 2, ..., k$. The ensemble T classifiers are given by $(\hat{H}(D)_i)_i^T = 1$. The class assigned to an unlabelled sample is given by D. $\hat{H}(D)$ is the prediction of the t-th ensemble member, I is an indicator function and $Y = y_1, ..., y_k$, where T_i is the number of members in the ensemble that predict the class y_i for the sample to be classified.

The entire methodological process pipeline for the identification of species of *Gyrodactylus* is summarised in 3.4 and is given by the following algorithm:

1. Given $D = X_1 Y_1, ..., X_n Y_n$ where $Y_j \in (1, 2, ..., 9)$.

2. Perform the feature selection step. In addition, the original 25 features are retained for use later on in the classification of *Gyrodactylus* specimens.

    a) Sequential forward selection (select the 21 features of the 25 original features). Starting from an empty set, sequentially add the feature A^+ that maximises $E(B_j + A^+)$ when combined with the features B_j that have already been selected.

        i. Start with an empty set $B_0 = (\emptyset)$.

        ii. Select the next best feature $A^+ = \arg\max_{A \notin B_j} E(B_j + A)$.

        iii. Update $B_j + A^+; j = j + 1$.

        iv. Go to ii (repeat until the optimum set of features is acquired).

    b) Sequential backward selection (select the 20 features of the 25 original features). Starting from the full set, sequentially remove the feature A^- that least reduces the value of the objective function $F(B - A^-)$.

i. Start with the full set $B_0 = A$.

ii. Remove the worst feature $A^- = \arg\max_{A \epsilon B_j} F(B_j - A)$.

iii. Update $B_{j+1} = B_j - A^-; j = j + 1$.

iv. Go to ii (repeat until the optimum set of feature is obtained).

3. Then, performs ensemble classification (LDA and K-NN classifiers) to the different feature sets (*i.e.* all 25 features, 21 features as determined by SFS and 20 features as determined by SBS). Four classifiers are applied to each set of features, where:

   a) Linear Discriminant Analysis (LDA)

   $\hat{N}_i = W_0 + W_{i1}D_1 + W_{i2}D_2 + ... + W_{in}D_n$

   b) K-Nearest Neighbor (K-NN)

   $(X_t, X_u) = \sqrt{\sum_{mm=1}^{m} (X_t^{mm} - X_u^{mm})^2}$

4. The results from the statistical classification step are then combined within an ensemble system which then applies a majority voting model expressed by:

   $Q = \arg\max_y \sum_{t=1}^{T} I(\hat{H}_t(D) = y), y \epsilon Y$

5. In the event of an equal split in the voting calculation, then the result from SFS-LDA is added to the system (the technique outperforming all others in an earlier study by [8]).

6. The identity of the specimen Q is determined i.e. as being *G. arcuatus*, *G. cichilidarum*, *G. derjavinoides*, *G. gasterostei*, *G. kherulensis*, *G. salaris*, *G. sommervillae*, *G. thymalli*, or, *G. truttae*.

3.5.3 *Results and discussion*

Combining the outputs of several predictors can improve performance over a single generic one [63], [21]. Good ensemble members must be both accurate and diverse [71], which poses the problem of generating a set of predictors with reasonably good individual performances and independently distributed predictions for the test points .

To demonstrate the proposed algorithm using ensemble based majority voting, the experiment used 557 specimens of the morphometric point to point measurement dataset of nine classes in all the following experiments. Three sets of point to point measurement features are selected, where three of them are selected using the feature selection techniques (*e.g.* SFS and SBS).

As combining the outputs of several predictors improves the performance of a single model, formal support to enable this is provided by majority voting. Majority vote is a decision rule approach that selects one of many alternatives, based on the predicted classes with the most votes [77], as discussed previously. To prevent voting conflicts, the best single model accuracy will be added to the ensemble system for determining the true species.

Two classifiers are used in the ensemble voting system. For each feature set, two classifiers are used and then the result put into the voting system. The overall classification accuracy of the ensemble model is 97.29% $\pm$1.98, demonstrating that the ensemble system has improved overall classification performance, compared to single classification approaches. Although the accuracy of classification is not improved significantly, the numbers of species misclassifications is improved in comparison to the best single model. This result is presented through the confusion matrix of the ensemble system (Table 3.25). Here, seven species in total are shown to be misclassified. Overall, in the ensemble system, 15 specimens have been misclassified.

For comparison, amongst the experimental classification using a single model and feature set, the best performing models are the MLP classifier (Table 3.10) and the LDA classifier (Table 3.12). Referring to the confusion matrices, the MLP model has eight misclassified species and total number of individual misclassification instances is 13. Similarly, using the LDA model, eight species in total show examples of misclassification, and 18 individual specimens from 557 have been misallocated to the wrong species. Although the MLP model has a slightly smaller number of misclassified species, and appears to be very suitable for use as part of the ensemble model, it was ultimately decided that it was more suitable to make use of different approaches.

The MLP approach was considered as a potential component of the ensemble method, however, the results of the MLP method were not found to be significantly better than the LDA and the K-NN classifiers, and do not provide robust variation with regard to performance of

classification on different species. By this it is meant that as discussed previously, the ensemble method should be made up of classifiers that produce different errors, in order to be as robust as possible, and there is more difference in the errors between the LDA and K-NN classifiers than between those and the MLP, making the former two classifiers more suitable for use. In addition, the MLP is more complex to create, optimise and train than the simpler classifiers, and so in summary, it was felt that despite the individual results, the overall ensemble method not found to lead to an enhancement in the ensemble classification performance.

When ensemble based majority voting is applied, we can see that 68 from 71 specimens of *G. salaris* would be classified correctly as being *G. salaris* and 83 from 85 of *G. thymalli* would be classified correctly as being *G. thymalli*. In ensemble based majority voting, the *G. salaris* has been misclassified as *G. tnymalli* for two specimens out of the total of 71. While *G. thymalli* was misclassified as *G. derjavinoides* and *G. truttae* with every one having one misclassification score out of 85 specimens. Misclassifications, particularly in the case of *G. salaris*, would have serious consequences should they allow this species to slip through undetected [18].

Among these nine species of *Gyrodactylus*, *G. derjavinoides* and *G. truttae* remain misclassified. Overall, the comparison between an ensemble approach with individual classifiers and feature sets shows that an ensemble approach based on majority voting performs consistently better than any individual classifier and feature set and across all other classifier and other feature sets. It is thus concluded that an ensemble of classifier and feature sets together is more effective than only combining features sets.

Although the LDA and MLP classifiers with the set of 21 features have demonstrated good performance in classifying species, the objective of the study is still not achieved, which is to reduce the misclassification error in classifying the *Gyrodactylus* species. For those reasons, an ensemble LDA and K-NN with three different set of features was proposed and experimented with. The evidence clearly supports the conclusion that combining varieties of classifiers (LDA and K-NN) and different feature sets (25 features, 21 features and 20 features) is of benefit to solving this problem. Although the misclassification errors are not improved significantly, the minimising of the number of errors has been achieved.

	G. art	*G. cic*	*G. der*	*G. gas*	*G. khe*	*G. sal*	*G. som*	*G. thy*	*G. tru*	**Sum**	**Precision (%)**
G. art	**24**	0	0	0	0	0	0	0	0	24	100
G. cic	0	**70**	0	0	0	0	0	0	0	70	100
G. der	0	0	**130**	0	0	**2**	0	0	5	137	94.89
G. gas	0	0	0	**30**	0	0	0	0	0	30	100
G. khe	0	0	0	0	**30**	0	0	0	0	30	100
G. sal	0	0	0	0	0	**68**	0	**2**	1	71	95.77
G. som	0	0	0	0	0	0	**30**	0	0	30	100
G. thy	0	0	1	0	0	0	0	**83**	1	85	97.65
G. tru	0	0	3	0	0	0	0	0	**77**	80	96.25
Sum	24	70	134	30	30	70	30	85	84	557	
Recall (%)	100	100	97.01	100	100	97.14	100	97.65	91.67		

Table 3.25: Confusion matrix of ensemble model.

Morphometric data taken from the sclerotized structures of *G. salaris* and related species that co-occur on salmonids were subjected to analysis. Statistical methods provided the function for a rapid automated diagnostic system that was robust enough to allow for the perfect discrimination of *G. salaris*. These results suggest that it is now feasible to develop an automated system for the identification of *G. salaris* and its discrimination from other closely related gyrodactylids that occur on the same hosts [76]. We believe that this methodology will find successful applications within other biological systems, not only for the identification of the monogenean *G. salaris* discussed here in detail but the discrimination of other pathogens such as *Myxosporea cerebralis*. Organisms bearing hard parts in the form of sclerotized hooks, copulatory structures or skeletal body components, all of which represent a challenge to taxonomic discrimination, equally lend themselves to identification by statistical classification. Indeed, any organism, free-living or parasitic, possessing a structure of constant size and shape in the key stages of its life cycle could be subjected to classification [76].

In this chapter, four machine learning classifiers and three feature selection methods were used to assess their performance in correctly classifying nine species of *Gyrodactylus* using morphometric data extracted from their attachment hooks. The correct identification of one species, *G. salaris*, a notifiable pathogen of salmonids throughout Europe, however is paramount. It is essential, therefore, to employ a method that does not generate misclassifications where *G. salaris* is concerned. From the results presented here, it can be seen that a single classifier is not sufficient for the accurate classification of all *Gyrodactylus* species to their true class and an ensemble approach has been proposed.

This chapter also presented the application of an ensemble method based on majority voting. In the application of an ensemble based majority voting strategy, the 557 image morphometric dataset was used. In this dataset, the nine different species and 25 original features were provided. An ensemble model has been constructed for classification. The LDA and K-NN classifiers have been applied with three different feature sets for morphometric *Gyrodactylus* species identification. The results indicate that a single classifier with different feature sets cannot achieve higher performance than an ensemble approach, which correctly

identifies multiple type of *Gyrodactylus* species. In order to target the problem, a new ensemble model, which combines the classifiers and feature sets for classification and identification, and majority voting, is applied to assign the class label. Experimental results demonstrate that using the ensemble technique is an effective way to combine different classifiers and feature sets for better classification performance. In this research, it is shown that it is possible to take advantage of an ensemble framework for combining different classifiers and feature sets to boost overall performance.

This work continues in the subsequent chapter, exploring a very pertinent and realistic research problem of classifying specimens based on Scanning Electron Microscope (SEM) images, which necessitates image pre-processing. Instead of performing classification using morphometric data manually extracted from slide mounted specimens, it is hoped that the data extraction process can be automated, accelerating the process of species identification.

4

GYRODACTYLUS SEM IMAGES IDENTITIFCATION

4.1 INTRODUCTION

In the previous chapter, a morphometric dataset has been used in classification and identification using machine learning classifiers and feature selection techniques. The data preparation, including the feature set, is prepared by the domain expert using a point to point measurement approach to the three parts of the sickle hooks of the *Gyrodactylus* species. Later, these experimental results will be compared with an image processing based feature extraction method, applied to Scanning Electron Microscope (SEM) images of *Gyrodactylus*.

Image analysis involves the use of image processing methods that are often designed in an attempt to provide a machine interpretation of an image, in a form that allows some decision criterion to be applied [20]. Pattern recognition uses a range of different approaches that are not necessarily based on any one particular theme or unified theoretical approach. In this chapter, an image processing technique has been introduced to apply to SEM images to predict the various classes of *Gyrodactylus* species groups.

The overall aim of this thesis is to find a potential method or model to be used to discriminate between multiple species of *Gyrodactylus* using Scanning Electron Microscope (SEM) images. The aim is to solve the same problem as discussed in the previous chapter, that of finding the best solution for identification of these species. In this chapter, SEM images types have been used, and of these images it was decided to use only the section of the image containing the marginal hook of the sickle attachment.

The reminder of this chapter is divided into a number of sections. In section 4.2, a review of a number of potential feature extraction techniques is presented, with the motivation and the framework summarised. After discussion of the various possible models for image feature extraction, the potential of Active Shape Model (ASM) and classification for separating species

is then described in section 4.3. Section 4.4 discusses the topic of materials and methods for extracting the SEM images of three species of *Gyrodactylus*. Results and discussion are provided in section 4.5, While in the section 4.6, the new potential feature extraction tool is investigated. The Complex Network model is explored to find the potential in extracting the SEM of *Gyrodactylus* species. Finally, the last section will concludes this chapter.

4.2 IMAGE PROCESSING

Much research has been devoted to the recognition of digital images, especially microscope images, but so far it is still an unresolved problem [110], [83], due to distortion, noise, segmentation errors, overlap and occlusion in colour images. Recognition and classification techniques have gained a lot of attention in recent years due to many scientists utilising these techniques in order to enhance their own problem domains.

Computerised parasite recognition and classification is still a new area with regard to the aquaculture domain, and is considered to have much potential application for encouraging and pushing aquaculture research ahead. Improved software and hardware technical advancements offer the chance and opportunity to apply recognition and classification technology to this domain, where it can help to improve the efficiency and accuracy of parasite identification. In the classification of parasite fish species it is important to achieve good accuracy, so that the correct particular treatment may be provided to prevent destruction to human health.

To provide a potential solution to the problem described above, image analysis is explored. Image analysis is a field of science which allows scientists to explore a complex assortment of images and effectively predict structure from the images autonomously. According to Kasturi [75], image analysis refers to algorithms and techniques that are applied to images to obtain a computer readable description from a pixel data. Instead of image analysis, image processing techniques have also been developed. In contrast with image analysis, image processing involves the use of electronic tools which allow the user to define changes within the parameters of the electronic signal [154]. This approach is needed to increase the pictorial information for human interpretation. One example of image processing is removing the illumination from images.

In terms of image analysis, feature extraction has been explored. Recognising the species group from the hook features makes the species recognition process more accurate and effective. Feature extraction is the key to both object segmentation and recognition, as it is to any pattern classification task. Examples of the features that might be of interest to extract include length, width, shape and angle. In the manual measurement of features, these tasks heavily depend on the concentration of the person taking the measurement; otherwise, the result of morphometric analysis will be false. And of course, the temporal duration of the manual measurement process is substantial. With state-of-the-art computer processing techniques, it is possible for this process to be performed efficiently and effectively, and thus, provide the prediction in a shorter time, and more accurately.

The final task for an image processing system is to take an object region in an image and classify or identify it. In other words, it needs to generate a collection of classes or objects, such as '*G. arcuatus*', '*G. salaris*', and then be able to take a region in an image and determine which, if any, of the classes that region falls into.

Before the class is determined, the information required for the purpose of classification, *i.e.* the extracted mathematical measurements (features) from that objects need to be provided. There are two categories of features: shape features (geometry measurements captured by both boundary and the interior region) and texture features (intensity of images). In this study, only shape feature information is considered due to providing more informative features than texture features. For identification of *Gyrodactylus* species, the shape information provides more unique features than texture. A number of related studies [76], [104], [126], [129], [130] also use only the shape features, but different information (different part of the hooks).

4.2.1 *Review of image processing techniques*

In recent decades, digital image processing, image analysis and machine vision have been significantly developed, and they have become a very important part of artificial intelligence and the interface between human and machine grounded theory, and applied technology. These technologies have been applied widely in industry, but rarely in the realm of aquaculture.

The identification of the edges or shapes of an object in an image scene is an important aspect of the human visual system because it provides information on the basic topology of the object from which an interpretative match can be achieved. Shape detection or segmentation is a pre-requisite for object identification in order to then perform further processing, feature extraction and classification.

Feature extraction is essential in many vision and biometric applications. The performance of feature-based face recognition algorithms relies heavily on the quality of the feature extraction. Selection of a feature extraction method is probably the single most important factor in achieving high recognition performance [142], [82]. In this study, accuracy in feature extraction is a must, since the majority of the *Gyrodactylus* species have a similar shape to each other, especially *G. salaris* and *G. thymalli* [130].

In human communication, shape description (features) have been used. It is one of the most important visual attributes of an object and the first used to perform object classification and identification [90], [141]. Specifically, in classification and identification of multiple species of *Gyrodactylus*, shape information has been used, although different methods of identification have been applied [104], [126], [129], [127].

Active contour modelling or snakes [74] is a feature shape extraction method that is mostly applied to medical image processing problems. The snake model has the ability to freely deform to fit the images instead of having a rigid shape application. Unfortunately, it has limitations that mean it is very challenging to apply to the case of SEM *Gyrodactylus* images. The physical appearance of the images that have been surrounded with the tissue make the segmentation procedure become complicated. It will therefore result in inaccuracy with regard to contour measurement.

Fish recognition based on the combination between robust features selection, image segmentation and geometric parameter techniques [11] has been demonstrated, using the scanned images for recognition of fish species. Measurements have also been done by measuring the size, shape, colour and geometrical parameters. In this study, the colour information plays the main part in fish species identification. Neural networks and decision trees have been used as their classifier tools to perform the subsequent classification after the initial image processing. Another paper that also considered colour information is by Du [44]. Here, the

colour information has been used to recognise 20 different types of plant species, using extracted features such as geometrical features and invariable moment features. Using the Move Median Centres (MMC) as classifier, their method was said to be more robust and an improvement on other approaches.

In addition, 2D gabor filter image processing has been successfully applied in the identification of mammals through the use of their hair [110]. The selection of this technique is due to the ability to perform rapid matching, carried out using either Hamming or Euclidean distance measures. This type of problem considers the whole object for performing identification.

Protozoan parasite extraction using basic techniques has achieved good performance in discriminating between species of protozoan [83]. When working with microscope images, there are similar problems that need to be overcome, such as illumination, noise, and the size of the target object in the image. In addition, there is also the complexity of the image content to be considered. Similar to the difficulties presented with *Gyrodactylus*, the different species of ectoparasites have different forms according to their maturity, or in the context of *Gyrodactylus*, their location and season. In classification of protozoan species using microscope images [83], the following methods have been applied sequentially: (a) color space transforming; (b) gamma-equalization; (c) two-mean filter; (d) two-classes edge enhancement; (e) two-means clustering filter; (f) morphological opening operation; and (g) largest independent component detection. Many steps have involved in extracting the valuable features for input to classification, and this procedure needs to be repeated each time for every image. Such a complex procedure is not efficient for detection and extraction of the object.

A similar problem to the *Gyrodactylus* species identification problem can be found in insect species recognition [92]. Recognition of species is not the same as recognition of objects. This is because in identification of the species, biologist expert knowledge is required for species recognition. In addition, object recognition is rather simple compared to human face recognition. In recognition of insect species, class specific sparse representation has been proposed [92]. In this study, the sparse representation is an expression of the input signal as a linear combination of base elements in which many of the coefficients are zero. In addition, SVM has been used as a classifier tool for classifying the species. Although this technique is dealing with a similar problem to *Gyrodactylus* species identification, it has the limitation

of not having the ability to perform scaling and rotation variance to the input images. SEM images of *Gyrodactylus* may come in varieties of rotation, with a particularly significant issue being the range of different scales for the same species.

Previously, it has been considered impossible to implement automated image processing for identification of parasite species due to several reasons. Firstly, the physical appearance of the scan images; the focused object has been overlapped with the tissue. Secondly, the recognition and identification of the fish parasite shape is made difficult because there are no clear boundaries defining the actual shape of the parasite species. As a result, *Gyrodactylus* species identification using image processing techniques is an area in need of significant development.

Many segmentation or feature extraction methods have been proposed and proved to be successful in implementation of a number of research problems. It is possible to apply many of these methods to the extraction of SEM *Gyrodactylus* images. Unfortunately, there are many criteria that might not exist in their problems, but do with regard to this specific challenge that need to be taken into consideration when applying image extraction techniques. The main contrast in SEM images of *Gyrodactylus* compared to images of other species is the brightness of the image quality and the interconnectedness of the focused object with surrounding tissue. Fig. 4.1 shows an example of images that have tissue surrounding of the object of focus.

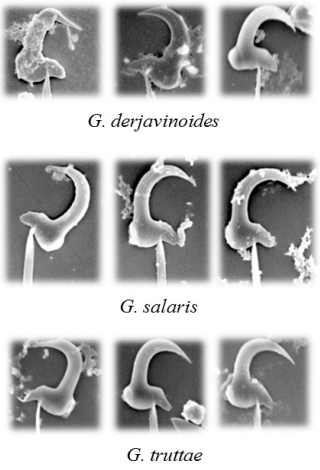


Figure 4.1: Example of images where the object of focus has been surrounded by tissue.

One of the objectives of this research is to identify and utilise an image processing technology that has ability to extract noisy images with similar pattern representation. For these reasons, the Active Shape Model (ASM) technique has been explored to evaluate the suitability of

using it for extracting informative features of multiple species of *Gyrodactylus*. In the case of SEM images of the fish parasite, only the shape features are considered, since, it was found that the texture information does not increase the accuracy of species prediction. Shape or contour refers to the boundary of the object, and that represents the shape of the object.

Devijver and Kittler [142] have highlighted that different feature extraction methods have been designed to take account of specific recognition problems and available data. Certain feature extraction methods may be found to perform successfully in one application domain, but might be not useful in another domain. In this research project, the Active Shape Model (ASM) technique has been identified as the most suitable to be adopted for recognising and identifying multiple *Gyrodactylus* species. Further discussion about this potential application is presented in the following section.

The shape model in ASM is given by the principal component of a vector of landmark points [143]. The principal component method has been used for extracting the most valuable features. Using the grey appearance model methodology proposed in ASM, it will ensure that the segmentation places the object at a location where the image structure around the border or within the object is similar to what is expected from the training images. While the fitting procedure is an alteration of landmark displacements.

4.2.2 *Existing methods of parasite identification and classification*

As discussed previously, identification and classification techniques have gained a lot of attention in recent years. Fish recognition and classification is an active area in the agriculture domain [11], while fish parasite identification using artificial intelligence techniques is considered to be a new area to explore and has much potential research application in this domain.

A number of statistical classification based approaches have been applied to morphological data [104], [130], [129], [126], and molecular-based techniques targeting specific genomic regions [32], [60], [105], have been developed to discriminate the pathogenic species, *G. salaris*, from other non-pathogenic species of *Gyrodactylus* that co-occur on salmonid hosts. While each technique is able to detect *G. salaris* within a population of specimens and to discriminate

it from its congeners with high levels of correct classification, the techniques can be time consuming [130]. If a system consisting of an image recognition model can be constructed and proposed to extract key discriminatory features from the attachment hooks of each species, then it is anticipated that the identification process could be accelerated with equivalent or better rates of correct identification.

4.3 THE POTENTIAL OF ACTIVE SHAPE MODELS (ASM) FOR SPECIES CLASSIFICATION

The ASM method has been successfully utilised for understanding of factors underlying morphologic and pitch-related functional variations affecting vocal structures and the airway in health and disease [106]. In addition, the ASM method was found to be the best method that can account for the varieties in variation [136].

Another successful application of ASM is for face recognition, as shown by [85]. In this study, ASM was applied to the alignment of the face, with four major improvements. These are: (1) a model combining a Sobel filter [122] and the 2D profile in searching for a face in an image; (2) application of the Canny [73] method for edge enhancement; (3) use of a SVM to classify the landmark points; and (4) automatic adjustment of the 2D profile according to the size of the input image. The introduction of these improvement has improved the process of finding landmarks and thus will save time during the training and testing of images.

ASM was also implemented for extracting features for plant recognition based on the leaf shape [136]. In this study, ASM was applied for recognising weed species, and due to using the ASM, it was found to be possible to not only take leaf shapes into account, but also the overall geometry of the seedlings.

With statistical shape models, shape can be characterised in terms of independent modes of variation. Variation in the image presentation is a key point that needs to focused on in this work. This is because a single species may come with much variation present, yet still be part of the same species. For example, location and water temperature can contribute to differences within the same species, although despite these variations, the overall shape of the hook remains the same.

The ASM technique permits users to construct a general shape model which is subsequently applied to all images in order to landmark the image area for every given image, providing a pattern that encapsulates the variation seen across the range of shape images. The subsequent ability (classification rate) of the developed model to separate "image classes" is in part based on the number of images used in the training set - in theory, the greater the number of images that are used in training and constructing the models, the better the classification ability of the resultant model. Given the success of ASM in resolving image-based shape recognition problems within the biomedical sphere, the research presented in this chapter aims to determine its utility when applied to SEM images of *Gyrodactylus* hooks.

The application of the ASM method to the analysis of *Gyrodactylus* attachment hooks is presented in Fig. 4.3. The input for the classification system is the specimen images, where a pre-processing step is applied to the required images. Once hooks have been processed to a common orientation, the ASM approach is then applied to extract informative features. These features are then reduced by a subsequent PCA step to select key features to be used as input features for each machine learning classification technique. Four machine learning classifiers have been used to evaluate the ASM performance. The breakdown of this framework will be presented at section 4.4.

The selection and consideration of the ASM method for feature extraction of SEM images is based on the following benefits. These are: (1) a shape model that ensures that the segmentation can only produce plausible shapes; (2) a gray-level appearance model is applied in ASM to ensure that the segmentation places the object at a location where the image structure around the border or within the object is similar to what is expected from the training images; (3) an algorithm for fitting the model by minimizing some cost function. With these advantages it can help the domain expert by aiding the processing and analysis of species immediately and accurately.

According to Ginneken *et al*. [143], [37], the main advantages of ASM compared to other model is its speed. It can save time during the image training, especially for those images that are considered to be noisy images. In the case of genus *Gyrodactylus* (Monogenea) species, most of the attachment hook (marginal hook) has been surrounded with tissue. Proper and

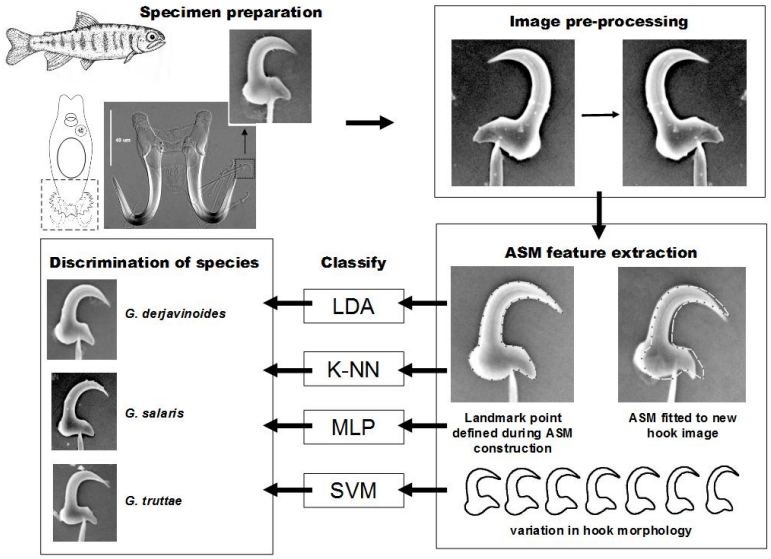


Figure 4.2: The methodological approach used in the current study. Specimens of *Gyrodactylus* were picked from the skin and fins of salmonids and their attachment hooks released by proteolytic digestion. Images of the smallest hook structures, the marginal hook sickles which are the key to separating species and typically measure less than 0.007 mm in length, were captured using a scanning electron microscope. The images were pre-processed before being subjected to an Active Shape Model feature extraction step to define 45 or 110 landmark points and to fit the model to the training set of hook images. This information is then used to train four classifiers (K-NN, LDA, MLP, SVM) and separate the three species of *Gyrodactylus* which includes the notifiable pathogen, *G. salaris*. Abbreviations: K-NN, K Nearest Neighbors; LDA, Linear Discriminant Analysis; MLP, Multi-Layer Perceptron; SVM, Support Vector Machine.

precise management with regard to image segmentation is needed since these hooks look very similar to each other.

4.4 MATERIALS AND METHODS

The focus of this research was on the genus of the *Gyrodactylus* species group. Specimens of *Gyrodactylus* (*G. derjavinoides* n = 25; *G. salaris* n = 34; *G. truttae* n = 9) were removed from their respective salmonid hosts and fixed in 80% ethanol. Subsequently specimens were prepared for scanning electron microscopy (SEM) by transferring individually, rinsed with distilled water, and the specimen were then mounted onto 13 mm diameter round glass coverslips,

where they had their posterior attachment organ excised using a scalpel, and the attachment hooks released using a proteinase-K based digestion fluid (*i.e.* 100 µg/ml proteinase K, 75 mM Tris-HCl, pH 8, 10 mM EDTA, 5% SDS). Once the hooks were freed from enclosing tissue, the preparations were flushed with distilled water, air-dried, sputter-coated with gold and then examined and photographed using a JEOL JSM5200 scanning electron microscope operating at an accelerating voltage of 10 kV. This data was collected and prepared by the parasitiology team, Institute of Aquaculture, Stirling, Scotland.

The study by Shinn *et al.* [127] has proven that in discrimination of the *Gyrodactylus* species, only marginal hook or hamuli information of features are useful enough sections of the specimen to be used for species classification. In this study, marginal hook of attachment hook will use in extraction and thus classify the species.

4.4.1 *Pre-processing*

In the pre-processing step, an image may need adjustment of its rotation to standardise the form of fish parasite object. As a number of the initially supplied images are provided in different rotation, to minimise the complexity of ASM implementation, all the images must have the same pose. For this purpose, image flipping has been applied to pre-process the images.

4.4.2 *ASM construction*

ASM were originally developed for the recognition of landmarks on medical x-rays. Landmark points can be acquired by applying a sample template to a "problem area", which appears to represent a better strategy over edge-based detection approaches [95], as any noise or unwanted objects within the image can be ignored in the selection of the shape contour. The shape variations in a training set are described using a Point of Distribution Model (PDM). The shape model is used to generate new shapes, similar to those found in the training set, which are fitted to the data using a model of the local gray value structure [37], [106], [138].

In the study presented in this chapter, the shape of each attachment hook image is presented by a vector of the position of each landmark, $G = (g_1, h_1, ..., g_r, h_r)$, where $(d_s h_s)$ denotes the 2D image coordinate of the s^{th} landmark point. The shape vector of the hook is then normalised into a common coordinate system. Procrustes analysis [64] is then applied to align the training set of images. This aligns each shape so that the sum of distances of each shape to the mean $\hat{F} = \sum |G_s - \bar{G}|^2$ is minimised. For this purpose, one hook image is selected as an example of the initial estimate of the mean shape and scaled so that $|\bar{G}| = 1$, which minimises the $\hat{F}$.

Assuming $\hat{G}$ sets of landmark points G_s which are aligned into a common shape pattern for each species, if this distribution can be modelled, then new examples can be generated similar to those in the original training set s, and then these new shapes can be examined to decide whether they represent reasonable examples. In particular, $G = M(b)$ is used to generate new vectors, where b is a vector of the parameters of the model. If the distribution parameters can be modelled, $p(b)$, these can then be limited such that the generated G's are similar to those in the training set. Similarly it should be possible to estimate $p(G)$ using the model. To simplify the problem, principal Component Analysis (PCA) is applied, to reduce the dimensionality of the data. PCA summarizes the variation seen across the data, allowing one to approximate any of the original points using a model. The model constructed here was based on 68 SEM hook images, each with 45 points and 110 points determined as the optimal number of landmark points to effectively characterise the shape of each hook. The subsequent PCA step reduced the number of extracted shape features to 22 and 49, removing redundant features and retaining those that best characterise morphological differences between the true species of *Gyrodactylus*.

PCA is used to find the major axes of a cloud of point in high dimensional space. PCA attempts to find a linear subspace of lower dimensionality than the original feature space, where the new features have the largest variance [146].

Consider a dataset G_i where $i = 1, 2, ..., N$ and each G_i is a D dimensional vector. The goal is to project the data onto an M dimensional subspace, where $M < D$. We assume the projection is denoted as $y = AG$, where $A = [u_1^T, ..., u_M^T]$, and $u_k^T u_k = 1$ for $k = 1, 2, ..., M$. We aim to maximise the variance of y_i, which is the trace of the covariance matrix of y_i.

Thus, the aim is to find $A^* = \frac{argmax}{A} tr(S_y)$, where ${}^{S}{}_{y} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{y})(y_i - \bar{y})^T$, and $\bar{y} = \frac{1}{N}\sum_{i=1}^{N} G_i$. Let S_G be the covariance matrix of G_i. Since $tr(S_y) = tr(AS_GA^T)$, by using Lagrangian multiplier and taking the derivative, we get ${}^{S}{}_{G}{}^{u}{}_{k} = \lambda_k u_k$, which means that u_k is an eigenvector of S_G. Now G_i can be presented as $G_i = \sum_{k=1}^{D}\left(G_i^T u_k\right)u_k$. G_i can be also approximated by $\tilde{G}_i = \sum_{k=1}^{M}\left(G_i^T u_k\right)u_k$, where u_k is the eigenvector of S_G corresponding to the kth largest eigenvalues.

The shape model in the ASM is given by the principal components of vectors of landmark points. The grey-level appearance model is limited to the border of the object and consists of the normalised first derivative of profiles centred at each landmark that run perpendicular to the object contour.

The use of the ASM method with regard to extraction of marginal hook features has been explored. One key point behind ASMs is that the landmark points need to be initially defined during the creation of the model and the training process. In this study, the initial landmark points are defined using a manual approach. During the initial creation of the model, the points are placed one by one. These points can then be extracted when new shapes are presented to the model. This use of randomly chosen landmark points is also utilised in other related work by others [38], [48].

It was decided to experiment with two different numbers of points, 45 points, and 110 points, both chosen arbitrarily through a trial and error basis. These two different numbers of points are then extracted and used as features, with the two sets compared to each other in order to provide the best prediction of the *Gyrodactylus* species group. This evaluation is performed by using the extracted features as inputs into a number machine learning classifier.

4.4.3 *ASM fitting*

Once the ASM model has been constructed, it is important to fit the defined model to a series of new input images to determine the parameters of the model that are the best descriptors of hook shape. ASM finds the most accurate parameters of the defined model for the new hook images. The ASM fitting attempts to "best fit" the defined model parameter to each image. Cootes *et al.* [27] explained that adjusting each model parameter from the defined model will

permit an extraction pattern of the image series to be created. Ginneken et al. [143] explain that, the fitting procedure is an alteration of landmark displacements and model fitting in a multiresolution framework. During the model fitting process, it measures newly introduced images and uses this model to correct the values of current parameters, leading to a better fit.

4.4.4 *Texture extraction*

In Active Appearance Model (AAM) only the texture information is observed and implemented. The texture means the pattern of intensities or colours across an image patch [27]. The texture extraction in AAM is refer as gray level appearance [143]. It describes the typical image structure around each landmark is obtained from pixel profiles, samples around each landmark, perpendicular to the contour. The texture eigenspace is spanned by the ℓ principle modes of $\hat{T}_i$. The texture model is $\hat{T} = \bar{T} + V_t$, where V is the matrix consisting of ℓ principal orthogonal modes of the variance in $\hat{T}_i$. While the $\bar{T}$ is the main texture and t is the vector of texture parameters.

4.4.5 *Machine learning classifiers*

Following ASM based feature extraction, the data was assessed using four models of machine learning classifiers. Selection of the models was based on the performance achieved in classifying the *Gyrodactylus* species using 25 feature morphometric dataset. The explanation of the machine learning classification used in separating the SEM images species was previously discussed in section 3.4.

The aim here is to evaluate classification with image recognition. Therefore, for consistency, the same classifiers are used as in section 3.4. The difference is that rather than the time consuming measurements used in the previous chapter, these measurements are entirely automated, and so much less time consuming. The time spent for feature extraction will be significantly shorter and reliance on the domain expert during the analysis will be reduced. Although the comparison cannot be precisely carried out, since the number of species used

are different, the results presented here aim to show that automated feature extraction and identification of multiple species of *Gyrodactylus* is possible.

4.5 RESULTS AND DISCUSSION

Although the attachment apparatus of *Gyrodactylus* consists of three main elements (*i.e* two larger centrally positioned anchors or hamuli; two connecting bars between the hamuli; and, 16 peripherally distributed marginal hooks), this study sets out to classify species based on features extracted from the sickles of the marginal hooks only. As the study is based on the analysis of biological structures, these require subsequent in order to standardise the position and format of the image. Processing to standardise the orientation of the image is applied to reduce processing time and complexity during the training and construction of the ASM model.

There are several methods that can make the computer able to recognise and to understand the images that the user presents to it. One approach for this is by using Artificial Intelligence (AI) techniques. Using an AI approach such as machine learning classification will help to augment parasitology expertise in identification and classification of images, which will provide a big contribution in the aquaculture domain, particularly *Gyrodactylus* recognition and identification.

The ASM is used in feature extraction from SEM images of the *Gyrodactylus* species. Four types of machine learning classifiers are then implemented for classification and identification in order to separate the three different species. For each approach, a 10-fold cross validation was used *i.e.* the data were divided into $k(10)$ subsets, where $k-1$ subsets were used for training and the remaining subset used as the test set. This process was repeated 10 times using a different test set on each run and the average classification performance computed. Four types of machine learning classifiers have been used for classifying and identifying the three different species of ectoparasite of genus *Gyrodactylus*.

Feature extraction information is a fundamental basis of image processing, it is necessary to point out the true information in feature extraction to get the best results from the classification process. That is why it is important to choose the right feature information. In this study, the

two features (*e.g* shape and texture) are compared to determine the best features to identify the multiple species of *Gyrodactylus*. Table 4.1 shows the comparison results between shape and texture feature information.

Table 4.1: Classification rate for multiple species of *Gyrodactylus* using ASM approach based on the texture feature extraction.

Classifier / Features information	Shape	Texture
LDA	95.71% ±6.90	79.76% ±10.68
KNN	98.33% ±5.27	81.25% ±7.39
MLP	97.06% ±4.87	92.65% ±9.01
SVM	95.59% ±5.79	94.12% ±8.66

When comparing shape and texture feature information, overall performance has shown that shape information results in better performance for all the identified classifiers. None of the classifiers used for texture information exhibit a higher performance when compared to the equivalent using shape information. These results confirm that texture features have only limited information for predicting SEM of *Gyrodactylus* species. Other research has confirmed this, as only shape information have been used in species identification [104], [126], [129], [127].

Therefore it has been decided to use shape feature information throughout this study. Two different defined number of points have been used for experimentation. This is to see if there is any difference in classification performance when varying the number of defined points during model construction. For measuring the classification and identification of the true species, four machine learning classifiers have been used in the classification procedure. The classifiers are LDA, K-NN, MLP and SVM. The results are shown in Table 4.2.

The results in Table 4.2 show that using 110 points performs better than 45 defined points. The biggest accuracy was achieved by the LDA classifier at 98.57%. Using the extracted data from 45 points, the best results were presented by the K-NN classifier at 98.33%. These results have demonstrated that the ASM method is a successful application for extraction of accurate features. The breakdown of the misclassification for every single model has been presented through confusion matrix tables.

Table 4.2: Classification rate for multiple species of *Gyrodactylus* using ASM approach.

Feature set / Classifier	**LDA**	**KNN**	**MLP**	**SVM**
45 points	95.71% ±6.90	98.33% ±5.27	97.06% ±4.87	95.59% ±5.79
110 points	98.57% ±4.52	93.57% ±8.33	98.53% ±4.64	97.06% ±5.95

The LDA classifier, using 45 defined points, was able to correctly classify the *G. truttae* to their true species, except for one specimen of *G. derjavinoides* which was classified as *G. salaris)*. This is similar to *G. salaris*, where two species have been misclassified to *G. derjavinoides* (Table 4.3). The K-NN classifier improved upon the classification of *G. salaris* and *G. derjavinoides* specimens with all being correctly classified (Table 4.4), however, one of the nine *G. truttae* specimens was misallocated as *G. salaris*. The two non-linear approaches MLP (Table 4.5) and SVM (Table 4.6) were also able to achieve high rates of correct classification, both with 97.06% and 95.59% but they were not able to improve upon the results obtained using the K-NN approach *i.e.* a correct classification rate of 98.53%. Typically, MLP and SVM classifiers provide good results if their parameters are chosen carefully. K-NN by comparison is a non-parametric approach requiring less training than MLP and SVM, which is easy to use and works well with both linear and non-linear datasets.

The equivalent confusion matrixes of classification performance using 110 defined points is presented in Table 4.7, 4.8, 4.9 and 4.10. Using the LDA classifier, almost all specimen are correctly identified (*G. derjavinoides* and *G. truttae*). Only one specimen from *G. salaris* was identified as *G. truttae* using the LDA classifier. The same goes to the K-NN performance where two species have fully correct classification. These are *G. derjavinoides* and *G. salaris*. Two specimens of *G. truttae* have been misclassified as *G. derjavinoides* and three specimens have been identified as *G. salaris*. There is very little difference in the confusion matrixes for MLP and SVM (shown in Table 4.9 and 4.10 respectively). Each one has only one species that has been misclassified. Using the MLP classifier, one specimen of *G. truttae* has been misallocated to *G. derjavinoides*. Using SVM classifier, two species are also fully correctly classified, similarly to the MLP classifier. However, in the SVM classifier, two specimens of *G. truttae* have been misclassified as *G. salaris*. Among these machine learning methods, LDA classifier has the best performance, with the MLP classifier also producing good results.

Table 4.3: A confusion matrix of *Gyrodactylus* species identification applied to the ASM extracted features implemented with the LDA classifier using 45 points. The three species are *G. derjavinoides* (*d*), *G. salaris* (*s*) and *G. truttae* (*r*).

	G. der	***G. sal***	***G. tru***	**Sum**
G. der	**24**	1	0	25
G. sal	2	**32**	0	34
G. tru	0	0	9	9
Sum	26	33	9	68

Table 4.4: A confusion matrix of *Gyrodactylus* species identification applied to the ASM extracted features implemented with the K-NN classifier using 45 points. The three species are *G. derjavinoides* (*d*), *G. salaris* (*s*) and *G. truttae* (*r*).

	G. der	***G. sal***	***G. tru***	**Sum**
G. der	**25**	0	0	25
G. sal	0	**34**	0	34
G. tru	0	1	**8**	9
Sum	25	35	8	68

Table 4.5: A confusion matrix of *Gyrodactylus* species identification applied to the ASM extracted features implemented with the MLP classifier using 45 points. The three species are *G. derjavinoides* (*d*), *G. salaris* (*s*) and *G. truttae* (*r*).

	G. der	***G. sal***	***G. tru***	**Sum**
G. der	**25**	0	0	25
G. sal	1	**33**	0	34
G. tru	0	1	**8**	9
Sum	26	34	8	68

Table 4.6: A confusion matrix of *Gyrodactylus* species identification applied to the ASM extracted features implemented to the SVM classifier using 45 points. The three species are *G. derjavinoides* (*d*), *G. salaris* (*s*) and *G. truttae* (*r*).

	G. der	***G. sal***	***G. tru***	**Sum**
G. der	**24**	1	0	25
G. sal	1	**33**	0	34
G. tru	0	1	**8**	9
Sum	25	35	8	68

Table 4.7: A confusion matrix of *Gyrodactylus* species identification applied to the ASM extracted features implemented with the LDA classifier using 110 points. The three species are *G. derjavinoides* (*d*), *G. salaris* (*s*) and *G. truttae* (*r*).

	G. der	***G. sal***	***G. tru***	**Sum**
G. der	**25**	0	0	25
G. sal	0	**33**	1	34
G. tru	0	0	**9**	9
Sum	25	33	10	68

Table 4.8: A confusion matrix of *Gyrodactylus* species identification applied to the ASM extracted features implemented with the K-NN classifier using 110 points. The three species are *G. derjavinoides* (*d*), *G. salaris* (*s*) and *G. truttae* (*r*).

	G. der	***G. sal***	***G. tru***	**Sum**
G. der	**25**	0	0	25
G. sal	0	**34**	0	34
G. tru	1	3	**5**	9
Sum	26	37	5	68

Table 4.9: A confusion matrix of *Gyrodactylus* species identification applied to the ASM extracted features implemented with the MLP classifier using 110 points. The three species are *G. derjavinoides* (*d*), *G. salaris* (*s*) and *G. truttae* (*r*).

	G. der	***G. sal***	***G. tru***	**Sum**
G. der	**25**	0	0	25
G. sal	0	**34**	0	34
G. tru	1	0	**8**	9
Sum	26	34	8	68

Table 4.10: A confusion matrix of *Gyrodactylus* species identification applied to the ASM features implemented with the SVM classifier using 110 points. The three species are *G. derjavinoides* (*d*), *G. salaris* (*s*) and *G. truttae* (*r*).

	G. der	***G. sal***	***G. tru***	**Sum**
G. der	**25**	0	0	25
G. sal	0	**34**	0	34
G. tru	0	2	7	9
Sum	25	36	7	68

The current study is based on a small set of higher quality SEM images, due to limitations with the availability of data. Despite this, the average correct classification is higher (*i.e.* 98.53% using 45 points and 98.57 using 110 points) than that achieved using the LDA method applied to the equivalent 25 point-to-point measurements manually extracted from light micrographs of 557 specimens (*i.e.* 92.59%) [8]. This approach appears promising and will be applied to hooks prepared for light microscopy hopefully with equal or better rates of correct classification. The ASM as extraction tool and the machine learning classifiers based approach applied to SEM images of the hook sickles of *Gyrodactylus* appears to outperform in comparison to the manual measurement point to point feature extraction that was applied previously. This application has been tested to identify and discriminate this species with confidence. Classification has been demonstrated to have been successfully performed.

With these successful results for extraction and classification, the difficulties faced by domain experts can be minimised. These difficulties include manual classification, a tedious and time consuming process. Another challenge in the manual approach, inaccurate point to point measurements, which results in inaccurate species identification, can also be overcome. Now, with this newly applied combination of techniques, domain experts can use these methods for feature measurement and species identification.

4.6 COMPLEX NETWORKS

Complex Networks can be described as the intersection between graph theory and statistical mechanics, which confers a truly multidisciplinary nature to this research, since it integrates computer science, mathematics and physics [15]. Nowadays, Complex Networks have become a topic of great interest in many fields of science. The main reason for the popularity of Complex Networks lies in their flexibility and generality to represent any given structure, natural or discrete, including those undergoing dynamic changes of topology [29], [30].

The main idea of this concept is to represent a shape in terms of a Watt-Strogatz network model [15], followed by analysing its topological and dynamic characteristics. This network model presents what is called the small-world property, *i.e.*, that all vertices can be reached from any other through a small number of edges.

The dynamic model of a small-world network is obtained artificially by sequential thresholds on vertices of a shape model. It has been demonstrated that shape format is correlated with small-world network structure on many stages of network growth. The study of its dynamical properties (the metrics derived from the dynamics of the network growth, based on the variation of the number of connected components) produce a shape signature, which can be used for image analysis and classification processes [15].

Recently, Complex Network based shape representation has been shown to be effectively and widely used in shape and image recognition and retrieval [15] [14], [16]. In general, this method consists of the following two steps:

1. Shape representation with Complex Network model

   First, G landmark (key) points should be extracted from the shape contour. Then, with these landmark points, the construction of a Complex Network will be designed $\hat{T} =< \hat{V}, \hat{E} >$, where node $\hat{V}_i \in \hat{E}$ and edge $(\hat{V}_i, \hat{V}_j) \in \hat{E}$ denotes the pair of neighboring vertices. Each landmark point is represented as a vertex in the network. For each pair of vertices, there is an edge with the corresponding weight $w_e f$ representing the Euclidean distance between them. Therefore, the network can be represented by a $G \times G$ weight matrix W, normalized into interval $[0, 1]$ [15], [14].

2. Feature extraction

   There are two main kinds of characteristics (measurements) that can be used to characterise topological connectivity of the Complex Network. One is static statistic measurement, and the other is dynamic evolution [15], [14]. The five static measurements used in this work are the maximum degree, average degree, average joint degree, average shortest path length, and entropy. These measurements have been used to express the topological features and their subsequent classification. The reason for this is that each Complex Network has specific topological features that characterise connectivity.

Dynamic evolution is also an important characteristic for Complex Networks. In this research, the evolution process was used as proposed in [15], [14]. By concatenating measurements the network achieved at different instants from the same underlying dynamic, the trajectories can be obtained and this value can provide more comprehensive characterization with which to analyse and classify the network. In addition, considering measurements as

a function of time, it would also be interesting to try to characterise classes of Complex Networks by considering the dynamics of the respectively defined feature space [30].Figure 4.3 shows the Complex Network representation and its dynamic evolution process.

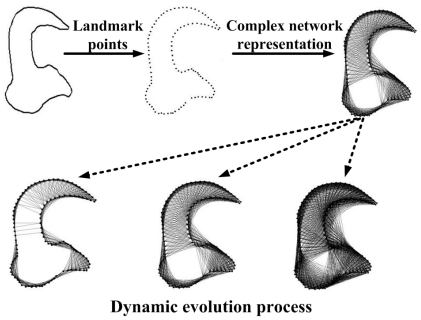


Figure 4.3: Shape representation and the dynamic evolution process of a Complex Network.

4.6.1 *Degree*

The method to describe the characteristic of vertices is degree. From the measurement of degree, other types of measurements can be carried out. $\hat{k}_i$ of node i is the number of edges directly connected to node, and it is defined in terms of the adjacency matrix $\hat{A}$ as

$$\hat{k}_{max} = \frac{max}{i}\hat{k}_i \tag{4.1}$$

and the average degree

$$\hat{k}_{avg} = \frac{1}{N}\sum_{i=1}^{N}\hat{k}_i \tag{4.2}$$

where $\hat{k}_i$ is the degree of node i.

To calculate the amount of order, disorder and chaos in a system, an entropy has been applied. It is defines as

$$\hat{n} = \sum_{\hat{k}=1}^{\hat{k}_{max}} p(\hat{k})logp(\hat{k}) \tag{4.3}$$

where $p(\hat{k})$ is the degree of distribution.

4.6.2 *Joint degree*

Once the average degree has been defined, then the correlation between degree of two vertices are quantified. Correlation is a consequential role in many properties of structural and dynamic networks. This can represented by the joint degree distribution $p(\hat{k}, \hat{k}')$, *i.e.*, the probability that an edge connects to a vertex of degree $\hat{k}$ with a vertex of degree $\hat{k}'$. Here, consider the case in which $\hat{k} = \hat{k}'$, where they have same degree. Predicted on the joint degree probability, various measurements can be extracted such as entropy, energy and the average joint degree.

The entropy for the joint degree distribution is defined as

$$\hat{h}_d = -\sum_{\hat{k},\hat{k}'=1}^{\hat{k}_{max}} p(\hat{k}, \hat{k}') \log p(\hat{k}, \hat{k}') \tag{4.4}$$

The enegy for the joint degree distribution is defined as

$$\hat{E} = \sum_{\hat{k},\hat{k}'=1}^{\hat{k}_{max}} p(\hat{k}, \hat{k}') \tag{4.5}$$

4.6.3 *Shortest path*

The other measurement is distance. This is defined by computing the mean value of geodesic distance for each pair of vertices as

$$d_{\hat{G}} = \frac{1}{N(N-1)} \sum_{i \neq j} d_{ij} \tag{4.6}$$

where d_{ij} is the geodesic distance between vertex i and vertex j.

4.6.4 *Feature Extraction using Active Shape Model and Complex Network Model*

This section describes feature extraction algorithm using ASM combined with a Complex Network model. A Complex Network is considered as a feature extraction tool because of their successful use in extracting and classifying an object, as reported in [15] [14], [16]. In this proposed method, the ASM method is used to plot the landmark point in order to get the contour of an image. ASM was found to be a successful technique to use for segmentation

of an image. As explain previously, the SEM images of the *Gyrodactylus* species are difficult to segment due to the tissue surrounding the focused object. Instead of using PCA as the extraction tool, a Complex Network based on graph theory is used.

The proposed feature extraction algorithm consists of the segmentation of an object, feature extraction, and classification. Fig. 4.4 shows the process involved in extracting the features from SEM images of multiple species of *Gyrodactylus*. The attributes of the framework are:

(1) Segmentation. The segmentation has to be performed correctly in order to provide a clean and clear representation of the focused object. As discussed before, the SEM images are considered difficult images to work on, since they are often surrounded with tissue. For that reason, the ASM model was used to segment the images using the plotting of landmark points. Using the ASM method, the contour of the images can be produced.

(2) Feature extraction. Feature extraction is required to be performed after segmentation. Feature extraction is therefore executed to produce data that can be used for classification and identification purposes. A complex model based on graph theory has been decided to be used for extracting informative features for classification.

(3) Classification. In a similar manner to other experiments, four machine learning classifiers were tested to measure the feature extraction approach. The four machine learning classifiers are LDA, K-NN, MLP, and SVM.

The entire process, from image segmentation using ASM to final classification, is presented in Fig. 4.4 and is described in detail as follows:

1. Given $G = g_1h_1, ..., g_rh_r$, where g_sh_s are the coordinate points.

2. Perform image segmentation using the ASM method. Then, produce the set of landmark points as $\hat{G} = G_1, G_2, ..., G_s$.

    a) Construct the ASM model.

    b) Fit the model.

3. Once the SEM images are ready for feature extraction, the contour images are then processed using the Complex Network model. The Complex Network method is used for extracting feature information for the purpose of species identification.

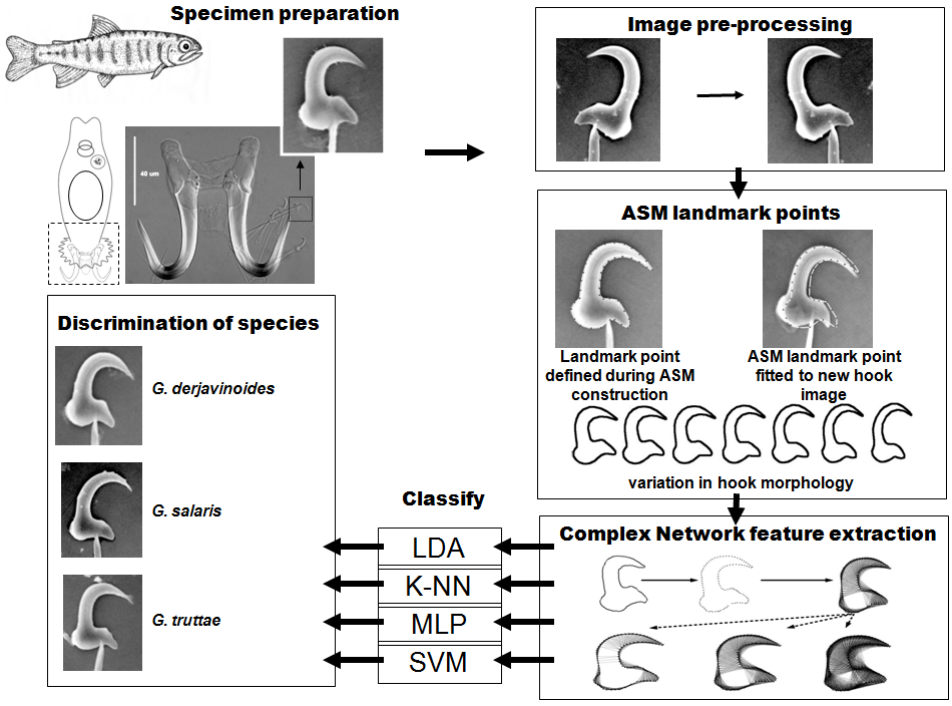


Figure 4.4: The methodological approach used for extracting the features from the marginal hook. The images were pre-processed before being subjected to an Active Shape Model and Complex Network for the feature extraction step. These features were used to train 4 classifiers (K-NN, LDA, MLP, SVM) and separate the three species of *Gyrodactylus*, which includes the notifiable pathogen, *G. salaris*. Abbreviations: K-NN, K Nearest Neighbors; LDA, Linear Discriminant Analysis; MLP, Multi-Layer Perceptron; SVM, Support Vector Machine.

a) Shape representation, where a Complex Network model is designed so that $\hat{T} =< \hat{V}, \hat{E} >$.

b) Feature extraction. The dynamic evolution process [15] has been selected to be used for feature extraction for the Complex Network model.

4. Classification can only be performed when the features are available. Four machine learning classifiers are used to evaluate the feature extraction strategy. These four machine learning classifiers are:

a) Linear Discriminant Analysis (LDA)

$$\hat{N}_i = W_0 + W_{i1}D_1 + W_{i2}D_2 + ... + W_{in}D_n$$

b) K-Nearest Neighbor (K-NN)

$$(X_t, X_u) = \sqrt{\sum_{mm=1}^{m} (X_t^{mm} - X_u^{mm})^2}$$

c) Multi-layer perceptron (MLP)

$$\hat{R}_j(X_p) = \varphi_{\breve{R}} \left(\sum_{s=0}^{z} V_{rs} \varphi_s \left(\sum_{c=0}^{zz} V_{sc} D_{cn} \right) \right)$$

d) Support vector machine (SVM)

$$\hat{S} \sum_{i=1}^{ii} Y^i \breve{K}(X^i, X) + \breve{b}$$

5. The results of the classification are analysed and the classifier performance is measured. Three species from 68 specimens of *Gyrodactylus* species are used for identification. These are: *G. derjavinoides*, *G. salaris* and *G. truttae*.

4.6.5 *Results and discussion*

This section contains the results obtained by classifying 68 SEM images of *Gyrodactylus* specimens using the ASM - CN method using linear and non-linear machine learning classifiers. The experiment strategy for the implementation of ASM - CN feature extraction to the SEM images of *Gyrodactylus* species is very similar to the experiments presented in section 4.5. The only difference is the method used for feature extraction. In this experiment, the same number of images are considered, with the same number of landmark points. The objective of the implementation of Complex Network model as a feature extraction tool is to apply the ASM method together with Complex Network model in order to extract and correctly classify

multiple species of *Gyrodactylus*. The results of the implementation of the combined ASM and Complex Network model are produced by the following parameters, as shown in Table 4.11.

Table 4.11: The parameter settings for Complex Network model feature extraction.

Parameter Name	**Value**
Initial threshold	0.025
Final threshold	0.5
Separation	0.075
Number of points	45 / 110

Table 4.12 presents the average classification rate with standard deviation of three species of *Gyrodactylus* using 10-fold cross validation for the training and testing strategy. The four selected machine learning classifiers are LDA, K-NN, MLP and SVM. The results in Table 4.12 show a comparison of classification performance of machine learning classifiers applied to different numbers of feature extraction landmark points. Using the 45 landmark points, the SVM classifier has achieved the highest accuracy (97.06%), followed by the MLP classifier at 94.12%. On the other hand, when considering 110 landmark points, the highest performance was found with the MLP classifier. These results conclude that the selection of the number of landmark points plays an important role in considering the method of extraction and classification.

Using the 45 landmark points, the SVM classifier improved upon the classification of *G. salaris* specimens with all being correctly classified (Table 4.16), while two more species remain misclassified; such as *G. salaris* specimen being misallocated as *G. truttae* and *G. truttae* being

Table 4.12: The average classification rate of *Gyrodactylus* species; performance with Linear (*i.e.* LDA and K-NN) and non-linear (*i.e.* MLP and SVM) machine learning classifiers from the hooks of each parasite, extracted using ASM and Complex Network approaches.

Feature set / Classifier	**LDA**	**KNN**	**MLP**	**SVM**
45 points	86.19 ±11.49	90.00 ±15.13	94.12 ±7.35	97.06 ±6.09
110 points	83.57 ±14.57	92.80 ±10.20	98.36 ±4.71	95.59 ±8.87

Table 4.13: A confusion matrix of the 45 points of ASM - CN feature extraction implemented to the SEM *Gyrodactylus* images using LDA classifier.

	G. der	***G. sal***	***G. tru***	**Sum**
G. der	**25**	0	0	25
G. sal	4	**28**	2	34
G. tru	0	2	**7**	9
Sum	29	30	9	68

Table 4.14: A confusion matrix of the 45 points of ASM - CN feature extraction implemented with the SEM *Gyrodactylus* images using the K-NN classifier.

	G. der	***G. sal***	***G. tru***	**Sum**
G. der	**22**	1	2	25
G. sal	1	**33**	0	34
G. tru	3	0	**6**	9
Sum	26	34	8	68

misclassified as *G. salaris*. Other classifier model results are shown for LDA (Table 4.13), K-NN (Table 4.14) and MLP (Table 4.15).

In addition to the 45 landmark points considered for extracting the features for classification, 110 points were also considered. The objective is to see if there is any improvement in classification if the number of landmark points increased. With the increment in the number of points, more features are therefore added. Amongst all the classifiers considered, the MLP classifier(Table 4.19) has performed well, with only one specimen misclassified as *G. truttae*. Other confusion matrixes of different classifiers have been presented in Table 4.17, Table 4.18 and Table 4.20.

This performance is same as that obtained using ASM-PCA presented by Ali *et al.* [9], and this is better than the 25 point-to-point measurements manually extracted from light micrographs of 557 specimens (*i.e.* 92.59%) [8], this approach appears promising and will in future, be applied to hooks prepared for light microscopy with the expectation of equal or better rates of correct classification. The ASM and Complex Network based approach applied to SEM images of the hook sickles of *Gyrodactylus* appears to outperform or equal other methods that have been tested to identify this species with confidence.

Table 4.15: A confusion matrix of the 45 points of ASM - CN feature extraction implemented with the SEM *Gyrodactylus* images using the MLP classifier.

	G. der	*G. sal*	*G. tru*	**Sum**
d	**25**	0	0	25
s	0	**32**	2	34
t	0	2	**7**	9
Sum	25	34	9	68

Table 4.16: A confusion matrix of the 45 points of ASM - CN feature extraction implemented with the SEM *Gyrodactylus* images using the SVM classifier.

	G. der	*G. sal*	*G. tru*	**Sum**
d	**25**	0	0	25
s	0	**33**	1	34
t	0	1	**8**	9
Sum	25	34	8	68

Table 4.17: A confusion matrix of the 110 points of ASM - CN feature extraction implemented with the SEM *Gyrodactylus* images using the LDA classifier.

	G. der	*G. sal*	*G. tru*	**Sum**
d	**24**	0	1	25
s	1	**28**	5	34
t	0	2	**7**	9
Sum	25	30	13	68

Table 4.18: A confusion matrix of the 110 points of ASM - CN feature extraction implemented with the SEM *Gyrodactylus* images using the K-NN classifier.

	G. der	*G. sal*	*G. tru*	**Sum**
d	**23**	1	1	25
s	0	**34**	0	34
t	2	1	**6**	9
Sum	25	36	7	68

Table 4.19: A confusion matrix of the 110 points of ASM - CN feature extraction implemented with the SEM *Gyrodactylus* images using the MLP classifier.

	G. der	*G. sal*	*G. tru*	**Sum**
d	**25**	0	0	25
s	0	**33**	1	34
t	0	0	**9**	9
Sum	25	33	10	68

Table 4.20: A confusion matrix of the 110 points of ASM - CN feature extraction implemented with the SEM *Gyrodactylus* images using the SVM classifier.

	G. der	*G. sal*	*G. tru*	**Sum**
G. der	**25**	0	0	25
G. sal	0	**32**	2	34
G. tru	0	1	**8**	9
Sum	25	33	7	68

Referring to the experimental results of the previously discussed ASM-PCA approach, there is not a large difference in the accuracy of classification, either using only ASM as the feature extraction tool, or a combined ASM and Complex Network model. In the case of extracting the SEM *Gyrodactylus* species images, the ASM approach was determined to be the best method for segmentation of the images. Other approaches [95], [83], [92], [123] are not suitable due to the physical appearance of the SEM images, that have additional tissue surrounding the object of interest (the attachment hook). In addition, the integration of a ASM and Complex Network to create a new feature extraction method appears to produce promising initial results that will be expanded upon in future work.

Backes *et al.* [16] used different parameters to those that were used in this study, where the direct images were used in the feature extraction procedure instead of the landmark points. In the case of *Gyrodactylus* species, using the SEM images directly is not possible because the target object is surrounded with the tissue. Another segmentation approach is needed to remove these noises. Otherwise, the main object is not accurately analysed. As discussed before, even though the classification performance is slightly poorer (0.17% different) than using the ASM-PCA technique discussed earlier, the combination of ASM and Complex Network model represents an improvement in terms of image segmentation. Backes *et al.*

[16] has introduced a method that applies a Complex Network to extract images using black (background) and white (contour). However, this method is not fully relevant to the problem of Gyrodactylus. This can be seen in figure 4.5, which shows the two images of the same subject. The segmentation of the images is incorrect as the images also include the noise (tissue) present in the SEM images. This will naturally lead to poor performance with regard to prediction of the true species.

In addition, in work by Tang *et al.* [137], they apply a Complex Network to model the graph structure. In order to model this graph structure, they extract the topological and dynamic characteristics of the Complex Network. To conclude, this work focuses on representation of the images rather than image extraction, while our focus is on extraction of the valuable features for the purpose of classification of multiple species.

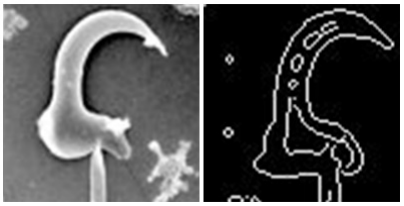


Figure 4.5: Original image (left) and segmentation of image (right).

The combination of ASM with a Complex Network model has created significant potential for improving the functionality and performance of the Complex Network model approach. As discussed before, in a Complex Network model, there is no function to segment the images before the feature extraction process. Conventionally, this step will be performed manually. As discussed previously, when the process is performed manually, it can very easily lead to inaccuracy in segmentation. An additional drawback is if there are many images that require segmentation, which means that this process will therefore take a significant time period to complete. The combination of the two approaches for successfully extracting the SEM images of multiple species of *Gyrodactylus* has been shown to be feasible, and allows for a reliable and automatic feature extraction process, rather than being reliant on a time consuming and labour intensive manual feature extraction approach. With the combination of ASM and a Complex Network, the issue of segmentation is solved. The current limitation with this approach is that the accuracy of this combined method is slightly less when compared to the ASM-PCA approach (0.17%), but the drop in performance is very small, and the convenience

of this approach, and the potential for future improvement, makes this approach viable and promising.

4.7 CONCLUSIONS

In this chapter, research into image analysis and processing has been discussed and presented. While many linear and non-linear machine learning classifiers have previously been demonstrated to successfully classify multiple species of *Gyrodactylus*, this chapter demonstrated *Gyrodactylus* species identification using two different numbers of point sets, generated by a new application of ASM feature extraction to this domain, linked with machine learning classification. The primary aim of this research was the determination of which input information is required to produce robust species discrimination.

Extraction methods based on statistical models, such as ASM, have been shown to obtain good results in many applications [106], [85], [136]. With those successful results taken into account, and other options being considered to be less well suited to this problem domain, the ASM method was chosen to be applied for extracting the SEM *Gyrodactylus* images. There are three species that were tested for correct identification. These are *G. derjavinoides* (n = 25), *G. salaris* (n = 34) and *G. truttae* (n =9). The experiments presented in this chapter showed that the approach presented here was successful and efficient with regard to extracting features and accurately classifying multiple species of fish parasite.

To the knowledge of the author, this is the first computer-based classification of ectoparasite of genus monogenea *Gyrodactylus* species that has been successfully demonstrated through the application of pattern recognition techniques of marginal hook pattern SEM images. ASM applied to 68 SEM images of the marginal hook sickle were able to overcome the limitations and difficulties of extracting feature information from the hooks. The best approach, which used 110 points and was identified to be the LDA method, was able to improve upon the performance of previous manual point to point measurement approaches (*i.e.* 98.57% cf. 92.59% using an LDA-based classifier applied to manually extracted morphometric data).

In this chapter, we also explored the utility of a novel combined ASM and Complex Network based approach in extracting and thus classifying the ectoparasites of genus *Gyrodactylus*

species. A novel ASM and Complex Network model was applied to the 68 SEM images of the marginal hook sickle, and was able to overcome the limitation and difficulties in extracting feature information from the hooks. The result indicates that the best classifier in identification of these multiple species is the MLP classifier with 110 landmark points. It was decided to integrate a Complex Network with the ASM method because Complex Network models have much potential for future use and are a novel subject of interest in the image processing research domain.

Although the combination of ASM and Complex Network does not improve the classification performance significantly, (*i.e.* ASM-PCA has achieved competative accuracy of classification performance between 98.57% and ASM-Complex Network has achieved 98.36%), there is the opportunity to improve the performance further. The chief benefit of our new combined approach is that the segmentation of the images for the purpose of feature extraction has been improved. Before, the segmentation was required to take place manually, which was time consuming and prone to human error. In this work, the ASM was used as segmentation by providing the landmark points, then a Complex Network was used to extract the features.

The final chapter concludes the thesis as a whole, providing an overview of the research and the original contributions presented in this work, and outlining some proposed future research directions.

5

CONCLUSIONS

5.1 SUMMARY

In this research, a number of intelligent approaches have been proposed for the identification of multiple monogenean parasites of genus *Gyrodactylus*, whose members are a common fish ectoparasite within aquaculture and wild capture fisheries, with more than 409 species identified to date [61]. The newly designed ensemble classification and signal image processing research for *Gyrodactylus* specimens presented in this thesis was motivated by several factors. Firstly, as an industry, aquaculture continues to expand worldwide, but this expansion has been accompanied by increased disease problems including those associated with ectoparasite monogenean worms. A second motivational factor was that the discrimination of pathogenic from non-pathogenic species is a key current requirement in order to allow for control and management of pathogens of wild and farmed fish at domain expert, industry and government level. However, a shortage of taxonomic experts and the shortcomings of molecular methods often make this requirement difficult to achieve in practice. Another significant motivation was the desire to make use of a novel combination within this research domain of an intelligent ensemble model and signal image processing, including techniques that have been nominally described as AI, to provide the opportunity to develop state-of-the-art automated or semi-automated models (including more intelligent models) for pathogen recognition. The state-of-the-art combination of these techniques will allow for rapid, consistent and secure initial identification of pathogens by field workers and non-expert users.

Based on the motivations described above, the goal of the research reported in this thesis was primarily to develop an ensemble based majority voting method by combining a number of successful classifiers and a combination of different feature sets, and also applying a number of advanced state-of-the-art image processing techniques. These have been investigated and

evaluated in this research project for providing the basis for the extraction of features that contribute to a rapid, secure and accurate recognition of species. The novel proof of concept framework presented in this thesis can be divided into two parts, as due to limitations with availability, two separate and distinct datasets have been used. As a consequence of these data limitations, resulting in a smaller number of SEM images available, it was found that a full and comprehensive demonstration of the whole framework was not possible, and so the evaluation was divided into a number of sections to take account of the strongest data availability. The first part of the final evaluation is the ensemble based majority voting, which is applied to the 557 *Gyrodactylus* images. The second stage of the evaluation process focuses on the feature extraction methodology, and is applied to the 68 SEM *Gyrodactylus* images. Both of these evaluations share the same objective, to provide a consistent and accurate classification and identification of the *Gyrodactylus* species.

In this thesis, chapter 2 presented the background to the inter-discplinary research problem. As described, the aim of this advanced research is to provide a novel mechanism for predicting the correct class of *Gyrodactylus* species, making use of state-of-the-art technology. This chapter summarised the background to the problem, focusing initially on introducing and providing a full description of the *Gyrodactylus* species. This chapter also discussed the spread of the species and the significant global impact in terms of both food supply and economic consequences.

After discussing the motivation and the background to this thesis, chapter 3 presented an original contribution of this thesis. This chapter presented an initial solution to the issue of identifying the multiple species considered in this project, using artificial intelligence techniques. In this chapter, a number of machine learning classifiers and feature selection methods have been reviewed and applied to evaluate their suitability with regard to solving the target problem. To evaluate these approaches, 557 specimen of nine different species of *Gyrodactylus* have been used. The results showed the benefits of the different classification approaches, which to the best knowledge of the author, have not been applied to this research problem previously. The conventional method used for extracting features for this morphometric dataset was based on point-to-point measurements, which is labour extensive and time consuming. For this reason, it was decided that it would be appropriate for intelligent signal

image processing to be explored, to investigate if applying this technology for assisting the domain expert in making quick and accurate identification of the species was viable.

The initial limitation of the preliminary experimental results was identified to be the misclassification errors present. Investigation of other state-of-the-art research approaches influenced development of a refined ensemble based majority voting approach, and also improvement of the feature extraction method. The ensemble-based model proposed and implemented in chapter 3 is designed to allow for rapid, consistent and secure initial identification of pathogens by field workers and non-expert users. To construct an ensemble model, two types of classifiers (*i.e.* LDA and K-NN) and three sets of features (*i.e.* 25 features, 21 features and 20 features) were used as components in the system. These were augmented by the use of a majority voting method as part of an ensemble strategy to combine these methods and produce an improved output. New experimental results presented in this chapter demonstrated that by using an ensemble approach, performance accuracy was increased and the number of misclassification errors was minimised, representing a further refinement of the initial results presented previously.

Chapter 4 presented another original contribution. In this chapter, a number of intelligent image processing techniques have been reviewed by focusing on feature extraction. The chapter demonstrates the application of intelligent image processing applied to feature extraction. For this, ASM based processing was applied to the 68 SEM images of three different species. Unlike the morphometric features, for SEM images, only features from the marginal hook sickle have been considered rather than the addition of both the hamuli and ventral bar features. ASMs applied to SEM images of the marginal hook sickle were able to overcome the limitations and difficulties of extracting feature information from the hooks, and demonstrated that this approach can be used successfully. The same classifiers as discussed in chapter 3 were used for measuring the success of the ASM method. The best approach, which used 110 feature points and was identified to be the LDA classifier method, was able to improve upon the performance of previous manual point-to-point measurement approaches (*i.e.* 98.57%, compared to 92.59% using a LDA-based classifier applied to manually extracted morphometric data). With this improvement, it was shown in this research that it is possible to apply ASMs to assist the domain expert with regards to feature extraction, and thus, successfully performing speech

classification. As an additional point of investigation, to assist the domain expert in analysis of SEM images, a refined approach for image processing was also explored. The initial ASM model was refined by integrating a Complex Network. This was implemented in a similar way to the ASM based method also presented in chapter 4, except that in this research, the ASM has been used only for identification of the landmark points. After this initial identification, the Complex Network was then used to extract the features, which to the knowledge of the author, was the first integration of these techniques to solve this research problem. The initial ground-breaking results indicate that it is possible to achieve good classification results using this novel method.

Overall, this thesis presents an investigation of novel state-of-the-art ensemble classification and image processing for classification of the genus *Gyrodactylus* (Monogenean). Although two separate datasets have been used due to data availability limitations outwith the control of the author, the overall objective of the thesis is achieved. These are, the combination of different models for classification and identification through the ensemble method, and feature extraction using ASM with the novel integration of a Complex Network model. A number of classification techniques were investigated, and these were then paired with state-of-the-art image processing approaches. These were combined with an ensemble approach, which used majority voting to produce improved performance compared to using individual classifiers or feature points.

5.2 CONCLUSIONS

There are a number of key conclusions that can be drawn from this research:

1. The 557 image morphometric dataset discussed in this research has been successfully used for classification and identification using machine learning classifiers and feature selection techniques. This dataset consisted of images from nine different sub-species of *Gyrodactylus*. These were *G. arcuatus*, *G. derjavinoides*, *G. gasterostei*, *G. kherulensis*, *G. salaris*, *G. sommervillae*, *G. thymalli*, *G. truttae* and *G. cichilidarum*. The results of carrying out classification using this data were assessed using LDA, K-NN, MLP and SVM. In addition to the classification, four feature sets have also been evaluated. These consisted

of the original 25 features, 21 selected features using SFS, the 20 features identified using the SBS method, and seven features that were found by using the SFFS method. The highest classification result using this dataset was found to be 97.67% accuracy, obtained by using the MLP classifier and the full set of 25 features.

2. One issue that this research identified was that when applying a single classification technique, the misclassification rate was generally found to be high and therefore, arguably, there could be a lack of confidence in the accuracy of the final result. To resolve this issue, an ensemble-based majority voting system was proposed. In this proposed algorithm, two classifiers (LDA and K-NN) together with three sets of features (the set of 25, 21, and 20 features respectively) have been combined into an integrated classification system, with majority voting applied to make the final classification decision by considering all of these variables rather than one single approach. In this work it was decided to make use of LDA and K-NN because when comparing these approaches to the MLP and SVM techniques, the misclassification errors were found to be similar, meaning there was little additional value from the integration of all four classifiers. The principle of ensemble system is the combination of the models must have independent errors, so that the ensemble system can perform better.

3. In addition to the morphometric dataset, the use of an image processing dataset has also been explored in this research. The contents of this dataset were taken from SEM images. Only the marginal hook or contour information of each *Gyrodactylus* was used to extract the features required in order to perform species identification. To extract these features, a combination of two techniques were applied to each image in the dataset. The first technique to be applied was the ASM approach. ASM was used to segment the images by taking into account only the shape of the object. This is a critical task, since the target image (the hook) is surrounded by tissue, which may act as noise in any classification using the raw images. Only then, after the successful removal of the additional tissue and the extraction of only the relevant object, was the second method applied, the Complex Network model. The Complex Network Model was used to extract the features for classification and prediction of the species class. This novel application was found to successfully assist scientific research in the Aquaculture domain to accurately identify

the ectoparasite species in a shorter time, than taking the time consuming measurements required otherwise. In addition to this, in the computer science domain, the proposed combined application of ASM and the Complex Network model has a great deal of potential, as it allows the segmentation process to be integrated within the Complex Network model for extracting feature information. This is an exciting area of future research.

5.3 FUTURE WORK

There are a number of potential future research directions that can follow on from the research presented in this thesis. These include:

1. The number of images within the dataset could be increased. This would make it possible for further application of the complex models. Currently, the SEM image dataset only consists of 68 images from three different species. Once an increased quantity of images becomes available, additional research and a considerably more detailed and comprehensive, including statistical, evaluation can be performed, a benchmarked against a range of state-of-the-art signal image processing approaches that were not employed in this thesis.

2. Other than the ensemble methodology, there are a number of other techniques that could be explored in order to extract the most informative features. One potential technique that could be explored is skeleton graph matching [17]. In this method, a skeleton graph is matched by comparing the geodesic paths between skeleton endpoints. The main idea is that it will identify matches based on the similarity of the shortest paths between each pair of endpoints. It could be possible to apply this method in the case of SEM *Gyrodactylus* images used in the research presented in this thesis, where the images have varieties of scale affected by variations in temperature, even though the original shape remains the same for similar species. One example of this is the case of *G. salaris* from Norway and *G. salaris* from Italy. The size of these two may be slightly different, but the shape remains the same for both.

3. Further enhancements need to be explored to the novel combined ASM and Complex Network based feature extraction algorithm proposed in this thesis. In addition, alternative unsupervised machine learning based approaches, such as novel incremental, slow feature analysis [96] and manifold-based machine learning algorithms (including newly developed locally optimised laplacian eigenmap and globally optimised isometric projection algorithms [97], Malik *et al.* [98]) can be used for simultaneous *invariant* feature extraction, dimensionality reduction, and real-time clustering and visualization (in appropriately projected low-dimensional latent space). Newly developed unsupervised canonical correlation based machine learning methods [100] could also be employed, to identify features that maximise correlations, and hence determine similarities between the extracted features and their cumulative effects on the classification outcomes. A newly developed, cognitively-inspired graph theory-based real-time clustering technique [3] could be employed to analyse and extract any connected clusters of features that may otherwise go undetected with traditional cluster extraction techniques.

4. For more enhanced prediction and classification, other state-of-the-art techniques in addition to ensemble based majority voting employed in this study, could be applied in future research. These would enable an investigation into the possibility of further improvements in the classification results, and thus a reduction in misclassification errors. The current problem that is not fully solved is that certain classifiers are very good at classifying certain species, while performing poorly with other species. To solve this issue, it could be possible to implement the recently proposed approach termed multi-target regression with rule ensembles [6]. The rule ensemble approach is used to ensemble decision trees into a large collection of rules. An optimization approach is then used to select the best subset of these rules and to determine their respective weights.

5. As noted earlier, a detailed comparative performance-complexity trade-off analysis needs to be carried out of the various classifiers employed in this study, in order to optimise the developed ensemble based classifier approach. Further, other state-of-the-art supervised machine learning based classifiers could also be employed and comparatively evaluated as part of ensemble or multi-stage classifiers, including feedforward neural networks such as adaptively regularized multi-class logistic regression models

[103], incremental linear discriminative analysis with extreme learning machines [99] and temporal recurrent neural network models [96], [100]. A range of state-of-the-art *unsupervised* machine learning based classifiers can also be employed and comparative evaluated in the future.

6. Finally, the versatility and flexibility of the work presented in this thesis can be investigated. This can be accomplished by evaluating whether the groundbreaking integrated model, which combines state-of-the-art feature extraction, a range of classification techniques, all integrated with an ensemble method to successfully identify *Gyrodactylus* can be successfully applied to identify data from a different species of Monogenea, such as *Dactylogylus*.

BIBLIOGRAPHY

[1] Nobanis - invasive alien species fact sheet gyrodactylus salaris. URL `www.nobanis.org`.

[2] Monogenean parasites of fish. URL `http://edis.ifas.ufl.edu/fa033`.

[3] A. Abdullah and A. Hussain. A cognitively inspired approach to two-way cluster extraction from one-way clustered data. *Cognitive Computation*, pages –, 2015.

[4] E. Acuña, F. Coaquira, and M. Gonzalez. A comparison of feature selection procedures for classifiers based on kernel density estimation. *Computer, Communication and Control Technologies*, 1:468–472, 2003.

[5] A. Ahmad and L. Dey. A feature selection technique for classificatory analysis. *Pattern Recognition Letters*, 26:43–56, 2005.

[6] T. Aho, B. Zenko, and T. Elomaa.

[7] T. Aho, B. Zenko, and S. Dzeroski. Rule ensemble for multi-target regression. *Data Mining*, pages 21–30, 2009.

[8] R. Ali, A. Hussain, J. E. Bron, and A. P. Shinn. Multi-stage classification of *Gyrodactylus* species using machine learning and feature selection techniques. *Intelligent Systems Design and Applications*, pages 457–462, 2011.

[9] R. Ali, A. Hussain, J. E. Bron, and A. P. Shinn. The use of asm feature extraction and machine learning for the discrimination of members of the fish ectoparasite genus *Gyrodactylus*. *Neural Information Processing*, 7666:457–462, 2011.

[10] R. Ali, J. Bo, A. Hussain, L. Bin, J. E. Bron, and A. P. Shinn. Classification of fish ectoparasite genus gyrodactylus sem images using asm and complex network model. *Neural Information Processing*, 8836:103–110, 2014.

[11] M. K. S. Alsmadi, K. Omar, S. A. Noah, and I. Almarashdah. Fish recognition based on the combination between robust features selection, image segmentation and geometrical

parameters techniques using artificial neural network and decision tree. *Computer Science and Information Security*, 6(2):215–221, 2009.

[12] J. Alvarez-Borrego and M. C. Chávez-Sánchez. Detection of IHHN virus in shrimp tissue by digital color correlation. *Aquaculture*, 194:1–9, 2001.

[13] A. Ault, X. Zhong, and E. J. Coyle. K-nearest neighbor analysis of received signal strength distance estimation across environments. *Wireless Network Management*, 2005.

[14] A. R. Backes and O. M. Bruno. Shape classification using complex network and multi-scale fractal dimension. *Pattern Recognition Letters*, 31(1):45–51, 2010.

[15] A. R. Backes, D. Casanova, and O. M. Bruno. A complex network-based approach for boundary shape analysis. *Pattern Recognition*, 42(8):54–67, 2009.

[16] A. R. Backes, A. S. Martinez, and O. M. Bruno. Texture analysis using graphs generated by deterministic partially self-avoiding walks. *Pattern Recognition*, 44(8):1684–1689, 2011.

[17] X. Bai and L. J. Latecki. Path similarity skeleton graph matching. *Pattern Analysis and Machine Intelligence*, 30(7):1282–1292, 2008.

[18] T. A. Bakke, J. Cable, and P. D. Harris. The biology of gyrodactylid monogeneans: the "Russian-doll killers". *Advances in Parasitology*, 64:161–376, 2007.

[19] Kristin P. Bennett and Ayhan Demiriz. Semi-supervised support vector machines. In *Advances in Neural Information Processing Systems*, pages 368–374. MIT Press, 1998.

[20] J. M. Blackledge and A. Dubovitskiy. Object detection and classification with applications to skin cancer screening. *Intelligent Systems*, 1(1):34–45, 2008.

[21] H. Bouziane, B. Messabih, and A. Chouarfia. Profiles and majority voting-based ensemble method for protein secondary structure prediction. *Evolutionary Bioinformatics*, 7: 171–189, 2011.

[22] M. Bramer. Ensemble classification. *Principles of Data Mining*, pages 209–220, 2013.

[23] Y. Chen, F. Chen, J. Y. Yang, and M. Q. Yang. Ensemble voting system for multiclass protein fold recognition. *Pattern Recognition*, 22(4):747–763, 2008.

[24] Qi Cheng, P. K. Varshney, and M. K. Arora. Logistic regression for feature selection and soft classification of remote sensing data. *Geoscience and Remote Sensing Letters*, 3(4): 491–494, 2006.

[25] G. S. Choudhury and C. G. Bublitz. Electromagnetic method for detection of parasites in fish. *Aquaculture Food Product Technology*, 185(2):883–893, 1994.

[26] C. Colak and C. Isik. Feature subset selection for blood pressure classification using orthogonal forward selection. *Bioengineering*, pages 122–123, 2003.

[27] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. *Pattern Analysis and Machine Intelligence*, 23(6):681–685, 2001.

[28] Luigi Pietro Cordella, Ro Limongiello, and Carlo Sansone. Network intrusion detection by a multi-stage classification system. In *In: Roli, Kittler, and Windeatt (Eds.): Multiple Classifier Systems, LNCS 3077*, pages 324–333. Springer, 2004.

[29] L. da. F. Costa. Complex network. *Simple Vision*, 2004.

[30] L. da. F. Costa, Rodriguesm F. A., G. Travieso, and P. R. Villas Boas. Characterization of complex networks: a survey of measurements. *Advances in Physics*, 56(1):167–242, 2007.

[31] S. F. Cotter, K. Kreutz-Delgado, and B. D. Rao. Feature subset selection for blood pressure classification using orthogonal forward selection. *Signal Processing*, 81:1849–1864, 2001.

[32] C. O. Cunningham, D. M. McGillivray, K. MacKenzie, and W. T. Melvin. Discrimination between *Gyrodactylus salaris*, *G. derjavini* and *G. truttae* (Platyhelminthes: Monogenea) using restriction fragment length polymorphisms and an oligonucleotide probe within the small subunit ribosomal RNA gene. *Parasitology*, 111:87–94, 1995.

[33] C.O. Cunningham, D.M. McGillivray, K. MacKenzie, and W.T. Melvin. Identification of *Gyrodactylus* (monogenea) species parasitizing salmonid fish using DNA probes. *Fish Diseases*, 18:539–544, 1995b.

[34] P. Cunningham and S. J. Delany. K-nearest neighbor classifiers. *Technical Report UCD-CSI, University College, Dublin*, 2007.

[35] A. Dasgupta, P. Drineas, and B. Harb. Feature selection methods for text classification. *Knowledge Discovery and Data Mining*, pages 230–239, 2007.

[36] V. H. C. de Albuquerque, A. R. de Alexandria, P. C. Cortez, and J. M. R. S. Travers. Evaluation of multilayer perceptron and self-organizing map neural network topologies applied to microstructure segmentation from metallographic images. *NDT & E International*, 42:644–651, 2009.

[37] M. de Bruijine, B. van Ginneken, M. A. Viergever, and W. J. Niessen. Adapting active shape models for 3d segmentation of tubular structures in medical images. *Information Processing in Medical Imaging*, 18:136–147, 2003.

[38] V. de Silva and J. Tenenbaum. Sparse multidimensional scaling using landmark points (technical report). *Standford University*, pages 1–41, 2004.

[39] C. Demir and B. Yener. Automated cancer diagnosis based on histopathological images: s systematic survey. *Technical Report TR-05-09, Computer Science Department at Rensselaer Polytechnic Institute*, 2005.

[40] T. G. Dietterich. An experimental comparison for three methods for constructing ensembles of decision trees: bagging, boosting, and randomization. *Machine Learning*, 40:139–157, 2000.

[41] T. G. Dietterich. Ensemble methods in machine learning. *Multiple Classifier System*, 1857: 1–15, 2001.

[42] D. Dolmen. *Gyrodactylus* salaris (monogenea) in norway; infestations and management in: Parasites and diseases in natural waters and aquaculture in nordic countries stenmark, a & malmberg. g. (eds.). *Zoo-Tax-Symposium*, pages 63–69, 1987.

[43] S. Doraisamy, S. Golzari, N. M. Norowi, M. N. B. Sulaiman, and N. I. Udzir. A study on feature selection and classification techniques for automatic genre classification of traditional malay music. *Music Information Retrieval*, 2008.

[44] J-X. Du, X-F. Wang, and G-J. Zhang. Leaf shape based plant species recognition. *Applied Mathematics and Computation*, 185(2):883–893, 2007.

[45] R. Duangsoithong and T. Windeatt. Bootstrap feature selection for ensemble classifiers. *Advances in Data Mining. Applications and Theoretical Aspects*, 6171(, YEAR =).

[46] R. O Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*, pages 1–19. Wiley Interscience Publication, second edition, 2001.

[47] H. Fan and K. Ramamohanarao. A bayesian approach to use emerging patterns for classification. *Australian Database Conference*, 17, 2003.

[48] Q. Fang, L. Gao, V. Guibas, de Silva, and L. Zhang. Gradient landmark-based distributed routing for sensor networks. *Communications Society*, pages –, 2005.

[49] K. Farooq, J. Karasek, H. Atassi, A. Hussain, PeiPei Yang, C. MacRae, M. Mahmud, Bin Luo, and W. Slack. A novel cardiovascular decision support framework for effective clinical risk assessment. *Computational Intelligence in Healthcare and e-health*, pages 117–124, 2014.

[50] E. Frank, M. Hall, L. Trigg, G. Holmes, and I. H. Witten. Data mining in bioinformatics using weka. *Bioinformatics*, 20(15):2479–2481, 2004.

[51] C. O. A. Freitas, J. M. de Carvalho, and J. J. Oliveira. Confusion matrix disagrement for multiple classifiers. *Progress in Pattern Recognition, Image Analysis and Applications*, 4756: –, 2007.

[52] Zhouyu Fu, A. Robles-Kelly, and Jun Zhou. Mixing linear svms for nonlinear classification. *Neural Networks*, 21(12):1963–1975, 2010.

[53] D. Garrett, D. A. Peterson, C. W. Anderson, and M. H. Thaut. Comparison of linear, nonlinear, and feature selection methods for eeg signal classification. *Neural Systems and Rehabilitation Engineering*, 11(2):141–144, 2003.

[54] P. Geurts, M. Fillet, D. d. Seny, M-A. Meuwis, M-P. Malaise, M. Merville, and L. Wehenkel. Proteomic mass spectra classification using decision tree based ensemble methods. *Bioinformatics*, 21(15):3138–3145, 2005.

[55] I. A. Gheyas and L. S. Smith. Feature subset selection in large dimensionality domains. *Pattern Recognition*, 437:5–13, 2010.

[56] D.I. Gibson. Monobothrium wageneri: another imported tapeworm established in wild british freshwater fishes? *Fish Biology*, 43:281–285, 1993.

[57] I. Gokce and J. Peng. Comparing linear discriminant analysis and support vector machines. *Advance in Information Systems*, 2457:104–113, 2002.

[58] M. H. Goldbaum, P. A. Sample, K. Chan, J. Williams, T-W. Lee, E. Blumenthal, C. A. Girkin, L. M. Zangwill, C. Bowd, T. Sejnowski, and R. N. Weinreb. Comparing machine learning classifiers for diagnosing glaucoma from standard automated perimetry. *Investigative Ophthalmology and Visual Science*, 43(1):162–169, 2002.

[59] M. A. Hall and L. A. and Smith. Feature subset selection: A correlation based filter approach. *Neural Information Processing and Intelligent Information Systems*, pages 885–858, 1997.

[60] H. Hansen, L. Bachmann, and Bakke T. A. Mitochondrial DNA variation of *Gyrodactylus* spp. (Monogenea, Gyrodactylidae) populations infecting Atlantic salmon, grayling, and rainbow trout in Norway and Sweden. *Parasitology*, 33:1471–1478, 2003.

[61] P. D. Harris, A. P. Shinn, J. Cable, and T. A. Bakke. Nominal species of the genus *Gyrodactylus* v. Nordmann 1832 (Monogenea: Gyrodactylidae), with a list of principal host species. *Systematic Parasitology*, 59:1–27, 2004.

[62] C. C. Homes and N. M. Adams. A probabilistic nearest neighbor method for statistical pattern recognition. *Royal Statistical Society*, 66(2):295–306, 2002.

[63] S-L. Hsieh, S-H. Hsieh, and P-H. Cheng. Design ensemble machine learning model for breast cancer diagnosis. *Medical Systems*, 36(5):2841–2847, 2012.

[64] Laura Igual, Xavier Perez-Sala, Sergio Escalera, Cecilio Angulo, and Fernando De la Torre. Continuous generalized procrustes analysis. *Pattern Recognition*, 47(2):659–671, 2014.

[65] I. Inza, P. Larranaga, R. Blanco, and A. Cerrolaza. Filter versus wrapper gene selection approaches in dna microarray domains. *Artificial Intelligence in Medicine*, 31:91–103, 2004.

[66] A. Jayachandran and R. Dhanasekaran. Automatic detection of brain tumor in magnetic resonance images using multi-texton histogram and support vector machine. *Imaging Systems and Technology*, 23(2):97–103, 2013.

[67] B. O. Johnsen, E. P. Ieshko, A. Karasev, A. J. Jansen, and I. Schurov. Report on joint research on gyrodactylus salaris in the northern region of norway and russia. *NINA-NIKU Project Report*, 9(1):1–20, 1999.

[68] B. O. Johnsen, A. J. Jensen, and P. I. Mϕkkelgjred. Gyrodactylus salaris pålaks i norske vassdrag, statusrapport ved innagengen til år 2000. (in norwegian). *NINA Oppdragsmelding*, 617:1–129, 1999.

[69] K. Jong, J. Mary, A. Cornuéjols, E. Marchiori, and M. Sebag. Ensemble feature ranking. *Knowledge Discovery in Databases*, pages 267–278, 2004.

[70] M. F. Kabir, D. L. Schmoldt, P. A. Araman, M. E. Schafer, and S-M. Lee. Classifying defects in pallet stringers by ultrasonic scanning. *Wood and Fiber Science*, 34, 2003.

[71] L. Kainulainen. Ensembles of locally linear models: Application to bankcruptcy prediction. *Data Mining*, pages 280–286, 2010.

[72] M. Karagiannopoulos, D. Anyfantis, S. B. Kotsiantis, and P. E. Pintelas. Feature selection for regression problems. *Hellenic European Research on Computer Mathematics & Its Applications*, pages 20–22, 2007.

[73] Kailash Jagannath Karande and Sanjay Nilkanth Talbar. Canny edge detection for face recognition using ica. In *Independent Component Analysis of Edge Information for Face Recognition*, pages 21–33. Springer, 2014.

[74] M. Kass, A. Witkin, and D. Terzopoulos. Snakes: active contour models. *Computer Visions*, 1(4):321–331, 1988.

[75] R. Kasturi, L. O'Gorman, and V. Govindaraju. Document image analysis: A primier. *Sadhana, Indian Academy of Sciences, Special Issue on Indian Language Document Analysis and Understanding*, 27(1):3–22, 2002.

[76] J. W. Kay, A. P. Shinn, and C. Sommerville. Towards an automated system for the identification of notifiable pathogens: using *Gyrodactylus salaris* as an example. *Parasitology Today*, 15(5):201–203, 1999.

[77] Hyunjoong Kim, Hyeuk Kim, Hojin Moon, and Hongshik Ahn. A weight-adjusted voting algorithm for ensembles of classifiers. *Korean Statistical Society*, 40(4):437 – 449, 2011.

[78] R. Kohavi and G. H. John. Wrappers for feature subset selection. *Artificial Intelligence*, pages 273–324, 1997.

[79] V. Kolodyazhniy, S. D. Kreibig, J. J. Gross, and W. T. Roth. An affective computing approach to physiological emotion specificity: Toward subject-independent and stimulus-independent classification of film-induced emotions. *Psychophysiology*, pages 1–15, 2011.

[80] P. Komarek and A. Moore. Making logistic regression a core data mining tools with tr-irls. *Data Mining*, page 4, 2005.

[81] S. B. Kotsiantis. Supervised machine learning: A review of classification techniques. *Informatica*, 31:249–268, 2007.

[82] G. Kumar and P. K. Bhatia. A detailed review of feature extraction in image processing systems. *Advanced Computing and Communication Technologies*, pages 321–331, 2014.

[83] C. H. Lai, S. S. Yu, H. Y. Tseng, and M. H. Tsai. A protozoan parasite extraction scheme for digital microscope images. *Computerized Medical Imaging and Graphics*, 34:122–130, 2010.

[84] B. M. Larsen and K. Haarsaker. Freshwater pearl mussel Margaritifera margaritifera and stocking of fish in the river hoenselve and bingselva, buskerud countary in norway. *NINA Fagrapport*, 56:1–33, 2002.

[85] HT. Le and NT. Vo. Face alignment using active shape model and support vector machine. *Biometrics and Bioinformatics*, 4(6):224–234, 2012.

[86] S. Li, R. Xia, and C. Huang. A framework of feature selection methods for text catagorization. *Natural Language Processing*, 2:692–700, 2009.

[87] T. Li, S. Zgu, and M. Ogihara. Using discriminant analysis for multi-class classification: an experimental investigation. *Knowledge Information System*, 10(4):453–472, 2006.

[88] T. Linderstrom, C. M. Collins, J. Breciani, C. O. Cunningham, and K. Buchmann. Characterization of a *Gyrodactylus* salaris variant: infection bilogy, morphology and molecular genetics. *Parasitology*, 127:165–177, 2003.

[89] M. Liu and C. Wan. Feature selection for automatic classification of musical instrument sounds. *Joint Conference on Digital Libraries*, pages 247–248, 2001.

[90] S. Loncaric. A survey of shape analysis techniques. *Pattern Recognition*, 31(9):983–1001, 1998.

[91] P. Love. Fisheries: While stocks last? OECD *Insights,* OECD P*ublishing*, pages 32–44, 2010.

[92] A. Lu, X. Hou, X. Chen, and C. Lio. Insect species using sparse representation. *British Machine Vision Association*, pages 1–10, 2010.

[93] J. Lu, K. N. Plataniotis, and A. N. Venetsanopoulos. Face recognition using LDA based algorithms. *Neural Networks*, 2002.

[94] R. Madhanagopal, R. C. Avinaash, and K. Karthick. Comparison of support vector machines and linear discriminant analysis for indian industries. *Statistika Mathematika*, 4(3):74–80, 2012.

[95] R. Maini and H. Aggarwal. Study and comparison of various image edge detection techniques. *Image Processing*, 3(1):1–60, 2009.

[96] Z. Malik, A. Hussain, and J. Wu. Novel biologically inspired approaches to extracting online information from temporal data (submitted revise). *Cognitive Computation*, 6(3):–, 2014.

[97] Z. Malik, A. Hussain, and J. Wu. An online generalised eigenvalue version of laplacian eigenmap for visual big data (in press). *Neurocomputing*, pages –, 2015.

[98] Z. Malik, A. Hussain, and J. Wu. A novel generalized isometric projection approach for on-line learning (revised submitted). *Knowledge and Data Engineering*, pages –, 2015.

[99] Z. Malik, A. Hussain, and J. Wu. A neural implementation of linear discriminant analysis and extreme learning machine. *Neural Networks and Learning System*, pages –, 2015.

[100] Z. Malik, A. Hussain, and J. Wu. Extracting online information from dual and multi-data streams (in press). *Neural Computation and Applications*, pages –, 2015.

[101] H. Mark, F. Eiba, H. Geoffrey, P. Bernhard, R. Peter, and H. Witten Ian. The weka data mining software. *Special Interest Group on Knowledge Discovery and Data Mining*, 11(1):–, 2009.

[102] G. Martínez-Muñoz, D. Hernandez-Lobato, and A. Suarez. The excretory systems and marginal hooks as a basis for the systematics of *Gyrodactylus* (Trematoda, Monogenea). *Arkiv för Zoologi Series 2-Band 23 (nr. 1)*, pages 1–235, 1970.

[103] T. Mazzocco and A. Hussain. Novel logistic regression models to aid the disgnosis of dementia. *Expert Systems with Applications*, 39(3):3356–3361, 2012.

[104] S. E. McHugh, A. P. Shinn, and J. W. Kay. Discrimination of *G. salaris* and *G. thymalli* using statistical classifiers applied to morphometric data. *Parasitology*, 121:315–323, 2000.

[105] M. Meinilä, J. Kuusela, M. Zietara, and Lumme J. Brief report: Primers for amplifying 820 bp of highly polymorphic mitochondrial COI gene of *Gyrodactylus salaris*. *Hereditas*, 137:72–74, 2002.

[106] Nicola A. Miller, Jennifer S. Gregory, Richard M. Aspden, Peter J. Stollery, and Fiona J. Gilbert. Using active shape modeling based on {MRI} to study morphologic and pitch-related functional changes affecting vocal structures and the airway. *Voice*, (0):–, 2014.

[107] T. Mitchell, 1997. Lecturer slides for textbook machine learning, McGraw Hill.

[108] T.A. Mo. Status of *Gyrodactylus* salaris problems and research in norway. *Parasitic Diseases of Fish*, pages 43–56, 1994.

[109] M. P. Morant, G. Prpich, E. Peeler, M. Thrush, S. A. Rocks, and S. J. T. Pollard. Assessment of consequences of notifiable fish disease incursions in england and wales. *Human and Ecological Risk Assessment*, 19(1):278–290, 2013.

[110] T. Moyo, S. Bangay, and G. Foster. The identification of mammalian species through the classification of hair patterns using image pattern recognition. *Computer Graphic, Virtual Reality, Visualisation and Interaction*, pages 177–181, 2006.

[111] G. Paladini, A. Gustinelli, M.L. Fioravanti, H. Hansen, and A.P Shinn. The first report of *Gyrodactylus salaris* malmberg, 1957 (platyhelminthes, monogenea) on italian cultured stocks of rainbow trout (oncorhynchus mykiss). *Vet. Parasitol*, 165:290–297, 2009.

[112] Hong-xia Pang, Wen-de Dong, Zhi-hai Xu, Hua-jun Feng, Qi Li, and Yue-ting Chen. Novel linear search for support vector machine parameter selection. *Zhejiang University SCIENCE*, 12(11):885–896, 2011.

[113] M. Pohar, M. Blas, and S. Turk. Comparison of logistic regression and linear discriminant analysis: a simulation study. *Netodoloski zvezki*, 1(1):143–161, 2004.

[114] P. Pudil, F. J. Ferri, J. Novovicova, and J. Kittler. Floating search methods for feature selection with nonmonotonic criterion functions. *Pattern Recognition*, pages 279–283, 1994.

[115] Z. Qiao, L. Zhou, and J. Z. Huang. Sparse linear discriminant analysis with application to high dimension low sample size data. *Application Mathematic*, 39(1):48–60, 2009.

[116] P. Reed, R. Francis-Floyd, and R.E. Klinger. Monogenean parasites of fish. *University of Florida, IFAS Extension*, 2005.

[117] P. Refaeilzadeh, L. Tang, and H. Liu. Cross-validation. *Arizona State University*, 2008.

[118] Lior Rokach. Ensemble-based classifiers. *Artificial Intelligence Review*, 33(1-2):1–39, 2010.

[119] F. Roli, J. Kittler, G. Fumera, and D. Muntoni. An experimental comparison of classifier fusion methods for multimodal personal identity verification systems. *Multiple Classifier Systems*, 2364:325–336, 2002.

[120] H. Roubus, M. Setnes, and J. Abonyi. Learning fuzzy classification rules from data. *Development in Soft Computing*, pages 108–115, 2001.

[121] Y. Saeys, I. Inza, and P. Larrañage. A review of feature selection techniques in bioinformatics. *Bioinformatics*, 23(1):2507–2517, 2007.

[122] Mohammad Salahshoor, Ali Broumandnia, and Maryam Rastgarpour. Application of intelligent systems for iranian license plate recognition. In *Intelligent Systems*, pages 1–6. IEEE, 2014.

[123] S. S. M. Sallah, F. A. Hussin, and M. Z. Yusoff. Shape-based road sign detection and recognition for embedded application using matlab. *Intelligent and Advanced Systems*, pages 1–5, 2010.

[124] A. P. Shinn, S. d. Clers, C. Sommerville, and D. I. Gibson. Multivariate analyses of morphometrical features from gyrodactylus spp. (monogenea) parasitising british salmonids: light microscope based studies. *Systematic Parasitology*, 33(2):115–125, 1995.

[125] A. P. Shinn, C. Sommerville, and D. I. Gibson. Distribution and characterization of species of gyrodactylus nordmann, 1832 (monogenea) paratizing salmonids in the uk, and their discrimination from g. salaris malmberg, 1957. *Natural History*, 29:1383–1402, 1995.

[126] A. P. Shinn, J. W. Kay, and C. Sommerville. The use of statistical classifier for the discrimination of species of the genus *Gyrodactylus* (Monogenea) parasitizing salmonids. *Parasitology*, 120:261–269, 2000.

[127] A. P. Shinn, D. I. Gibson, and C. Sommervillae. Morphometric discrimination of *Gyrodactylus* salaris malmberg (monogenea) from species of *Gyrodactylus* paratising british salmonids using novel parameters. *Fish Diseases*, 24:83–97, 2001.

[128] A. P. Shinn, J. E. Bron, C. Sommervillae, and D. I. Gibson. Comments on the mechanism of attachment in species of the monogenean genus *Gyrodactylus*. *Invertebrate Biology*, 122 (1):1–11, 2003.

[129] A. P. Shinn, H. Hansen, L. Bachmann, and T. A. Bakke. The use of morphometric characters to discriminate specimens of laboratory-reared and wild populations of *Gyrodactylus salaris* and *G. thymalli* (monogenea). *Folia Parasitologica*, 51:239–252, 2004.

[130] A. P. Shinn, C. Collins, A. García-Vásquez, M. Snow, G. Paladini, T. Lindenstrøm, M. Longshaw, I. Matějusová, D. M. Stone, J. F. Turnbull, S. M. Picon-Camacho, C. Vázquez Rivera, R. A. Duguid, T. A. Mo, H. Hansen, K. Olstad, J. Cable, P. D. Harris, R. Kerr, D. Graham, G. H. Yoon, K. Buchmann, R. Raynard, S. Irving, and J. E. Bron. Multi-centre testing and validation of current protocols for *Gyrodactylus salaris* (Monogenea) identification. *International Journal of Parasitology*, 40:1455–1467, 2010.

[131] M. H. Sigari, M. R. Pourshahabi, and H. R. Pourreza. Offline handwritten signature identification and verification using multi-resolution gabor wavelet. *Biometrics and Bioinformatics*, 5(4):234–248, 2011.

[132] A. R. Sofia Visa, B. Ramsay, and E. van der Knaap. Confusion matrix-based feature selection. *CEUR Workshop Proceeding*, 710, 2011.

[133] P. Somol, B. Baesens, and P. Pudil. Filter- versus wrapper-based feature selection for credit scoring. *Intelligent Systems*, 20:985–999, 2005.

[134] P. Somol, N. ŏvaăJ, and P. Pudil. Flexible-hybrid sequential floating searching statistical features selection. *Structural, Syntactic, and Statistical Pattern Recognition*, 4109:632–639, 2006.

[135] J. Song, Y. Huang, D. Zhou, H. Zha, and C. L. Giles. IKNN: Informative k-nearest neighbor pattern classification. *Springer-Verlag Berlin Heidelberg*, pages 248–264, 2007.

[136] Sϕgaard. Weed classification by active shape models. *Automation and Emerging Technologies*, 91(3):271–281, 2005.

[137] J. Tang, Bo. Jiang, C-C. Chang, and B. Luo. Graph structure analysia based on complex network. *Digital Signal Processing*, 22:713–725, 2012.

[138] R. Tedín, J.A. Becerra, and Richard J. Duro. Using classifiers as heuristics to describe local structure in active shape models with small training sets. *Pattern Recognition Letters*, 34(14):1710 – 1718, 2013.

[139] M.H. Terra and R. Tinós. Fault detection and isolation for robotic system using a multilayer perceptron and a radial basis function network. *Systems, Man, and Cybernatics*, 2:1880–1885, 1998.

[140] L. S. Thota and S. B. Changalasetty. Optimum learning rate for classification problem with mlp in data mining. *Advances in Engineering and Technology*, 6(1):35–44, 2013.

[141] R. daS. Torres, L. Falcäo, and L. da. F. Costa. A graph-based approach for multiscale shape analysis. *Pattern Recognition*, 37:1163–1174, 2003.

[142] O. D. Trier, A. K. Jains, and T. Taxt. Feature extraction methods for character recognition - a survey. *Pattern Recognition*, 29(4):641–662, 1996.

[143] B. van Ginneken, A. F. Frangi, J. J. Staal, B. M. ter Haar Romeny, and M. A. Viergever. Active shape model segmentation with optimal features. *Medical Imaging*, 21(8):924–933, 2002.

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

[145] S.K. Wajid, A. Hussain, and Bin Luo. An efficient computer aided decision support system for breast cancer diagnosis using echo state network classifier. *Computational Intelligence in Healthcare and e-health*, pages 17–24, 2014.

[146] Q. Wang. Kernel principal component analysis and its applications in face recognition and active shape models. *arXiv preprint arXiv:1207.3538*, pages –, 2012.

[147] T. Xiong and V. Cherkassky. A combined SVM and LDA approach for classification. *Proceedings of the International Joint Conference on Neural Networks, Montreal, Canada*, pages 1455–1459, 2005.

[148] A. Yazdani, T. Ebrahimi, and U. Hoffmann. Classification of eeg signals using dempster shafer theory and k-nearest neighbor classifier. *Neural Engineering*, pages 327–330, 2009.

[149] H. Yu and S. Xu. Simple rule-based ensemble classification for cancer dna microarray data classification. *Computer Science and Service System*, pages 2555–2558, 2011.

[150] U. Zakir, A. Usman, and A. Hussain. A novel road traffic sign detection and recognition approach by introducing ccm and lesh. *Neural Information Processing*, 7665:629–636, 2012.

[151] C. Zhang and M. Yunqian. Ensemble machine learning: methods and applications. *Computational Intelligence and Complexity*, 7:6–8, 2012.

[152] J. Zhang and I. Mani. kNN approach to unbalanced data distributions: a case study involving information extraction. *International Conference Machine Learning*, 2003.

[153] D. Zheng, M. Na, and J. Wang. Face recognition using an nnsrm classifier in lda subspace. *Pattern Analysis Application*, 10:375–381, 2007.

[154] C. S. Zuria, Ramirez J. M., D. Baez-Lopez, and G. E Flores-Verdad. Matlab based image processing lab experiments. *Frontiers in Education*, 3:1255–1258, 1998.