Automated Ladybird Identification using Neural and Expert Systems

Mohd Zaki Ayob

Doctor of Philosophy

University of York

Department of Electronics

August 2012

Abstract

The concept of automated species identification is relatively recent and advances are being driven by technological advances and the taxonomic impediment. This thesis describes investigations into the automated identification of ladybird species from colour images provided by the public, with an eventual aim of implementing an online identification system. Such images pose particularly difficult problems with regards to image processing as the insects have a highly domed shape and not all relevant features (e.g. spots) are visible or are fore-shortened. A total of 7 species of ladybird have been selected for this work; 6 native species to the UK and 3 colour forms of the Harlequin ladybird (*Harmonia axyridis*), the latter because of its pest status. Work on image processing utilised 6 geometrical features obtained using greyscale operations, and 6 colour features which were obtained using CIELAB colour space representation. Overall classifier results show that inter-species identification is a success; the system is able to, among all, correctly identify *Calvia 14-guttata* from *Halyzia 16-guttata* to 100% accuracy and *Exochomus 4-pustulatus* from *H. axyridis* f. *spectabilis* to 96.3% accuracy using Multilayer Perceptron and J48 decision trees. Intra-species identification of *H. axyridis* shows that *H. axyridis* f. *spectabilis* can be identified correctly up to 72.5% against *H. axyridis* f. *conspicua*, and 98.8% correct against *H. axyridis* f. *succinea*. System integration tests show that through the addition of user interaction, the identification between Harlequins and non-Harlequins can be improved from 18.8% to 75% accuracy.

Contents

List of Figures

List of Tables

Acknowledgements

The thesis is a contribution of all perseverance, perspiration and knowledge from people whom I know personally, and also indirectly. I would like to thank my sponsors, Majlis Amanah Rakyat (MARA) and Universiti Kuala Lumpur of Malaysia, for their financial support towards my full-time PhD research. I would like to thank everyone in the Department of Electronics for their continuous support in my research, especially to my thesis advisory panel, Professor John Robinson for his invaluable insights on my PhD work. I am grateful to my External Examiner, Dr. Jonathan Clark for his comments and sincere advice. I am deeply indebted to my supervisor, Dr. E.D. Chesmore firstly for giving me the opportunity to study computer aided taxonomy, and secondly, for his patience and guidance to shape a learning researcher. My gratitude goes to Bio-Applied Systems Engineering senior lab members- Dr. Oliver Bunting, Dr. John Stammers, Dr. Naoko Evans and Dr. James Schofield- for the memorable experience and assistance. I would like to thank Dr. Peter Brown from Centre for Ecology & Hydrology (CEH) for providing some stock ladybird images. I am grateful to my loving wife, Siti Noraini, for her constant support throughout my studies. Her devotion and the love of my children- Fauzan, Fatini, Farisya and Fawwaz- encouraged the completion of the tasks. My special thanks go to all my teachers and colleagues at Universiti Kuala Lumpur, Malaysia and close friends for all their constant and genuine support. I dedicate this thesis to my parents; without their constant advices, initiatives, effort and prayers I would not be able to achieve my targets in life so far, may Allah bless you.

Author’s declaration

I hereby declare that the research presented in this thesis is my own work and that it has not been submitted for any award elsewhere. Where other sources of information have been used they have been appropriately acknowledged and full references included. I declare that this research was carried out under the supervision of Dr. E. D. Chesmore in the Department of Electronics at the University of York and was co-sponsored by Universiti Kuala Lumpur and Majlis Amanah Rakyat, Malaysia.

CHAPTER 1

INTRODUCTION

CHAPTER 1

INTRODUCTION

This thesis explores the use of colour images for the identification of ladybirds in UK. Imagine people using their mobile devices for uploading images through entomological websites and receiving feedback from web queries. Nearly four years ago, the author embarked on a research to produce a framework for the automated identification of ladybirds in the UK. The author realised the massive amount of technical work involved, even though nowadays the use of computers in identification is far more advanced than those days when William Dawson of California took some photos of Lesser yellow-legs bird (*Totanus flavipes*) near Santa Barbara for identification on August 16, 1913 (Dawson, 1913). *T. flavipes* has always been confused with a Tennessee Warbler. Dawson realised using images from a camera did not actually improve identification process, except when the location of where the photograph was taken is known. This is completely understandable considering manual identification using camera was rare in those days; one would require the availability of equipment, skills, technology and

financial ability. The moral of the story is to have a computerised identification system that can perform as good as an expert, but with greater efficiency.

This research project has included the use of image processing, neural networks and expert system in the identification of UK ladybirds and Harlequin ladybirds (*Harmonia axyridis*). Due to potential impacts on economy and biodiversity, the task of monitoring the spread of *H. axyridis* is highly important to the UK (UK Harlequin Survey, n.d.). However it is a huge task to be done manually, considering various challenges in term of technicality, logistics, data collection and funding. Morphology seems to be the best approach and still useful, as other alternatives such as DNA-based identification is too expensive. Furthermore, it is not readily available for users, and best handled by the experts themselves. Automation using ladybird images as inputs would be beneficial, and this thesis will show the framework of implementation. The purpose of this work is to determine whether automated identification of ladybird species, including *Harmonia axyridis*, is possible.

1.1 Ladybirds in the UK

Ladybirds are beetles (Order: Coleoptera, family: Coccinellidae), and called ladybugs in the USA and some parts of the world. As many as 46 ladybird species have been identified in the UK (UK Ladybird Survey, n.d.). However, this project focussed on 26 most prominent ladybird species only. A list of 26 ladybird species with complete Latin names and authority is given in Appendix I. The list will be referred to in this thesis when abbreviations are used. From the twenty six species, investigations have concentrated on only the more common 6 species and 1 invasive species, *H. axyridis*.

Ladybirds are distinct from insects of other orders in two ways: they have hard forewings that cover the abdomen and meet centrally, and they have biting mouthparts (UK Ladybird Survey, n.d.). They also have a few general features that distinguish themselves from other families; the most obvious is their coloured spots. The body is divided into three parts; head, thorax and abdomen, as shown in Figure 1.1.

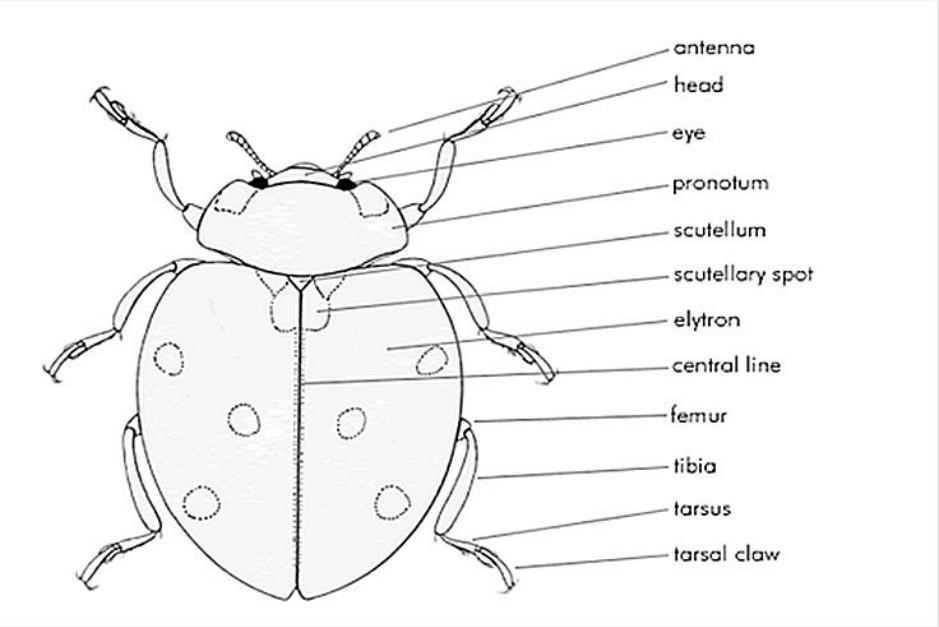
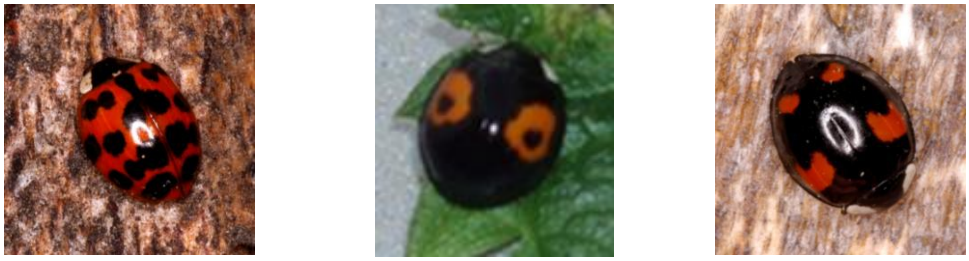


Figure 1.1: Anatomy of a 7-spot ladybird viewed from top (UK Ladybird Survey, n.d.)

The most visible parts of the ladybird body are the elytra (so called 'wing covers'), which cover the abdomen. The elytra are highly coloured and are predominantly red, black, brown or yellow. They almost invariably have spots of contrasting colour, for example, the 22-spot ladybird has a yellow background with black spots, whereas the 7-spot ladybird is red with black spots. In some species the pronotum is a different colour to the elytra and is generally black and white. Many ladybird species are polymorphic, and have a number of different colour forms. For instance, the 2-spot ladybird has two colour forms; one has two black spots on a red background, and the other has four red spots on a black background. In addition, the spotting patterns can be very variable with some specimens having no spots, for example, the Larch ladybird. Others have most spots joined together, for instance, the 24-spot ladybird.

1.2 Harlequin Ladybirds

The Harlequin ladybird has a large number of colour forms and spot patterns. The three most identifiable colour forms in the UK are form *succinea* (orange with 18 or 19 black spots), form *conspicua* (black with 2 red spots and black inner spots or "bull's eyes") and form *spectabilis* (black with 4 red spots or crescents); refer to Figure 1.2 for images of Harlequin ladybirds. For visual comparison purpose, readers can observe their similarities with some local ladybirds in Figure 1.3.



(a) Form *succinea* **(b) Form *conspicua*** **(c) Form *spectabilis***

Figure 1.2: Three different forms of *Harmonia axyridis*

(a) Orange ladybird

(b) Eyed ladybird

(c) 10-spot ladybird

Figure 1.3: Examples of native ladybirds commonly mistaken as *Harmonia axyridis*

In terms of distinguishing the Harlequin from other species, there are a few features or characteristics which can be helpful. In terms of size and shape, the Harlequin is generally large and the length is between 5 mm to 8 mm. It is generally quite round and domed. The elytron colour can be highly variable; common colours are pale yellow-orange, orange-red and black. The spots can vary between 0 to 21 orange-red or black spots, and may be in a grid pattern, as in the case of the form *succinea*. In UK the most common colour forms are orange with 15 to 21 black spots, and black with 2 or 4 orange or red spots. The pronotum pattern can be white or cream in colour. It can contain up to 5 spots or fused lateral spots forming 2 curved lines, M-shaped mark or solid trapezoid. The elytra have a wide keel at the back, and the legs are almost always brown. With regard to the above characteristics, the main species that the Harlequin can be confused with are: 10-spot ladybird, Orange ladybird, Eyed ladybird and Cream-streaked ladybird. A complete list of these, together with their scientific names, is presented with the other UK species in Table A1, Appendix I. As shown in Figure 1.3, these species are similar in colour to Harlequins and have many spots.

The Harlequin ladybird is an invasive species that originates from Asia (UK Harlequin Survey, n.d.). Scientifically known as the *Harmonia axyridis*, it was first detected in Britain in Essex on 19 September 2004 (Majerus, Strawson and Roy, 2006). Beforehand, this ladybird species has been sold as biological control throughout Europe since 1982 and becoming established in northern France, Germany, Luxembourg and Holland (Katsoyannos *et al.*, 1997; Iperti & Bertand, 2001). Despite its initial biological use, there have been studies in the USA reporting the adverse effects of *H. axyridis* that outweighs their biological potential (Majerus, Strawson and Roy, 2006). It was used as a biological control agent in the United States in 1988, where it is now the most widespread ladybird species. It is an aggressive and voracious predator that feeds on aphids as its main food source. When aphids are scarce they resort to intra-guild interactions, for instance, lacewings, hoverfly larvae and other ladybird species of which they dominate (Ware and Majerus, 2008). Averaging between 6 to 8 mm in size, they are bigger than many local species, which is advantageous to them. In fact, there is scientific evidence of attack on the following British species; *Coccinella septempunctata*, *Adalia bipunctata*, *Thea vigintiduopunctata* and *Propylea quatuordecimpunctata* (Majerus, Strawson and Roy, 2006; Ware and Majerus, 2008; Brown *et al.*, 2011). So far only the following species are likely to be the least threatened by the establishment of *H. axyridis* in UK: *Thea 22-punctata* (L.) (Coleoptera: Coccinellidae), *Subcoccinella 24-punctata* (L.) (Coleoptera: Coccinellidae) and *Coccinella magnifica Redtenbacher* (Coleoptera: Coccinellidae) (Pell *et al.*, 2008). Harlequins like to over-winter in large groups to hibernate within sheds, attics and parts of buildings where the locations are dry and protected. When they are disturbed, they emit a foul secretion to deter predators. This may stain fabrics and

may cause skin irritations (Just Green, n.d.). *H. axyridis* has now spread to most parts of England and requires considerable attention due to its impact on ecological and biological balance. The distributions of *H. axyridis* in UK from 2003 until 2010 are shown in Figure 1.4 (NERC/Field Studies Council, 2010).

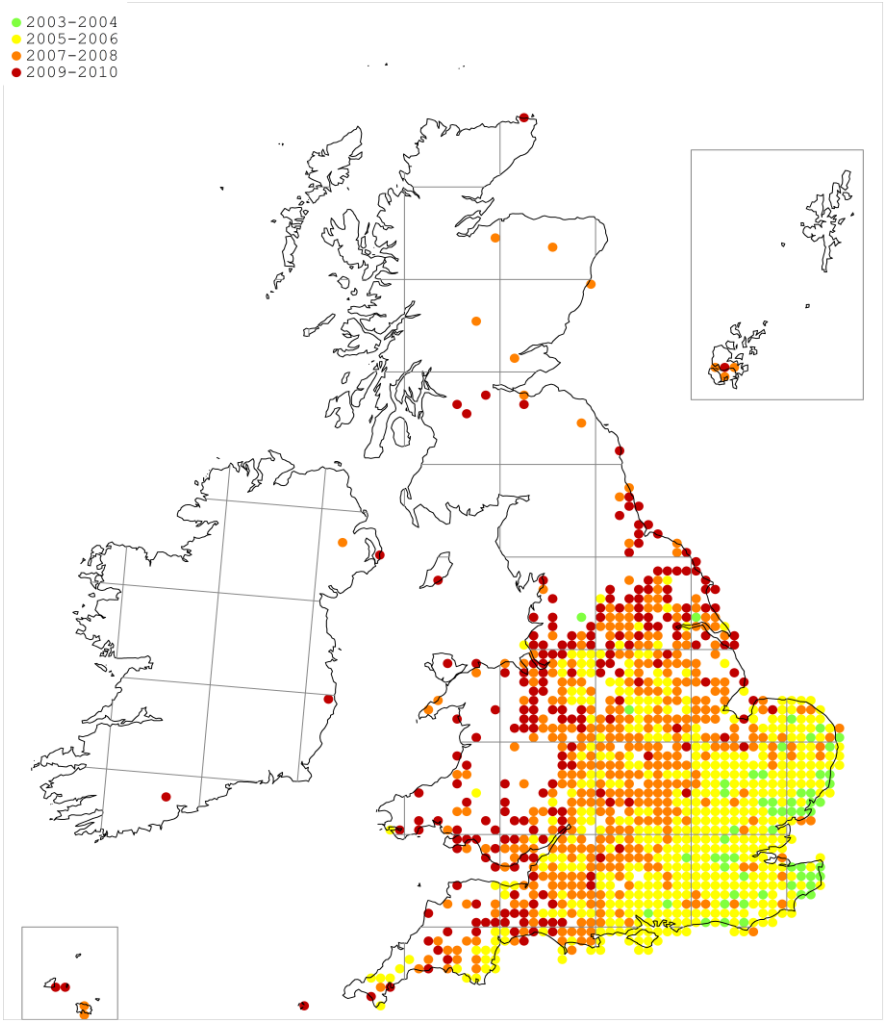


Figure 1.4: *H. axyridis* distributions in the UK (NERC/Field Studies Council, 2010)

1.3 Manual Identification vs. Automated Identification

Taxonomy is the science of classifying organisms which has been the foundation of all biological sciences (Mohamed, 2000; Simpson, 2010). It is a major part of systematics, and part of it is identification. Species identification involves either manual or automated identification. Manual identification involves use of physical observations on the insect's body; this includes capturing a few characteristics through detailed observations and physical measurements. Observations include the body segments, movement, shape, colour, shape of antennae, etc. Measurements include size, count of spots, etc. (Glickstein, 1987). The use of dichotomous keys has been a continuing practice among taxonomists due to the availability of references for identification. Once the keys are accessible, users are guided through step-by-step query and answers. As the ecology changes, major changes in identification techniques are required. For instance, the way taxonomists practice manually requires continuing need for rapid field identification and the need for identification of large numbers of organisms to provide ecological information (Boddy, Morris and Morgan, 1998). This has been affected by the general decline in the taxonomic workforce, which has been part of the taxonomic impediment to biodiversity studies (Mohamed, 2000; Hopkins and Freckleton, 2002). The issue is further aggravated by changes in the taxonomic community, where floristic and faunistic studies have become less attractive and most research funds and effort have been channelled towards phylogenetic reconstruction (Weeks *et al*., 1999). This taxonomic impediment will become serious unless solutions are explored to rectify it (Cotterill, 1995; Zakri, 2000; Macleod, 2007). As per 'Darwin Declaration' of 1998 in the context of the Convention on Biological Diversity (CBD), removal of the taxonomic impediment is an important step towards the conservation of biodiversity (Mohamed,

2000). It will require skilled workers and experts for urgent implementation. Other than human resources, the number of specimens will need to be large in order to perform routine species identifications (Gaston and O'Neill, 2004). As a countermeasure, the use of automated identification of ladybirds, may have enormous potential and has not been extensively explored.

1.3.1 Automated Species Identification (ASI)

Automated species identification is an application of general pattern recognition, and part of computer-aided taxonomy (CAT) (Chesmore, 1998; 2000). It is one of the provisions by Article 7 (Identification and Monitoring) and Article 12 (Research and Training) of the CBD (Zakri, 2000). Due to the importance of biodiversity at this level, ASI definitely is an interesting path in the application of pattern recognition. Pattern recognition itself has many applications including speech analysis, handwriting recognition, face recognition, human-computer interaction and condition monitoring of machines. There are two main levels of automation; the first is full automation meaning complete identification without user interaction, and the other is semi-automation**.** Semi-automation is more realistic than full automation as it allows prior sorting into higher taxonomic categories such as genera and more likely to be feasible in the short term (Chesmore, 2007). Knowledge-based systems have the ability to handle non-linear, fuzzy and incomplete data; therefore, they are more suitable as the core for any CAT system (Chesmore, 2007). In this thesis reference are made to the design of ANN systems, algorithms and methodologies which have been deployed in past literatures. This is elaborated in Chapter 2.

1.4 Research Aims

The specific aims of the research project are to:

a) develop algorithms to optimally distinguish Harlequin and other likely ladybird species using colour image processing and classification methods, including Artificial Neural Networks (ANN);
b) evaluate classification accuracy;
c) extend the classification methods to all 26 recognisable species.

1.5 Hypothesis

The hypothesis of the research is:

"It is possible to develop methods for automatically identifying ladybird species from colour images."

The hypothesis means that:

a) There will be a pre-sorting system to identify some available species of ladybirds in the UK, with provision for additional unknown species.
b) Techniques and algorithms will be developed based on colour images, to determine which class a ladybird should belong to. For instance, if an image of a ladybird has 2 black spots and red coloured background on its elytron, the species is without doubt a 2-spot ladybird. The same will apply to all ladybird images.

In practice, a prototype automated species identification system has been developed to distinguish UK ladybird species using techniques such as image processing, ANN and expert systems. This shows that the system has been able to perform pre-sorting of the questionable and/or incorrect species, and provides immediate feedback to the supplier of the image. Ultimately users will be able to freely access the system and

supply images according to specific requirements (viewing angles, size, resolution, etc.). The system will be able to process the images to provide automated response by pre-sorting the supplied images into a few categories: "definitely", "definitely not" and "pass on to the expert". As will be seen later in the coming chapters, the achievements are quantified through the use of confusion matrix, Receiver Operating Characteristic (ROC) curves, and some statistical measures. The block diagram of the proposed hybrid intelligent system is shown as in Figure 1.5, and the details will be explained in Chapter 6.

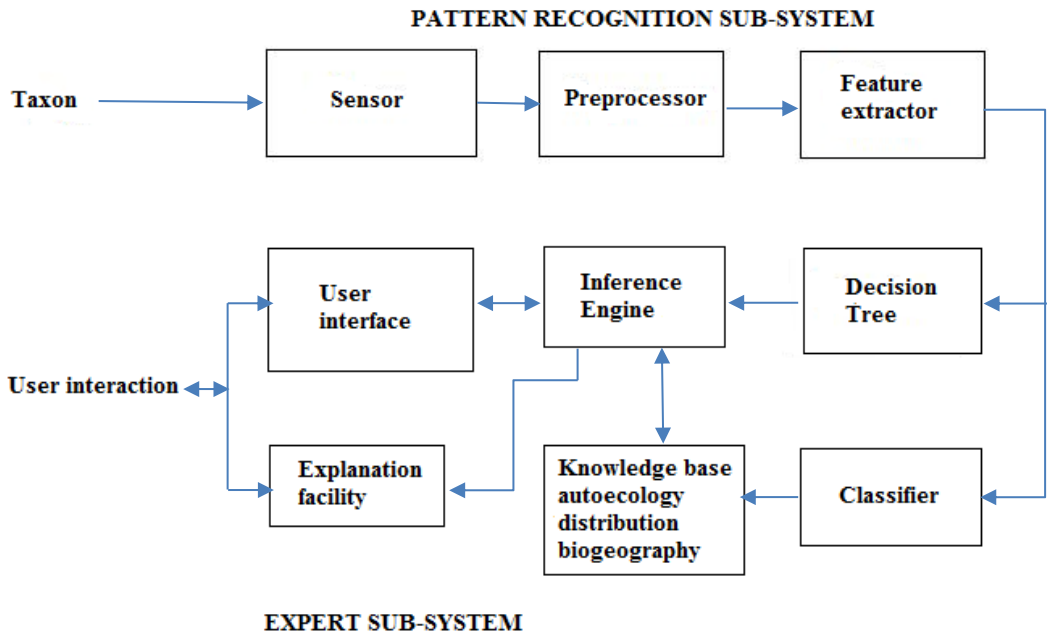


Figure 1.5: Block diagram of hybrid intelligent system, referred to as Automated Ladybird Identification using Expert and Neural Systems (ALIENS)

1.6 Issues and Challenges

In tackling the research hypothesis, the whole process of research did not shy away from technical challenges. Figure 1.6 shows various ladybird images, and the extent of the images quality supplied by members of the public to the Centre for Ecology & Hydrology (CEH) Wallingford, England.

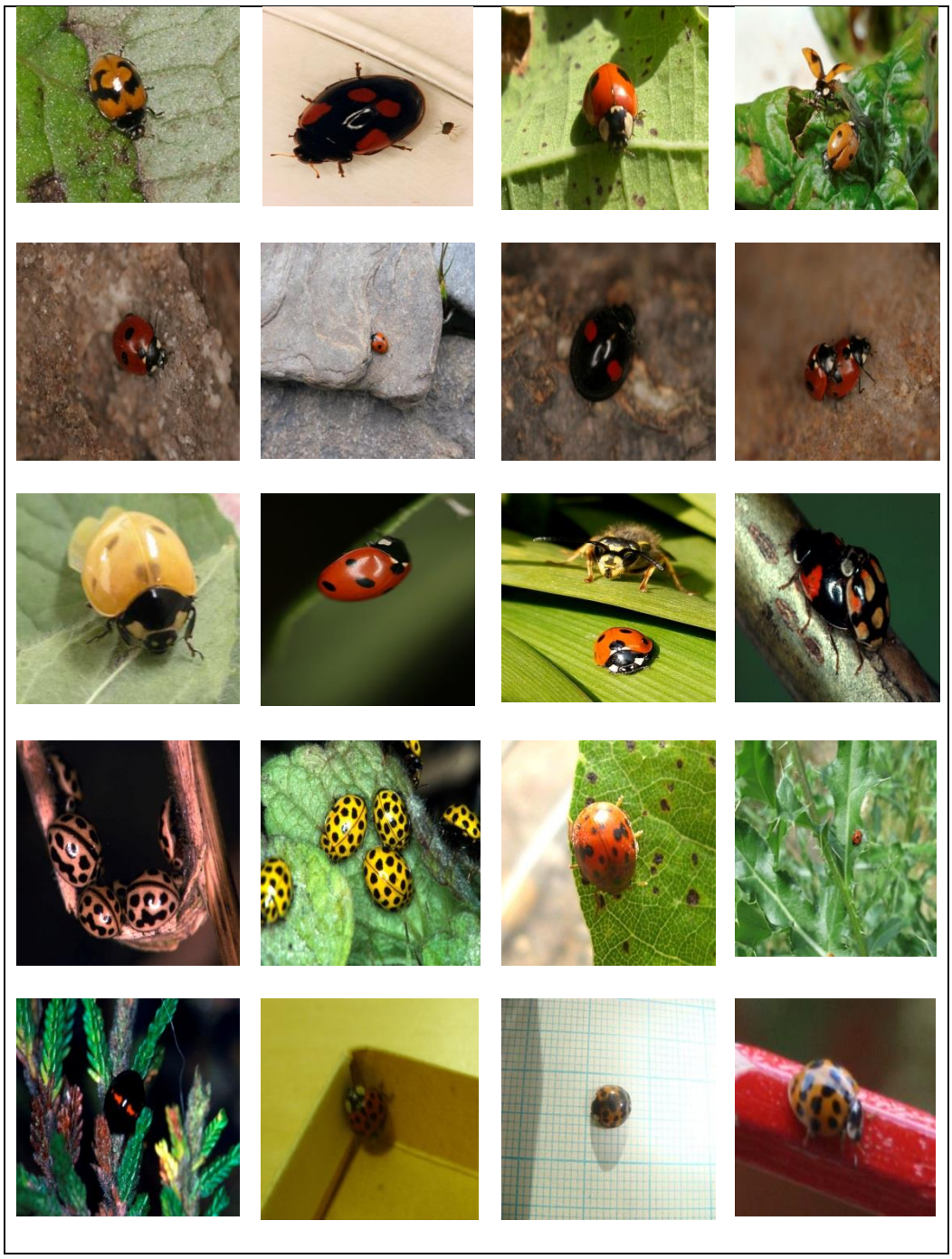


Figure 1.6: Sample images of ladybird species showing various qualities and pose

These photographs have been used extensively in the development of the ladybird identification system. Most of the photographs suffer from various issues such as poor illumination, complicated background, multiple objects, incomplete views etc. Furthermore, the ladybirds are three-dimensional insects and it is difficult to accurately capture the details of spots on both the elytra and the pronotum from one angle only. Another concern is size. Statistically, the size of the ladybirds in the UK ranges from 3 mm to 8 mm (UK Ladybird Survey, n.d.). With an automated system, a crude estimate of size is the best the image processing can produce, as the actual size and the depth of field are unknown parameters. It has been an interesting challenge for the system to tackle due to the many unknowns.

1.7 Cost saving benefits

In manual identification, a large number of unknown samples/images need to be manually examined and responded to by an expert. Morphological features such as measurement of body markings are still useful in comparison to DNA-based identifications. In fact, DNA-based identifications require higher costs associated with its availability, facility and expertise required (Will, Mishler and Wheeler, 2005; Will and Rubinoff, 2004; Chesmore, Bernard, Inman and Bowyer, 2003). In some cases, its use is not necessary such as in the identification of birds (Dunn, 2003). For the identification of Diptera which has wide overlap between intraspecific and interspecific genetic variability, the use of DNA-barcoding can be misleading (Meier et al., 2006).

In contrast, automated identification of ladybirds will reduce the number of images to be examined by pre-sorting the questionable and/or incorrect species, and provide immediate feedback to the supplier of the image. With the expert's opinion

embedded in the system, and proper system structure it will hopefully generate more interest from the general public to contribute ladybird images for automated identification, hence saving time and expenses in scientific data collection. Moreover, taxonomists' expertise are highly valuable to the identification loop and would not deteriorate or be devalued due to automation. Their input has been effectively used in this project, where users interact with the expert system through a graphical user interface (GUI). This identification task performed by digital tools will not replace the human expert; however, it helps in making faster progress in taxonomy (Page *et al.*, 2005). In short, an expert system approach is more superior to typical identification tools, like dichotomous keys in terms of efficiency and ease of use; tolerance of missing data, explanatory capability and provision for meaningful output when an unambiguous identification is not possible (Woolley and Stone, 1987).

1.8 Contribution towards field

This thesis has produced the following contributions towards knowledge:

(a) The whole thesis itself demonstrates a pioneering work in automated identification of ladybirds in UK. It contains working modules that can be implemented or re-engineered for future improvements.

(b) The use of CIELAB colour space in the image processing steps is novel.

(c) Application of decision trees to simplify the feature map, hence minimising number of features required for the next stages, is itself a novel technique.

(d) Experimental studies on the application of Multilayer Perceptron using back propagation algorithm, Learning Vector Quantisation (LVQ), Support Vector Machine (SVM) and Probabilistic Neural Network (PNN).

(e) Application of a fuzzy inference engine which links up with knowledge base, neural network and decision tree forming a hybrid intelligent system.

(f) The system is an improvement from existing automated identification systems, by having user inputs and explanation ability of the inference engine.

1.9 Thesis outline

The remainder of this thesis is outlined as follows: Chapter 2 explains the literature surveys containing work involving identification of insects using image processing. These will cover various general issues including image processing techniques, data representation, feature extraction and classification. Chapter 3 elaborates progress in image processing, the various theories applied to experiments, and the results obtained. Chapter 4 explains feature extraction and identification systems. Chapter 5 elaborates on each classifier in use, the algorithms and data partitioning. Chapter 6 provides detailed results from experiments and analytical discussions. Chapter 7 explains the overall integrated system between the image processing, neural network and the expert system. Chapter 8 concludes the thesis with suggestion on possible future improvements.

CHAPTER 2

AUTOMATED SPECIES IDENTIFICATION

CHAPTER 2

AUTOMATED SPECIES IDENTIFICATION

This chapter elaborates the concepts of automated species identification (ASI) and provides a review of automated identification systems which use image processing, highlighting the advantages and disadvantages of each system.

2.1 Computer Aided Taxonomy

The word 'taxonomy' means a system for naming and organizing plants and animals into groups which share similar qualities (Cambridge Dictionaries Online, 2009). According to the Oxford Online English Dictionary (2009), 'taxonomy' is a branch of science concerned with classification. In general, most identification methods can be divided into two groups. Comparison is normally done when specimen is compared with a museum collection or illustrations in a natural history guide book, and then will produce an estimate of the similarities between the unknown and a range of possible taxa. The taxon that best matches the similarities are selected (Pankhurst, 1998). In contrast, the elimination method performs diagnostic by asking questions regarding the states of one or more characters to be observed. The response acquired will effectively eliminate taxa that do not belong to the observation. The diagnostic

will ask more questions until all possible alternatives are eliminated, therefore will leave only one taxon left as the winner (Pankhurst, 1998).

Computer Aided Taxonomy (CAT) can be described as the use of computer technology to assist the taxonomy-related process. In the author's opinion, CAT was probably inspired by the use of interactive multiple-entry keys to enable biologist to identify a specimen with the aid of computers (Goodall, 1968). Multiple-entry keys differs from dichotomous keys as they permit the user to choose a subset of characters and the order they are used rather than pre-set characters by the maker of dichotomous keys. This leads to the development of more interactive keys in taxonomy. Interactive keys involve an interactive computer program, where a user will enter character-state values of a specimen and interact with the program. The program then eliminates taxa whose attributes do not match those of the specimen, and the process continues until only one taxon remains (Dallwitz *et al.*, 1998). The first international data standard for identification data was created by Dallwitz in 1980, which was named DELTA Format (DEscription language for TAxonomy). It was one of the first standards to be adopted by the Taxonomic Databases Working Group (TDWG). DELTA uses ASCII text, and encodes the descriptions of taxa, characters and states in a data matrix (Pankhurst, 1998). The DELTA project and INTKEY were two identification programs that were based on DELTA format (Dallwitz, 1980; Pankhurst, 1998; White & Sandlant, 1998). The next section elaborates some biological identification systems that have been deployed, and it will explain the various approaches to automate the identification process. It will also show some advantages, and disadvantages of each system.

2.2 Biological identification systems

As mentioned earlier, the principle of identification involves the process of comparing a representation of an individual specimen with taxa (Dallwitz *et al.*, 1998; Lebbe and Vignes, 1998). Automated species identification (ASI) involves the use of a computer to aid species identification. In this section some biological identification systems are reviewed. Unlike earlier computerised identification systems, these projects use images of the object at the input stage even though some other inputs can be used such as acoustic, radar, flow cytometry and sonar (Chesmore, 2007).

Some of the earliest applications of ASI using image processing involved the identification of marine zooplankton and bacteria (Katsinis *et. al.*, 1984; Blackburn *et al.*, 1998; Dorge et al., 2000; Walker and Kumagai, 2000; Foreroa *et al.*, 2004). There have been applications developed to perform photographic identification of mammals such as cheetah, zebras, giraffes, lions, chimpanzees, wildogs (Kelly, 2001), sea lion (McConkey, 1999) and sea otters (Gilkinson *et al.*, 2007). Some researchers reported work on image-based identification of insects, such as Lepidoptera (Chesmore and Monkman, 1994; Watson, O'Neill and Kitching, 2003; White and Winokur, 2003; Kipling and Chesmore, 2005). Image-based identification has also been extended on Hymenoptera, for example on braconid wasps (Weeks, O'Neill, Gaston and Gauld, 1997a, 1997b; Gauld, O'Neill and Gaston, 2000), honeybees (Daly, Hoelmer, Norman and Allen, 1982; Schroder, Drescher, Steinhage and Kastenholz, 1995; Steinhage, Kastenholz, Schroder and Drescher, 1997; Steinhage, Schroder, Lampe and Cremers, 2007), solitary bees (Roth, Pogoda, Steinhage and Schroder, 1999), ichnumonid wasps (Yu *et al.*, 1992), parasitic wasps (Angel, 1999) and leafhoppers (Dietrich and Pooley, 1994). There are also reported

works on identification of Arachnids via genitalia images (Do, Harp and Norris, 1999; Russell, Do, Huff and Platnick, 2007). Dai and Chesmore reported works on the location and description of wing venation in Diptera (Dai and Chesmore, 2005).

Since their introduction for ASI, ANN has been widely used to discriminate a few taxa, for instance, six basidiomycetes based on flow cytometric measurements of spores, five *Penicillium* species based on cultural characteristics and also 17 species of *Pestalotiopsis* and the closely related *Monochaetia karstenii* and *Truncatella truncate* from spore morphometric data. Boddy, Morris & Morgan (1998) reported that expert systems could be a further possible method for systematists. However, it was a very difficult and time consuming to extract the relevant rules and to encode the rules in a formal manner (Boddy, Morris and Morgan, 1998).

The following are specific examples of image-based ASI systems. They are elaborated to show the main similarities and differences with the current project in terms of approach and techniques used.

2.2.1 Automatic Bee Identification System (ABIS)

ABIS is designed to identify species of bees from images of their forewings. ABIS uses wing images from both existing collections and in the field (Steinhage, 1997). The structure of the wing venation is generally fixed and well-suited to the identification of bee species. ABIS uses diffuse background illumination to get the image structure of the wing venation. By analysing each wing image then a well-defined, characteristic morphometric feature can be formed (veins, vein junctions and cells) and used in a knowledge-based classification approach. Steps in automatic extraction of morphologic wing features are:

1) Detect edges in the pattern of wing venation, and formulate hypotheses regarding the location of key cells with the aid of genus wing template.

2) Once the cells are detected, ABIS generates numerical morphometric features which describe the cells and the topological relationships.

3) ABIS inputs the morphometric features to linear discriminant analysis (LDA) and non-linear kernel discriminant analysis (NKDA).

Using this technique generates too many features over a normalised wing image, even after down-sampling using second-order Gaussian filtering. The resulting intensity matrix is 120x20, forming 240 elements from down sampling called iconic features, and 50 morphometric features. In total ABIS produces 290 feature vectors. Due to this high dimensionality and small training sets, ABIS applies NKDA instead of LDA. When compared with a support vector machine (SVM), NKDA classifier is better for this situation because it allows data visualization, and faster due to simpler optimization and the ability to handle multiclass problems directly (Steinhage et al., 2007). The performance of the ABIS system was tested for identifying *Bombus lucorum*, *Bombus terrestris*, *Bombus cryptarum* and *Bombus magnus*. The system was initially trained with wing images of 70 individuals per species. It achieved more than 95% identification rate by combining both morphometric and iconic features. Similar results were obtained for German *Colletes*, *Andrena*, and American *Osmia* species (Steinhage, 2007).

2.2.2 SPIDA

SPIDA stands for 'Species Identified Automatically' (Russell *et. al.*, 2007). There are two versions of SPIDA; one is a stand-alone and the other is an Internet-based version called SPIDA-web. Both were designed with the aim to produce a tool to

make routine identifications of any group of organisms by non-experts accurate and efficient. An earlier version of SPIDA has proven successful in the identification of 5 species of Ichneumonid wasp, 6 species of Lycosid spiders, 12 species of North American bees and 121 species of the Australasian ground spiders of the family Trochanteriidae (Russell *et. al.*, 2007). They highlighted a few typical problems inherent to an automated ID system utilising specimen images. In the species identification of spiders of the family Trochanteriidae, the primary character is based only on the shape of genitalia which are visible without dissection. Species can be highly similar, possibly limited data per species and high intraspecific variation between individuals. From the spider images, features are encoded using wavelet transform where both Daubechies 4 and Gabor wavelet function were used. MLP was used and then fed with coefficients from both wavelet encoders. There was one ANN per species in the group, and each ANN has two output nodes representing positive and negative outputs (Russell *et. al.*, 2007). Consequently, the training set consists of 'pro' or positive training set and 'anti' or negative set. By having this structure of individually trained species-level ANNs and training sets, the network should be able to generalize or classify unseen images (novel species). For the novel species test, they randomly selected 20 images from species in related families as test images. Cascade correlation was used together with quick propagation to train the ANNs.

For SPIDA-web, the trained ANNs are stored in a server and deal with queries from users through an Internet interface. Its website interacts with users and displays dynamic data in response to the user's input. It uses JAVA servlets supported by Tomcat, an open source server. To use the system for query, a user would have to log in. The image for submission will be sent to the servlet which uses Java

Advanced Imaging library for proper scaling and formatting. The wavelet transformation is applied and the coefficients forwarded to the trained ANNs. The results of identification are then displayed via a Java server page along with other information such as distribution maps, line drawings of genitalia, whole-body images, technical descriptions, etc. Even though this process should take a few seconds to complete on a stand-alone machine, the overall speed of SPIDA-web actually depends on server speed and Internet connection speed. The results have been quite impressive, with 95% of unknowns correctly classified. It has also been tested on reprocessed images, such as effect of cropping and rotation. The system is robust for all tests, except for 4-degree rotation where the identification accuracy dropped to 40%. In term of performance measures, the developers suggested SPIDA-web is accurate, accessible, reasonably scalable and flexible. They suggested for a general automated ID system must have the capacity to expand or retrained without needing much computing time. They also suggested that a sensible auto ID system be hierarchical to some degree, and not generic to all groups of organisms to avoid accidental similarities between processed images of very different structures (Russell *et. al.*, 2007).

2.2.3 DAISY

The DAISY system was developed by O'Neill et al. with Biotechnology and Biological Sciences research Council (BBSRC) and UK Government Darwin Initiative funding (Weeks, 1997b). The objectives of DAISY were to overcome the taxonomic impediment, and to provide a system which would allow non-specialists to identify organisms within arthropod genera using a combination of both morphology and molecular data (O'Neill, 2007). A few tests on the completed

DAISY system have been performed, inclusive of British Bumblebees (*Bombus*, *Megabombus*) (Pajak, 2000), Costa Rican Hawkmoths (*Xylophanes sp.*), Costa Rican parasitic wasps (*Enicospilus sp.*), Palaearctic biting midges (*Ceratopogonids*) (Gauld, O'Neill and Gaston, 2000) and British Lepidoptera (Moths) (Watson, O'Neill and Kitching, 2003). On average, the system is able to identify taxa between 75% to 85% correct identification rate. This has been achieved even when the difference between taxa is not substantive, or the pose of specimens was arbitrary (O'Neill, 2007). The DAISY system used generic pattern matching technology based on plastic self-organising map (PSOM) (Lang and Warwick, 2001, cited by O'Neill, 2007). It is fast, not restricted to a single organismal group and pattern class, and may be trained in real-time, scalable and easy-to-use GUI. A sample GUI of DAISY system is shown in Figure 2.1.

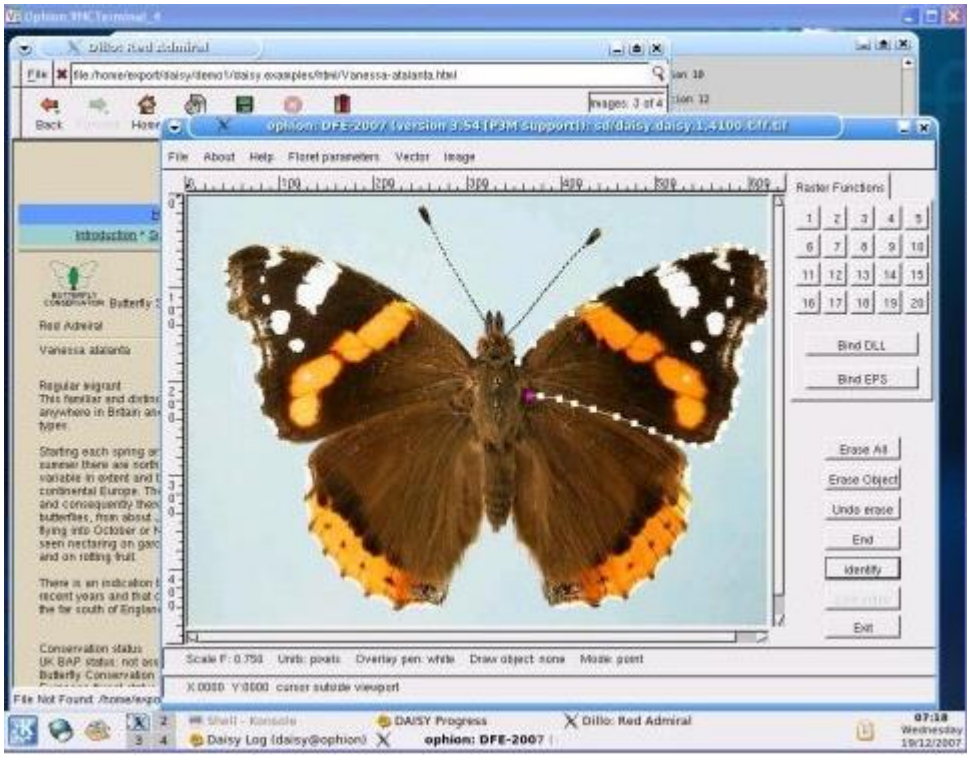


Figure 2.1: DAISY GUI in operation (O'Neill, 2007)

Despite its general ability to identify various insects, DAISY is generic as it tries to cover many taxa. DAISY also requires user to perform manual extraction of the region of interest (ROI) in the image, shown as white dots in Figure 2.1. Furthermore, DAISY operates well with 2D-images and has not had any success in dealing with three-dimensional (3D) images. It down samples images to a size of 32x32 pixels, which would reduce measurability when morphological features are extracted.

2.2.4 Moth ID

The work by Mayo and Watson on automatic species identification of live moths is considered next. They used image analysis on 774 live moth images and data mining techniques using WEKA machine learning toolkit to classify 35 UK species (Mayo and Watson, 2006). A support vector machine (SVM) was used and 85% accuracy was obtained using jackknife test. Other techniques tested were Random Forests, Instance-Based Learning 1 (IB1), Instance-Based Learning 5 (IB5), Naive Bayes and J48 with accuracies of 83.2%, 71.6%, 65.36%, 65.9% and 58.3% respectively. The results were obtained without manual specification of a region of interest on the images. Feature extraction was performed both globally and locally, where a total of 11,300 features were obtained. From that amount, 9,600 local features were obtained by placing a grid of 600x600 pixels over the centroid of the image. They have 400 square patches, and measured the mean, minimum, maximum and standard deviation of pixel values over the 30x30 pixels patches. This has been done on both RGB (Red, Green and Blue) and HSB (Hue, Saturation and Brightness) colour spaces. Colour features were obtained through measurements of global statistics on the image, both in binary and colour versions. They made separate measurements on

each RGB colour planes of the image. The same process was repeated on a HSB version of the image. Some global features were obtained from binary version of the image, which include the count of foreground pixels, the number of background pixels, the ratio between them, the interior density, the standard deviation of pixel positions, skew and kurtosis. These form the total global features obtained from binary, RGB and HSB versions of the image. The global features were reported to be useful as they are invariant to rotation (*Ibid*). To improve accuracies, they suggested the use of local features over the centroid and to make the features invariant to the size of a moth. The use of other colour spaces was also suggested to boost accuracy.

2.2.5 Identification of quarantine fungal pests

A research project on using image analysis to identify quarantine fungal pest *Tilletia indica* (Karnal bunt) was reported by Chesmore et al. (Chesmore, Bernard, Inman and Bowyer, 2003). The work aimed to identify *Tilletia indica*, a floret-infecting smut fungus causing Karnal bunt of wheat. It can be confused with other species such as *Tilletia walkeri* (ryegrass bunt) and *Tilletia horrida* (rice smut). The image analysis system used bleached spores of *T. indica*, *T. walkerii* and *T. horrida*. Bleaching showed additional characters for identification. Due to significant overlap of characteristics, identification is difficult to achieve for small samples. Molecular methods were commonly used, but took a much longer time to diagnose (in the order of weeks). For rapid identification, they proposed image analysis and the use of Principal Component Analysis (PCA) for character discrimination. It was designed to automatically locate and measure all spores in a given image. Only the greyscale version of the image was used. Spores were located by scanning the image from top left to bottom right. A spore is discriminated from the background and debris

through thresholding. After labelling, nine parameters were measured which include perimeter, average spine size, number of spines, aspect ratio etc. PCA showed good discrimination with little overlap between characters. It was concluded that the ratios of internal to external spore diameters were able to discriminate between *T. indica* and *T. walkeri* giving 97% identification accuracy.

2.2.6 CAT using Structural Image Processing and ANN

The work involves the use of computer aided taxonomy (CAT) to perform semi-automated identification of hoverflies and bumblebees using wing images. High quality venation diagrams were extracted through both traditional processes and novel methods. Traditional processes include grey scale transformation and edge detection, while novel methods involve filling, noise filtering, smoothing and cutting. Manual processing of vein images are performed through software interface to obtain accurate venation in close resemblance to the original vein image. The work produced a novel analysis technique based on the venation and relationships between veins using tree diagram. The work has also explored taxon identification by extracting characteristic features such as cell composition and vein fitting coefficients, and these features were then fed to multilayer perceptron (MLP) and learning vector quantisation (LVQ) neural networks. Using the tree-based identification 100% accuracy was reported for taxa to tribe level for nine hoverfly species and three bumblebee species, including two sub-species. Using the LVQ neural networks, recognition rates of 90.95% for hoverflies and 95.6% for bumblebees were achieved (Jing Dai, 2006). However, for MLP the average results were only 60% for hoverflies and 30% for bumblebees. It has been shown that the

semi-automated technique using structural image processing was user friendly, and can be applied to any insect groups with transparent wings.

2.2.7 Plant Identification Systems

Plant identification systems involve the identification of plants using their leaves. One of the pioneering work by JY Clark of the University of Surrey aims for the identification of mature specimens taken from the crown of the tree. The system uses characters obtained from cultivated species of the genus *Tilia*, commonly known as lime trees (Clark, 2004; 2007). Each specimen provides 22 morphological characters and 57 training records were used. Data is presented in ASCII tabulated numeric format to a multilayer perceptron ANN. Input vectors were normalised to the range -0.9 to 0.9. This was done to ensure the training time is reduced. Clark used the squared error percentage when evaluating training, testing and validation test sets. Three different partition pairs of training and validation sets were produced, however, only one record of each species was randomly chosen to make up a validation set. An optimized number of hidden nodes were obtained first after obtaining the lowest error on the validation data set. Next, the learning rate was obtained when the hidden nodes were fixed. Clark showed that a systematic methodology in applying character and measurement data into MLP results in effectively tuned system parameters, which could be useful for non-experts to use for plant identification. Results for species identifications were shown in term of confusion matrix. The identification performance of the MLP was improved by 16% after the inclusion of minimal geographic information, represented in term of 3 geographic characters in binary code.

The MORPHIDAS project (Morphometric Herbarium Image Data Analysis) is another project on automatic extraction of leaf information from digital images of whole herbarium specimens (MORPHIDAS, online). The by-product is a database of 1900 images of *Tilia* specimens from Kew and an automation system for specific research tasks. The tasks include locating individual leaves within images, locating margins and veins, extracting morphological features, measuring patterns of variation of morphological features within and between groups and virtual restoration of damaged leaves. MORPHIDAS use advanced image processing developed in MATLAB 7.10.0 (R2010a) to capture and measure morphological details from the leaves such as the length, width, veins and teeth. The system is able to determine the marginal teeth given the image of a single leaf by tracing the outline and locates local maxima and minima to find the teeth. Other features include the area of each tooth and the tip angle.

A recent publication by Clark et al. explained the use of multilayer perceptron (MLP) used as a tool for automatic plant identification (Clark, Corney and Tang; 2012). They used morphological characters obtained from images of 4 species of the genus *Tilia* in the Herbarium of the Royal Botanic Gardens, Kew, UK. A simple feed forward MLP with one input layer, one hidden layer and one output layer was used. There were 22 input nodes corresponding to the number of characters, and 4 output nodes representing the four species of *Tilia*. Input vectors were normalised to (-0.9, +0.9) and performed independently for each character over all training periods. It was claimed to reduce the training time and help prevent initial weighting of characters. Data were divided into three partitions called training, validation and test sets in a ratio of 70:20:10. The validation data set was used to reduce over fitting. Three different sets called A, B and C were created where stratified cross validation

was applied to enable each species be represented the same number of times in the training, validation and test set. A constant learning rate of 0.1 and single fixed random seed was initially used for training. After a number of trials, the optimum number of hidden nodes was determined from the configuration that gave the lowest error on the validation set. Once the number of hidden nodes is obtained and fixed, similar trial runs were performed to find the optimised value of the network's learning rate. Species identification is represented by a misidentification matrix. It also shows the percentage confidence of correct identification (*%Conf*). Even though the results from this work showed lower identification rates of 44% compared to earlier similar studies, it showed some significant achievement in innovation because some level of automation helped extracted information from images, as compared to earlier studies which used manual character extraction.

2.2.8 Leaf Recognition using Probabilistic Neural Network

As demonstrated in other researchers' work, in general, artificial neural networks have performed well as classifiers. The next work by Stephen Gang Wu *et al.* used image processing and Probabilistic Neural Network (PNN) to build general purpose automated leaf recognition for plant classification (Wu *et al.*, 2007). They were able to derive 12 leaf features from 5 basic geometric features, which were extracted after the implementation of image processing techniques on selected plant leaves. Leaf images, which were in 800x600 resolution, were converted from RGB to grey scale. The threshold level was selected according to RGB histogram of 3000 leaves. The image was subjected to smoothing using 3x3 averaging filter. Shape of the leaves was obtained after applying Laplacian filtering. The system was able to automatically extract 11 features automatically out of the 12 digital morphological

features. From the 12 features, only 5 orthogonalised features were adopted through the use of Principal Component Analysis (PCA). An interesting remark pointed out by Wu et al. was that since ANN can be treated as a "magical" black box; there is no need for a specified algorithm on how to identify different plants. PNN is an ANN using radial basis functions at the Radial Basis Layer. This layer evaluates vector distances between input vector and row weight vector in weight matrix. The RBF scales the distances nonlinearly. The next layer called Competitive Layer finds the shortest distance, hence finds the training pattern which is closest to the input pattern based on distance. PNN has many advantages: speed, robust to noise, easy to train and pose simple structure. Their PNN was trained using 1800 leaves for classifying 32 kinds of plants. The average accuracy was reported to be 90.312% (Wu *et al.*, 2007).

2.2.9 VeSTIS

VesTIS stands for 'Versatile Semi-Automatic Taxon Identification System from Digital Images' (Nikolaou *et al.*, 2010; Hart and Huang, 2011). It is an identification system built on Open Source platform, developed by Nikolaou et al. to classify 5 different species of marine annelid worms of class Polychaeta (*Nematonereis unicornis*, *Marphysa bellii*, *Polyophthalmus pictus*, *Armandia polyophthalma* and *Terebellides stroemi*) (Nikolaou *et al.*, 2010). The idea was to provide public access to such system, and broad enough in term of identification of taxonomic groups, unlike existing identification systems such as SPIDAweb, ABIS and DAISY. It uses digital image analysis, image enhancement and pattern recognition algorithms in an Open Source platform, so that it is freely accessible, extensible and not tied to any commercial software. The system applies Otsu binarisation method for image

segmentation. Fourier descriptors were used to mathematically describe the object's contour and parameterised the shape. These are then fed into feed forward ANN for identification. Multiple users can also work simultaneously due to the use of SQL-based database and client-server schema.

2.3 Summary

Table 2.1 provides a summary of past ASI projects and their contributions.

Table 2.1: ASI projects

System	Taxonomic Groups	Input Type	Classifier	Results/ Accuracy	References
ABIS	Bombus (*B. lucorum*, *B. terrestris*, *B. cryptarum*, *B. magnus*)	2D images of wing venation	NKDA	> 95%	Steinhage *et al.*, 2007
SPIDA	Ichneumonid wasps, Lycosid spiders, North American bees, Australian ground spiders of family Trochanteriidae	Images of genitalia	Multilayer perceptron using Back propagation algorithm	95%	Russell *et al.*, 2007
DAISY	British Bumblebees (Bombus, Megabombus), Costa Rican Hawkmoths & parasitic wasps, Palaearctic biting midges (Ceratopogonids), British Lepidoptera (Moths)	2D colour images	PSOM	Varies according to species: 94% for Costa Rican ichneumonoids, 85%-98% for Ceratopogonidae, 35%-100% for macrolepidoptera	O'Neill *et al.*, 2007
Moth ID	Moths	2D colour images	SVM J48 Random Forests Naïve Bayes IB1 IB2	85% 58.3% 83.2% 65.9% 71.6% 65.36%	Mayo and Watson, 2007
Quarantine fungal pests	Tilletia (*T. indica*, *T. walkeri*, *T. horrida*)	2D grey scale images	PCA	97%	Chesmore *et al.*, 2003
CAT using Structural image processing	Hoverflies, bumblebees	2D images of wing venation	ANN (MLP, LVQ)	MLP: 60% for hoverflies, 30% for bumblebees LVQ: > 90%	Jing Dai, 2006
VeSTIS	Marine annelid worms (Polychaeta)	Colour images	Feedforward ANN	70%	Nikolaou *et al.*,, 2010 Hart, 2011
Plant ID system incl. MORPHIDAS	*Tilia*	2D images of *Tilia* leaves	Multilayer perceptron	44%	Clark, 2004 Clark, 2007 Clark *et al.*, 2012
Plant Recognition System	32 species of Chinese plants	2D images	PNN	90.3%	Wu *et al.*, 2007

Some applicable concepts and techniques obtainable from past projects are vital to the success of the proposed automated ladybird identification system. The followings are some key areas to be considered for implementation:

- **Specific:** the system need not be generic (holistic) and may not need to identify all ladybird species, unlike DAISY, SPIDA-web.
- **Morphometric:** morphometric features have been useful for identification to some degree (ABIS, Moth ID, and CAT using Structural Image Processing).
- **Reasoning:** the need for a hybrid identification system of human-like learning abilities & explanation capabilities, utilising biogeographic information and ecological factors. A blend of expert system and ANN capabilities is required.
- **Online:** online implementation on mobile device will help taxonomists and the public.

CHAPTER 3

DIGITAL IMAGE PROCESSING

CHAPTER 3

DIGITAL IMAGE PROCESSING

There are two major development works in the investigation, which are image processing and intelligent systems. This chapter elaborates concepts and experimental work that has been carried out in image processing. The research into image processing involves two major processes, which are greyscale operations and colour image processing. These operations are implemented in MATLAB. Some strategies on image processing are discussed next.

3.1 Image Processing Strategy

From visual observations, ladybirds possess large variation in body colour. Some are quite obvious, for instance, there is a species commonly called ‘orange ladybird’ which has orange-coloured elytra and sixteen white spots. There are also ‘striped ladybirds’ with brownish elytra and cream-white stripes. Based on colour variation, the author has initiated investigating the use of colour as the leading feature for identifying ladybird species. This has become the hypothesis for the research. In doing so the focus is on two areas of the ladybird body: the elytra and the pronotum. Apart from its colour, a ladybird species may also be identified through various

physical characters, such as size, shape, number of spots, etc. (Southampton Natural History Society, 2005). Ideally, the system will need to evaluate whether the object is a ladybird, or not a ladybird. This can be first evaluated through detecting typical signs of colour composition of the image, and removing background clutter. The next process involves checking whether there are coloured spots on the object. If it does contain coloured spots (including white), it is confirmed a ladybird. This kind of pre- assurance is difficult to perform and this project has made an assumption that all colour images supplied by users of the system contain ladybirds. The system will also need to be able to deal with rotation and scaling of the colour images. Using this as a basic guideline, other properties are then evaluated through user interaction with the expert system which will be discussed in chapter 7.

3.2 Image Preparation

Before the images can be used for automated processing, they need to undergo a few pre-processing steps.

3.2.1 Image Capture & Specification

There are two sets of images; there is a set which contains ladybird images from UK including Harlequin ladybirds, and a second set comes from laboratory image capture. The first set has been provided by researchers from the Centre for Ecology & Hydrology (CEH), Wallingford. Most of these originated from photographs taken by members of the public. Some of the images suffer from various issues such as improper illumination, too low a resolution, complicated background, multiple objects, incomplete views, etc. This non-standard level of the image quality will be an interesting challenge for the system to tackle.

3.2.1.1 Hardware

The hardware for capturing images of ladybird has been setup to prepare samples, in light of the various constraints in image qualities. A 12 Megapixels SONY Cybershot DSC-HX1 digital camera of has been used to capture image of ladybirds in a laboratory setup. Other than this, initially a VEHO USB microscope has also been used with an intention to capture more details through its microscopic zooming abilities. Unfortunately, at certain magnification level some images can become noisy and grainy, subsequently losing details.

3.2.1.2 Software

From the input stage, 'MicroCapture' was used as an interface to the USB microscope. The images were saved in the hard disk for later use. For image manipulation, 'GIMP2' and 'MATLAB' was used. GIMP2 is a free GNU-based image manipulation program specifically employed to edit ladybird images. MATLAB 7 has been extensively used for the feature extraction and intelligent system stages. It provides easy access to a range of both elementary and advanced algorithms for numeric computing. The algorithms include operations for linear algebra, matrix manipulation, basic statistics, linear data fitting, and data reduction. MATLAB Toolboxes are add-ons that extend MATLAB with specialized functions and easy-to-use graphical user interfaces. The Image Processing Toolbox contains powerful built-in functions that allows user to perform data manipulation in image processing. This allows faster development of algorithms without the need for run-time compilation.

3.2.1.3 Sample Preparation

The second set contains images taken from ladybirds in a controlled setting. Ladybirds were captured around York, and multiple-view photos of each were taken with the background colour set to be light blue. The digital camera was set to macro mode for close-up capture. While the intention was to provide the 'ground truth' for comparison purpose, unfortunately some species are scarce such as five-spot ladybird and water ladybird (Southampton Natural History Society, 2005). Here only a small number of collections were gathered; these include orange ladybirds, 2-spot ladybirds, 10-spot ladybirds and Harlequins. This is due to difficulties in obtaining samples of ladybirds for photography. It was therefore decided to use the first set from CEH as the main source. The multiple-view ladybird images obtained through digital camera are shown in Figure 3.1, and the multiple-view images of ladybirds obtained through VEHO USB microscope are shown in Figure 3.2.

(a)

(b)

(c)

Figure 3.1: Example of 3D view photo from digital camera (a) top view, (b) left side view, and (c) right side view

(a)

(b)

(c)

Figure 3.2: Example of images from VEHO USB microscope (a) top view, (b) left side view, and (c) right side view

3.3 Process Workflow

The image processing steps follow the process flow as shown in Figure 3.3. It shows two major processes that an input image must go through to finally produce colour and morphological features.

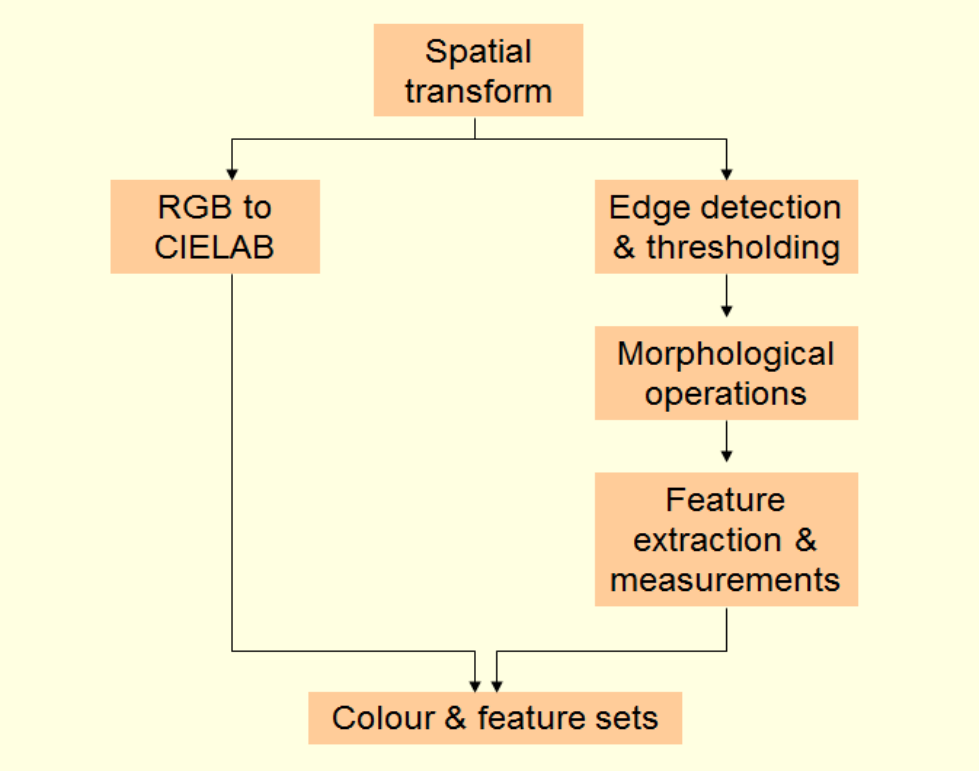


Figure 3.3: Workflow of image processing showing a modular approach

The idea is to get both geometrical measurements and colour measurements through a parallel and modular approach. Troubleshooting is easier this way.

3.4 Colour Image Processing

Colour is vital information to use for ladybird identification, mainly because it represents the natural characteristic of the ladybird. In fact, colour has been used in dichotomous keys for ladybird identification (Paul Mabbott, 2011; Southampton Natural History Society, 2005). The following subsections explain colour spaces and

how the use of CIELAB colour space has been useful for ladybird identification. Initial works on RGB and CIELAB are shown, where much effort has been put to determine which colour space to use as colour input to the identification system.

3.4.1 Colour spaces: RGB and CIELAB

It is important that colours in images are able to be represented visually, and colour space is the way to present them. The most frequently used is the RGB colour space. Note that RGB image is only used as input image at the early stage of the greyscale transformation operations. CIELAB, on the other hand, has been used to represent pixel colour values obtained from the elytra and spots. A series of experiments have been performed to finalise which is more suitable. This is based on the following criteria:

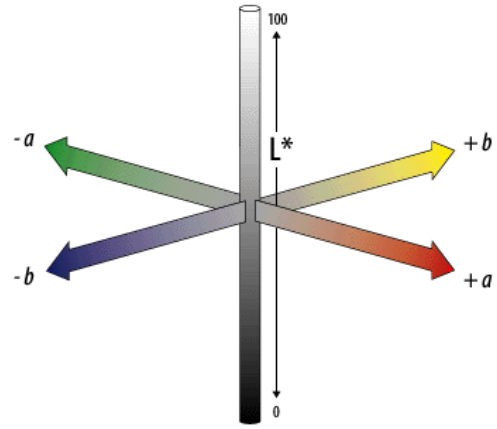
1. The ability of each to perform image processing process such as segmentation, in the most efficient and easiest way with minimum effects on output
2. The ability to perform when background clutter is present

Tests have been performed on both standard test images and ladybird images. The results on standard test images are given in Appendix II. CIELAB colour space is explained first, followed by test results on ladybird images.

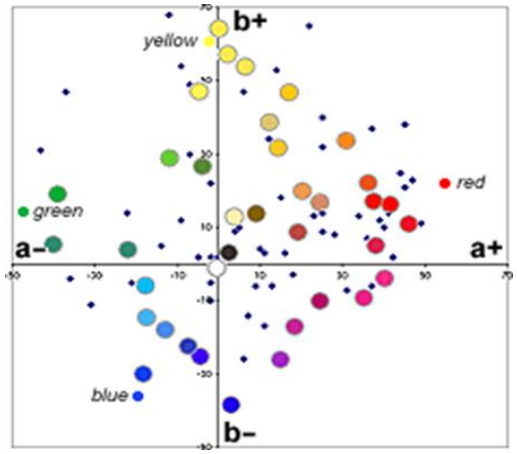
3.4.2 CIELAB Colour Plane

CIELAB is an approximate uniform colour scale to represent visual difference in the form of colour plane, and able to represent chroma separately from lightness (CIELAB colour models-Technical guides, n.d; Colour models, n.d). The CIELAB colour space separates lightness (L*) from the chroma components, a^* and b^*. It is

also able to represent the chroma values in the form of colour plane, as shown in Figure 3.4. This permits the visualisation of the clusters of colour. Unlike RGB, CIELAB is also device independent. The maximum values for *a** and *b** are +120, while minimum values are -120. The range for L-axis is 0-100.



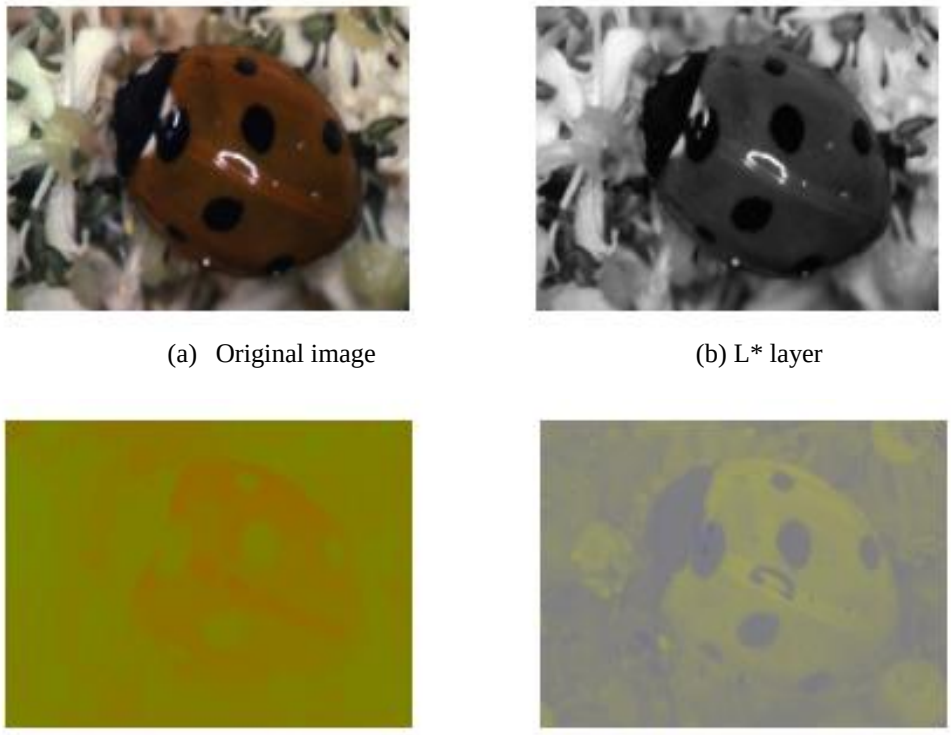
(a)



(b)

Figure 3.4: CIELAB colour plane
(a) 3-axes view, and (b) viewed from L* axis
(CIELAB colour models-Technical guides, n.d.; Colour models, n.d.)

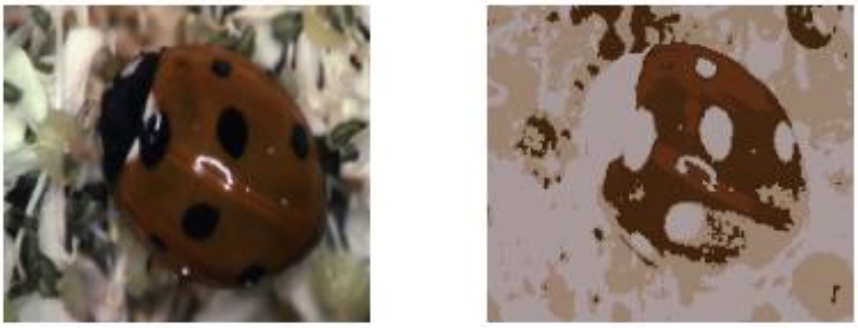
In this work CIELAB values have been used to represent pixel colours of both the spots and the elytra of ladybirds. Due to the majority of ladybird images were photographed in their natural habitat, CIELAB was selected as the colour space in order to counter-act illumination problems. These outdoor-type images are prone to illumination issues, which is a variable that is quite difficult to control. By separating lightness and chroma using CIELAB as the colour plane, subsequent work has been made simpler to model logically (Torres, Reutter and Lorente, 1999; Yip and Sinha, 2001). This is because one colour is distinct from another colour and the difference between chroma values can be calculated (CIELAB colour models – Technical Guides, n.d.; Vízhányó and Felföldí, 2000). Figures 3.5-3.7 show evidence of tests on a CIELAB version of a scarce 7-spot ladybird image.



(a) Original image (b) L* layer

(c) a* layer (d) b* layer

Figure 3.5: Image of scarce 7-spot ladybird (with background) after conversion to CIELAB from RGB

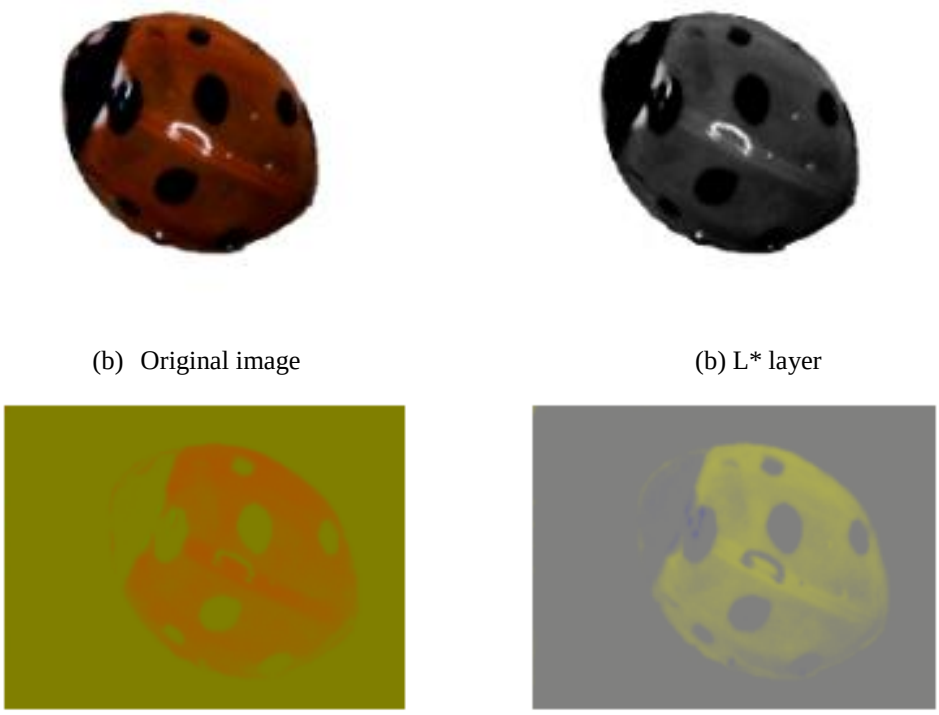


(a) Original image (b) Segmented image

Figure 3.6: Image of scarce 7-spot (with background) after colour segmentation showing background clutter

Figure 3.7: Magnified view of the binary version of scarce 7-spot ladybird showing complicated background and unintelligible image

Figures 3.8 - 3.9 show the resultant image of scarce 7-spot without background (after manual cut-off using GIMP2).



(b) Original image (b) L* layer

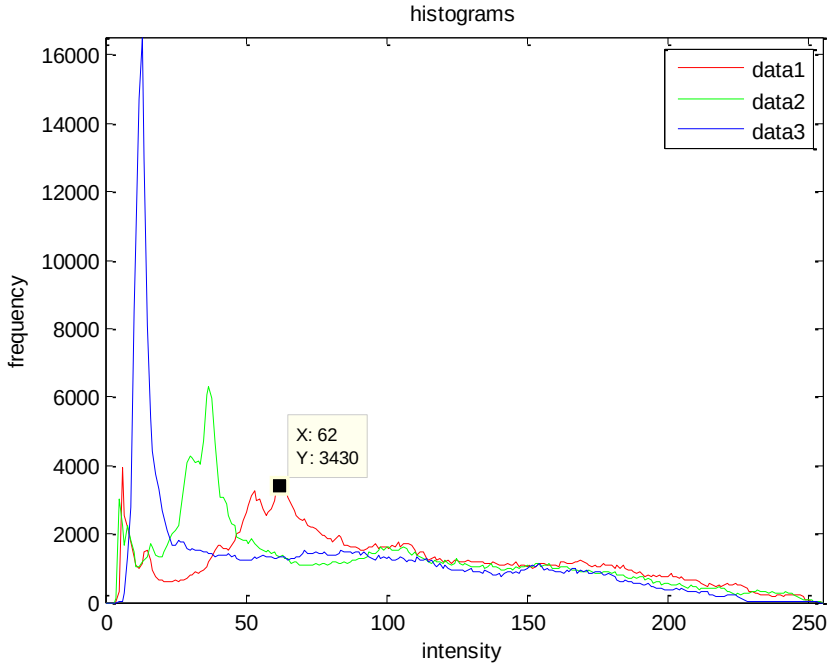
(c) a* layer (d) b* layer

Figure 3.8: Image of scarce 7-spot ladybird (without background) after conversion to CIELAB from RGB

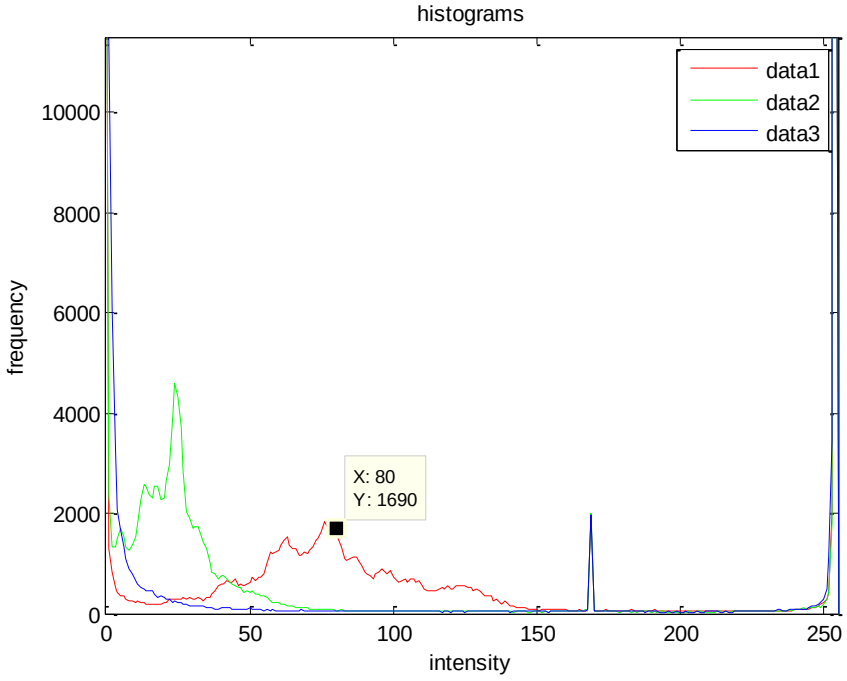
(b) Original image (b) Segmented image

Figure 3.9: Image of scarce 7-spot ladybird after colour segmentation showing rough segments of reddish colour and illumination effects

The RGB intensity histograms are compared, as shown in Figure 3.10 (a) and (b).



(a) With background



(b) Without background

Figure 3.10: Comparison of intensity histograms (RGB)

The histogram shows that background clutter has influence on the meaningful interpretation of the scarce 7-spot ladybird image. It suggests the image without background can be manipulated to obtain elytra markings, including spots and pronotum patterns. The whole test shows that:

- There is a limitation to the use of CIELAB for segmentation purpose. The scarce 7-spot image without background reveals elytra markings better than the image with background (refer Appendix II).
- Segmentation of ladybird images should be done in RGB, if required. The way forward is to perform body marking measurements in greyscale and convert the image into binary format. The colours of the body markings can be captured in CIELAB values.

3.4.3 Ladybird Colour Distributions

Based on the previous observations, it is conclusive that the spot colour and the elytra colour need to be captured in CIELAB colour space. For each image, CIELAB values are obtained by reading the average L^*, a^* and b^* values from a user-interactive pixel capture box. The size of this capture box is not fixed. It varies between 25x25 to 100x100 square pixels depending on the image resolution, hence quite user-dependent. Higher resolution images need only smaller capture box, and vice versa. If the size of the capture box were fixed, and the image is of low-resolution then some level of magnification make border pixels indistinct and blurry. Once the average values were obtained, each value was normalised to [-1,1]. Figures 3.11-3.14 show the representation of the spot colour and elytra colour (in CIELAB values) on normalised colour planes.

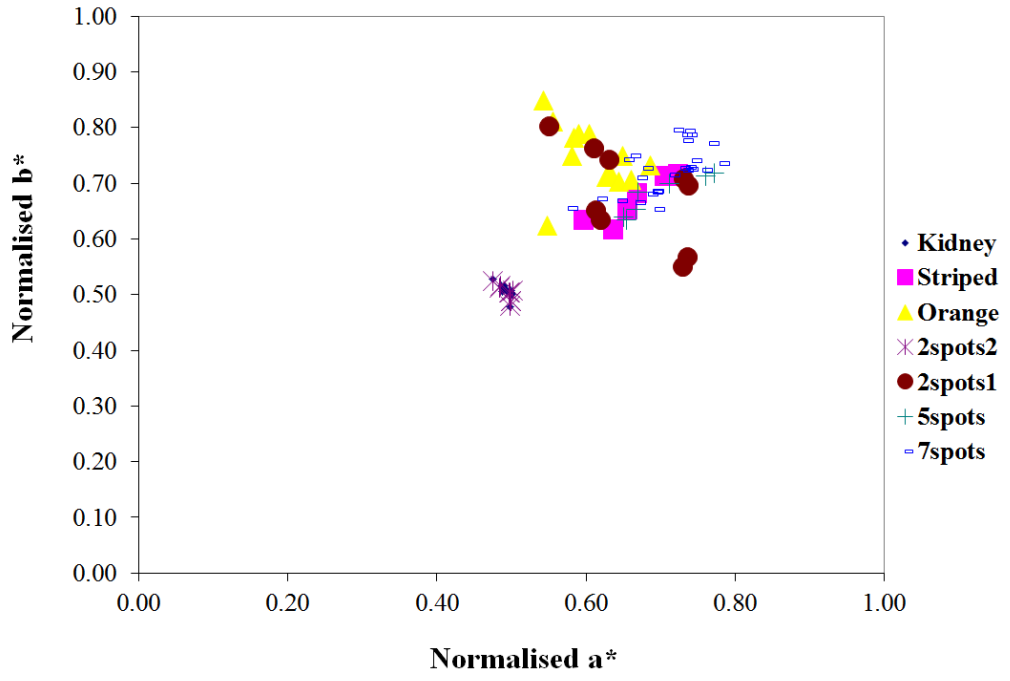


Figure 3.11: Elytra colour distributions among local ladybird species

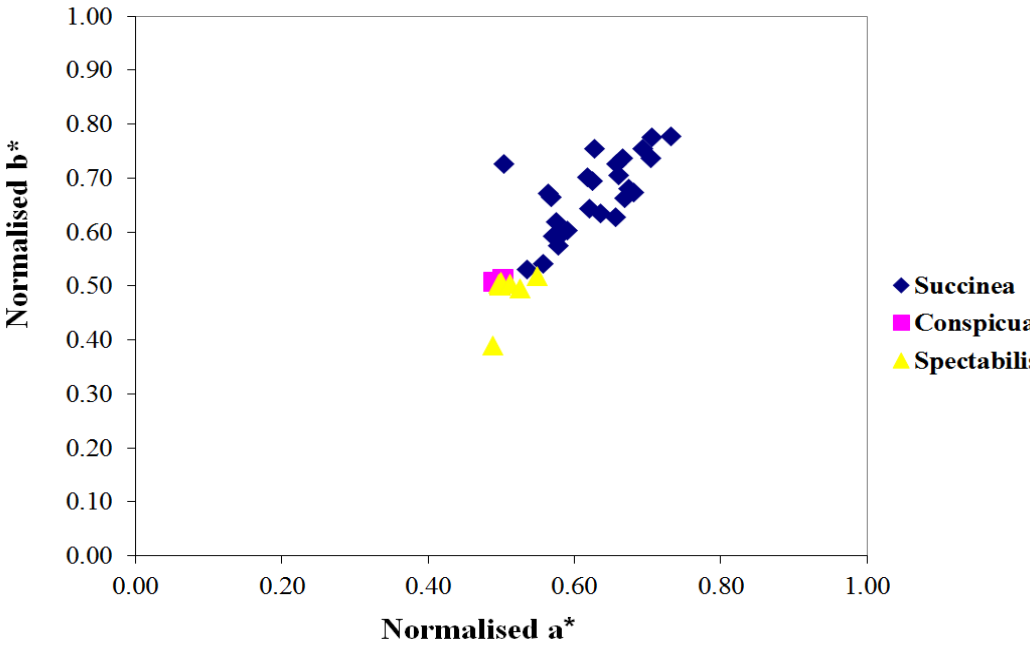


Figure 3.12: Elytra colour distributions among harlequins

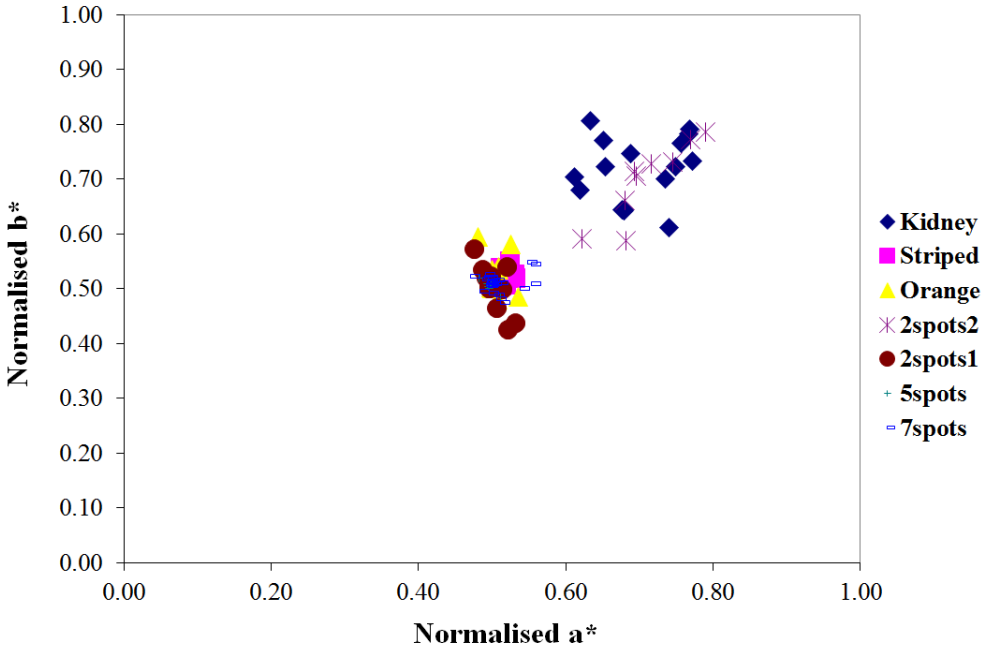


Figure 3.13: Spot colour distributions among local ladybird species

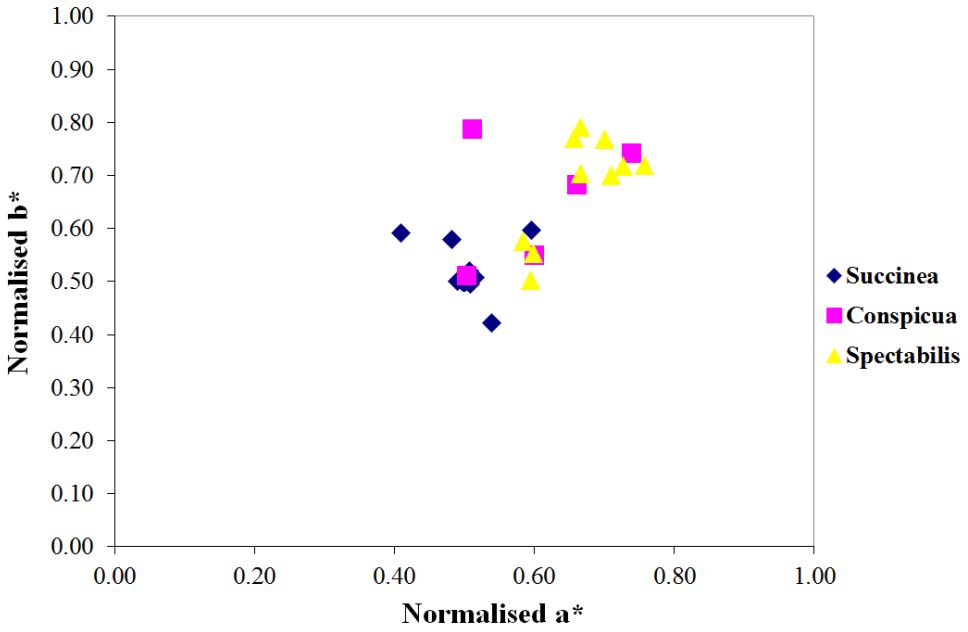


Figure 3.14: Spot colour distributions among harlequins

The normalisation was based on the following formulae:

$$Normalised\ L^* = (L^*/100) \qquad (1)$$

$$Normalised\ a^* = (a^*+120 / 240) \qquad (2)$$

$$Normalised\ b^* = (b^*+120 / 240) \qquad (3)$$

These values were obtained from various species of local ladybirds and harlequins. From the elytra colour planes, it shows how the colours of the elytra (wing case) are distributed. They are non-linearly clustered in some way. The pair of values clustered near the bottom left are darker, or close to black, than those clustered near the top right hand corner. This explains why values for kidney spot ladybird are located near the bottom left corner, and why species like two-spot and seven-spot ladybird appear to be near the top right corner. Intermediate values like orange/yellow group together around the top left corner, while reddish colours group together near the bottom right hand corner. The same discussion and explanation can be made for the spot's colour planes.

3.5 Greyscale and Binary Pre-processing Steps

For pre-processing the images to get geometrical measurements, greyscale and binary image processing were performed rather than using colour image processing as it involves minimal complications to perform binary processing in one channel. Initially images have been converted to greyscale via the MATLAB function 'rgb2gray' and resized to 640x480 pixels. Figure 3.15 shows the greyscale processing steps. The explanation of each block follows.

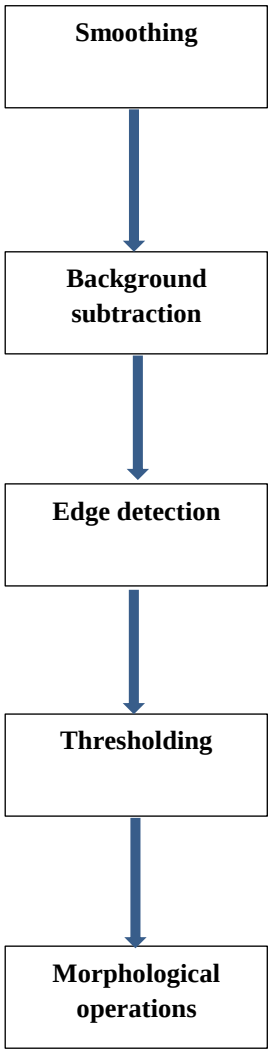


Figure 3.15: Greyscale operations

3.5.1 Smoothing

Smoothing is filtering an image from noise or distortions using neighbourhood filter operations (Gonzalez & Woods, 1992; Jähne, 1995). The form of filtering employed was spatial average filtering or box filter. The effect was to have neighbouring pixels been divided by a common scalar. For example:

$$M = \frac{1}{9}\begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix} = \begin{bmatrix} \frac{1}{9} & \frac{1}{9} & \frac{1}{9} \\ \frac{1}{9} & \frac{1}{9} & \frac{1}{9} \\ \frac{1}{9} & \frac{1}{9} & \frac{1}{9} \end{bmatrix}$$

3.5.2 Background subtraction

Some images have unwanted background. This requires background subtraction, apart from cropping, to reduce clutter for next stages use. This is manually done by cutting through the edge of the elytra using 'scissor' tool in GIMP2. The portion of unnecessary background is removed. Alternatively, 'free-select' and 'eraser' tools can be used sparingly depending on the complexity of image.

3.5.3 Edge Detection

Edge detection determines whether an edge exists between two neighbouring pixels by comparing their relative intensities (Bovik, 2000; Gonzalez and Woods, 1992). It produces rough shape out of the greyscale image. A number of edge detector operators were considered, including Roberts', Canny's and Sobel's.

The Roberts edge detector operator is based on two masks which give a measure of intensity changes in a diagonal direction. This gradient magnitude is calculated by computing the square root of the sum of the squares of the differences between diagonally adjacent pixels. An edge is detected when the gradient magnitude exceeds a threshold (Gonzalez and Woods, 1992).

The masks are:

$$\begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} \text{ and } \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}$$

Canny's edge detector is a multi-stage operator. First, smoothing of the image is performed by Gaussian convolution. Next, a 2-dimensional first derivative operator (for example, Roberts operator) is applied to highlight regions in the image (Canny, 1986). Edges are shown as ridges. Then, the algorithm tracks the top of the ridges and sets to zero all pixels that are not on top of the ridge. The resulting image will contain some thin lines due to the detected edges.

Sobel operator uses two 3-by-3 masks for searching a larger neighbourhood. Theoretically, this should give better results than Roberts operator. It calculates the gradient of the image intensity and weighs the pixels closer to the centre with higher values compared to others (Boyle & Thomas 1988, p.52). Given pixel (i,j), and masks as:

$$\begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix} \text{ and } \begin{bmatrix} 1 & 2 & 1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{bmatrix}$$

Calculation-wise, applying the masks on the image produces:

Δ_1 = (f(i+1,j-1) – f(i-1,j-1)) + 2(f(i+1,j) – f(i-1,j)) + (f(i+1,j+1) – f(i-1,j+1))

Δ_2 = (f(i-1,j+1) – f(i-1,j-1)) + 2(f(i,j+1) – f(i,j-1)) + (f(i+1,j+1) – f(i+1,j-1))

where the gradient,

$$S(i,j) = \sqrt{\Delta_1{}^2 + \Delta_2{}^2}$$

and the gradient direction,

$$\theta = \tan^{-1}\left(\frac{\Delta_2}{\Delta_1}\right)$$

In general, the algorithm for an edge detector performs the following steps:

For each pixel (i,j) in the image:

If (i,j) differs from (i-1,j) or (i,j-1) by more than a specified threshold then edge is found

Else

No edge found

End if

End for

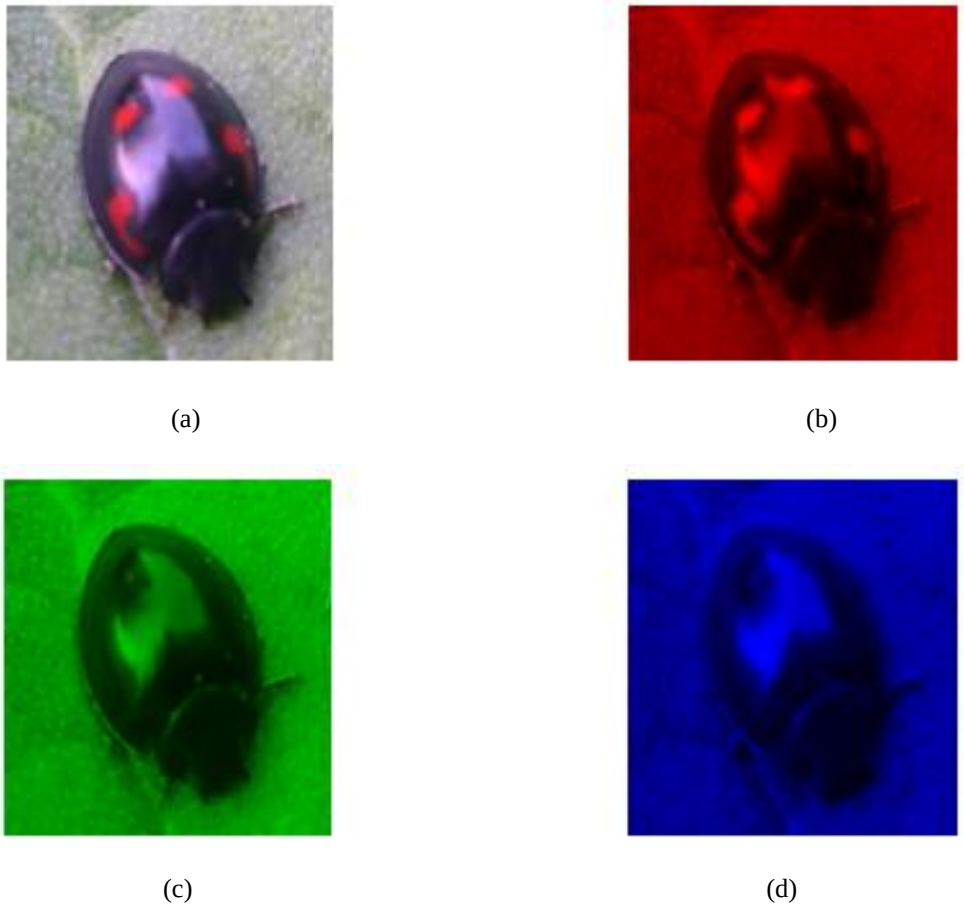
3.5.4 Thresholding

This is done by setting a threshold beyond which a pixel value is set to 0 or 1. Thresholding effectively produces binary, or black and white output. This effect is obtained by finding a suitable threshold using intensity histogram (Low, 1991). The intensity histogram shows two peaks; one belongs to the background, and the other is foreground peak. Ideally, the two peaks should not overlap because the background is uniform. Threshold is selected to be the optimum point between the two non-overlapping peaks.

3.5.5 Morphological Operations

At this stage edge detection and thresholding produces rough shape of the ladybird, including noisy dots. In some images noise can surmount a whole area, making it looks like a genuine ladybird spot. Morphological operations such as erosion and dilation were performed to reduce the binary noise. Both morphological operations work using Minkowski set addition or subtraction by applying structuring element (Haralick, 1987). In MATLAB it is referred to as 'strel'. Strel can have a user-specified 2D polygon shapes and radius. Common shape is circular or disk with specific radius.

Erosion performs Minkowski set subtraction of the structuring element from the image. The intersection of the structuring element and the image is found by looking for the overlap of the origin of the structuring element corresponding to a pixel in the image which belongs to the segmented region. For dilation, when the structuring element is stepped over the image it forms ‘union’ with that part of the image. In effect, it fills in small holes in the region and making the region expands slightly. Effectively, erosion removes spikes from the edges of a region, while dilation perform filling on a region’s edge valley (Low, 1991; MathWorks, 2012). This is shown in Figure 3.16.



(a) (b)

(c) (d)

Figure 3.16: Images of (a) an average filtered pine ladybird and (b), (c) and (d) its RGB constituents

When the two operations are performed one after another, they are called opening and closing. Opening is performed when dilation is performed after performing erosion. In reverse, closing performs dilation followed by erosion. Figure 3.17 shows some of the operations on a ladybird image.

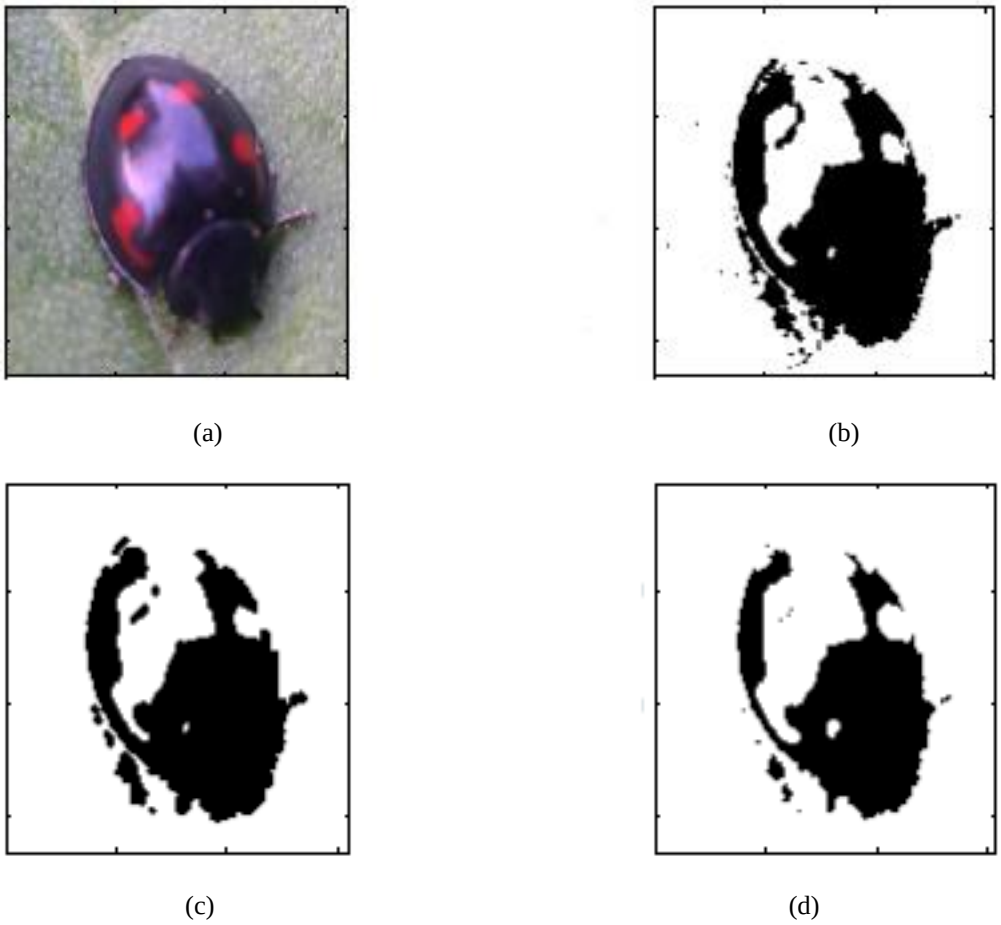
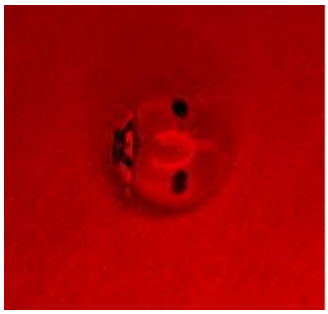


Figure 3.17: Images of (a) an average filtered pine ladybird and (b) Global thresholding, (c) Closing and (d) Dilation

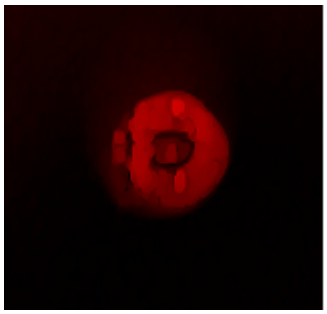
An image of *A. 2-punctata* (two-spot ladybird) is used to illustrate the resultant of each preceding processes. This is shown in Figure 3.18 and Figure 3.19.

(a) (b)

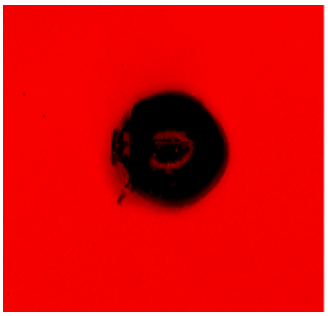
Figure 3.18: *A. 2-punctata* (a) original image (b) completed greyscale and binary pre-processing



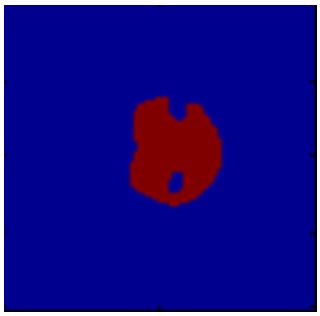
(a) Smoothing/Filter



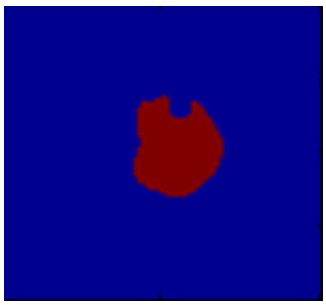
(b) Background subtraction



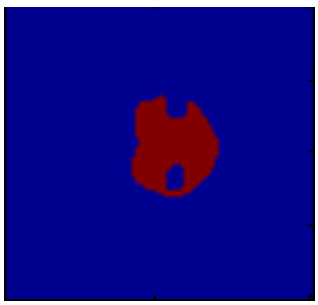
(c) After subtraction



(d) Opening (erosion, then dilation)



(e) Hole filling



(f) Erosion

Figure 3.19: Greyscale and binary pre-processing for *A. 2-punctata* (not to scale)

3.5.6 Geometrical Measurements

Once the image has completed the pre-processing stage and finally in a binary form, the next important process will be measurements of body markings. It includes the spots. At this moment only the shape is known, and no colour information is available. Working in binary made measurements simpler as there is only one channel to work with, therefore it minimises computations. The following steps are taken to produce geometrical measurements on body markings, first by obtaining geometrical properties of the objects in the image and then doing the same for spots (Shouche et al., 2001; Du and Sun, 2004; Eddins, n.d.):

1. Assuming the image is in binary form, use 'bwlabel' function to label all connected components using the default 8-connectivity. Alternative value is 4-connectivity. The function 'bwlabel' returns a matrix L of the same size as the binary image, which contains labels for the connected components in the binary image. The labels are set to a max value. The resultant label is optimum if objects are not touching each other, else they will be counted as one object.
2. The elements of matrix L contain integer values equal to 0, or greater. Background is labelled 0, while pixels labelled 1 belong to the first object, label 2 belongs to the next object, etc.
3. Apply the function 'regionprop' which extracts geometrical features by measuring a set of properties for each connected component in the binary image. The properties are geometrical measurements and pixel value measurements.
4. Apply 'bwboundaries' function to trace the region boundaries in the binary image. It returns a cell array, each cell contains the row and column

coordinates for an object in the binary image. Using the coordinates, plot the borders of all the spots on the original greyscale image. Using this function allows objects to be displayed in a particular colour, and holes in a different colour for better visualisation.

5. Measure the geometrical properties of the object(s) in the image. These are listed in Tables 3.1 and 3.2.
6. Repeat steps 1 to 5 to determine geometrical properties of markings in the object.

Table 3.1: Geometrical properties determined by 'regionprop'

Area	Convex image	Extrema	Minor axis length	Solidity
Bounding box	Eccentricity	Filled area	Orientation	Subarray index
Centroid	Equivalent Diameter	Filled image	Perimeter	
Convex area	Euler number	Image	Pixel index list	
Convex hull	Extent	Major axis length	Pixel list	

Table 3.2: Pixel value properties determined by 'regionprop'

Max intensity	Mean intensity	Min intensity
Pixel values	Weighted centroid	

Once body markings are labelled, counted and measured they are then sorted out to filter genuine spots from noise. Sorting is performed by setting a threshold to the markings' area ratio. Area ratio is the ratio between the area of the object (elytra) and the area of the marking. Any markings with area ratio greater than 5 will be discarded. This is a non-optimal technique as compared to automated elimination technique such as using circularity; however, it is able to perform well (Chang *et al.*, 2011). The flow chart in Figure 3.20 addresses all processes performed in this chapter.

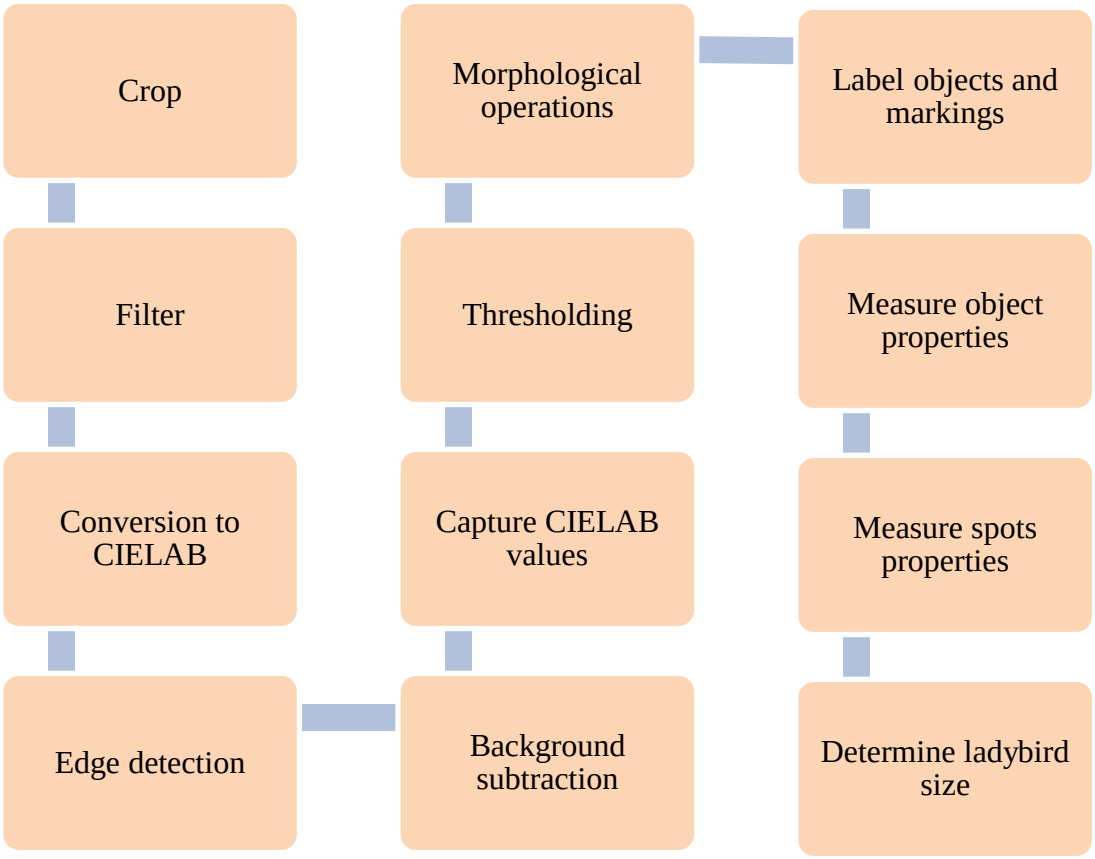


Figure 3.20: Flow chart showing image processing techniques used in the thesis

3.6 Summary

Image processing of the input images is not trivial due to various constraints. Binary processing was performed in order to get meaningful geometrical measurements. The CIELAB colour distributions for both elytra and spot of ladybirds are valuable information to be used in the next stages. These operations were only performed semi-automatically during which user interaction is required to crop and capture pixel colour information. While this seems non-ideal, as a piece of pioneering work in ladybird identification the use of geometrical measurements and the application of CIELAB colour space in this area are novel findings. The author foresees this as a breakthrough towards a workable feature representation for an automated identification system. This is discussed further in the next chapter on feature extraction.

CHAPTER 4

FEATURE EXTRACTION AND CLASSIFICATION

CHAPTER 4

FEATURE EXTRACTION AND CLASSIFICATION

Feature extraction is the most important part of the system, whereby features are selected and processed. Some typical features for images are shape, texture, size, colour, mean or standard deviation of RGB components. Image data is high dimensional, therefore it requires specific pre-processing to find a subset of variables based on the image data, therefore lowering computational cost (Egmont-Petersen, de Ridder and Handels, 2002; Nixon and Aguado, 2008). A perfect feature set is where each taxon has a 1:1 correspondence with a set of features. If feature overlap occurs then it will produce false identification (Chesmore, 2007). Clearly, feature extraction is a necessary step for image segmentation or object recognition to be successful.

4.1 Introduction

Other work dealing with manual species identification vary in many ways, even though the same concept of taxonomy applies. Some systems use single key inputs,

while others use multiple-keys. Dichotomous keys, for example, tend to use monothetic identification criteria (Woolley and Stone, 1987). Even though computerized keys have been developed and used, they suffer from the same inherent problem of being sensitive to individual characters. Furthermore, they give incorrect identification if characters are missing or assessed incorrectly (Woolley and Stone, 1987; Boddy *et al.*, 1998).

Using colour images have an added advantage for this kind of problem, apart from the visualisation for an observer. When CIELAB was introduced in the previous chapter, the colour planes showed some interesting clusters of colours. These clusters contain the vectors of a^* and b^* values. They have coordinates or location, and can be analytically determined such as calculating the distance between vectors using Euclidean distance measure. The same can be deduced for ladybirds from the same species, or perhaps from other species. To date there has not been any manual technique or automated system that has extracted this piece of visual data.

What an observer will notice from the CIELAB colour distributions are:

- Values representing ladybirds from the same species tend to group together.
- Values representing ladybirds from the same species but of different colour forms do not cluster together.
- Both elytra colours and spot colours show similar trends.

Apart from colour representation, there are also issues on which feature(s) is the best representation of the ladybird. These can not be determined visually, therefore some forms of analysis is needed. It would be ideal to have a single feature or character for species identification. A succesful single feature classification means there is no overlapping of feature distributions between taxa. Unfortunately, this demands tight

constraint on the input stage. Feature vectors from the same class may also be different. Theoretically, the differences are due to three factors (Looney, 1997):

1. Noise.
2. Bias in the measurement, data acquisition system and pre-processing.
3. The natural variation between objects within the same class (intra-class) due to unknown influence.

Classification is defined as the assignment of a signal or pattern to one of a number of classes based on features extracted (Schalkoff, 1992). Based on the various features, the system will next assign data to one or more specified classes. This chapter discusses methods of feature selection and classification in detail. The use of decision tree to select and simplify rules is explained, together with some mathematical treatments on Bayesian probability. There will be a section on classifiers, particularly on neural networks and the backpropagation algorithm.

4.2 Datasets

To get all feature datasets, the following steps were done:

Step 1: Read input image. Perform average filtering.

Step 2: Crop region of interest (ROI).

Step 3: Perform morphological operations to reduce noise and unnecessary segments.

Step 4: Obtain geometrical measurements.

Step 5: Perform colour space conversion from RGB to CIELAB.

Step 6: Get average values of a^* and b^* for both spots and elytra region.

Step 7: Normalise all feature sets.

Step 8: Repeat for next input image.

4.2.1 Geometrical features

There were six features obtained through the measurements on the elytra of each ladybird, four of which were primary measurements and the other two were derived from the primary. These features are listed in Table 4.1. Elytra measurements were also taken for normalisation purpose.

Table 4.1: Geometrical features and descriptions

Features	**Descriptions**
Spot area	Total count of pixel with binary value 1 in a spot
Spot perimeter	Total count of pixel along the circumference of a spot
Spot max axis length	The length of the longest line drawn between two points in the spot
Spot min axis length	The length of the longest line drawn between two points perpendicular to the max axis
Spot area ratio	Area divided by the product of max axis length and min axis length in a spot
Spot aspect ratio	Ratio between max axis length and min axis length

Generally, geometrical values tend to be bound by the effect of rotation and scale. Here, they are rotation and scale invariant due to the way measurements were taken. For example, quantities like area ratio and aspect ratio are made scale invariant by having measurements taken on both the longest and shortest distance of the spot, then the spot measurements are normalised by the size of the elytra.

4.2.2 Colour features

Colour features have been generated via the use of a capture box, both during initial development and data collection stage. This method was perceived to give good collection of pixel values within the box and around the vicinity of the capture box. The hue angles were derived from primary colour features a^* and b^*. Figure 4.1 shows an example of colour extraction utilising *'roipoly'* and *'impixel'*, which are two useful built-in MATLAB functions.

Figure 4.1: Example of elytra and spot colour acquisition from image

The function *'roipoly'* lets users to select a polygon region on the elytra and spot. This is done through clicking interactively using mouse to produce a polygon. The corresponding values of CIELAB quantities are acquired through the function *'impixel'* when user is finished with setting the polygon. The primary quantities are lightness, spot colour a^*, spot colour b^*, elytra colour a^* and elytra colour b^*. Consider an arbitrary point on the CIELAB colour plane. This point on the plane is actually a vector consisting of magnitude and angle made of both axes, a^* and b^*. The magnitude is called chroma $|C^*_{ab}|$ and hue angle called h_{ab}, given by the above formulae (Fundamentals of colorimetry, 2012).

$$Chroma, \ |C^*_{ab}| = \sqrt{a^{*2} + b^{*2}} \qquad (4.1)$$

$$Hue\ angle, \ \ h_{ab} = \tan^{-1}\left(\frac{b^*}{a^*}\right) \qquad (4.2)$$

In total there are eight quantities, however, lightness and chroma values are finally ignored to make only six usable colour features. The author rejected the two quantities for the sake of reducing the number of features, hence trying to reduce dimensionality.

4.3 Data trimming and normalisation

The idea is to check that the data or features have been stripped of outliers. This is done as a preprocessing stage before further work is done. Normalisation has been performed on the ladybird data by having each feature limited to a range of maximum and minimum values in the interval of [0,1], except for a^* and b^* which use [-1,1]. For example, a normalised version of Area Ratio is obtained through the following formula:

$$Normalised\ Area\ Ratio = \frac{Area\ Ratio - Min(Area\ Ratio)}{Max(Area\ Ratio) - Min(Area\ Ratio)} \tag{4.3}$$

By doing so, the feature values are limited to a predetermined range and have equal influence on the classifier as would other feature values. For the CIELAB colour quantities normalisation on the primary values have been shown in the previous chapter. For a secondary quantity like chroma, normalisation limits the values to [0,1]. For hue angle, normalisation means making the maximum values to be limited to 1, and minimum to -1. The normalisation formula are given as:

$$Chroma \quad |C^*_{ab}| = \left(\frac{\sqrt{a^{*2} + b^{*2}}}{\sqrt{2}}\right) \tag{4.4}$$

$$Hue\ angle \quad h_{ab} = \left(\frac{180}{\pi}\right) \tan^{-1}\left(\frac{b^*}{a^*}\right) \tag{4.5}$$

4.4 Feature selection

It has been shown in the previous sections how features have been generated from colour and geometrical properties, where both are the physical traits of a ladybird. These are multidimensional quantities projected in the feature space, in which case

there is a need for the features to be carefully selected. Some of these features are totally ‘unconventional’, unlike the manual key-based systems. Features like area ratio, aspect ratio and spot perimeter are some of which have never been obtained before for a typical ladybird.

It is quite difficult for a system to be trained to use all these features, especially when the number of features are large because the number of training exemplars need also be huge. For a classifier, this means searching for the best features which have the best discriminatory power during classification. Without this ‘weighting’ process a classifier would make poor decisions due to undertraining or overfitting, therefore affecting the overall system performance. Methods exist that help a researcher to select best features, and most of the time statistical techniques have been widely employed. In this work, decision tree has been used for feature selection. Before its inclusion in the whole identification system, it is important to consider a qualifying factor called the dissimilarity coefficient.

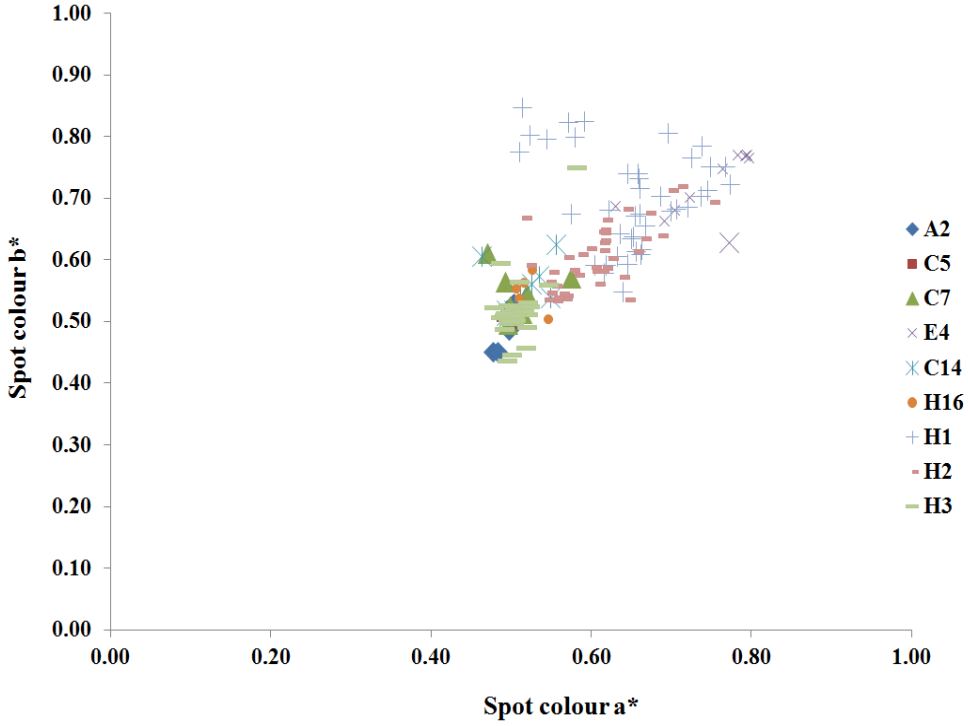
4.5 Dissimilarity coefficient estimation

There are a couple of factors that contributes towards the level of precision. There are some input errors obtained during the process of getting colour information from both spot and elytra. There are also variations of the CIELAB colour values among the different individuals in the same OTU. This is called intra-OTU variation. There is also variation of the CIELAB colour values among different OTUs, which is called inter-OTU variation. These factors can be estimated using dissimilarity coefficients, obtained from calculating the Euclidean distance between the points belonging to OTUs in CIELAB colour plane.

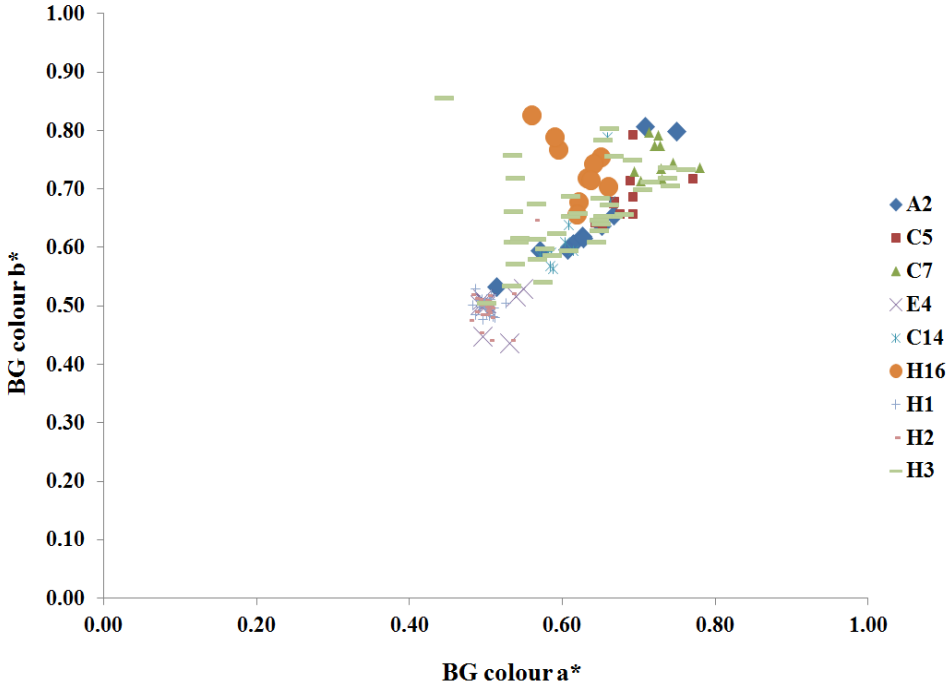
This method has been adopted from Liu (1996), who investigated the use of only vein points as a single character input to an expert system for the identification of Tortricinae (Lepidoptera) (Liu, 1996). Suppose the colour planes in Figure 4.2 are generated for all OTUs; one for the spot colour plane, and another for the elytra colour plane. To determine the dissimilarity coefficient between two arbitrary points on the colour plane, namely $S(a^*_{si}, b^*_{si})$ and $N(a^*_{ni}, b^*_{ni})$. The dissimilarity coefficient, *D(sn)*, is calculated using the Euclidean distance formula:

$$D(sn) = \sqrt{(a^*_{si} - a^*_{ni})^2 + (b^*_{si} - b^*_{ni})^2} \quad (4.6)$$

For the ladybird identification, 90 sample data were used with 10 samples from each OTU. The intra-OTU variation was calculated and averaged as variance within three forms of *H. axyridis*. For the inter-OTU variance, the dissimilarity coefficients were calculated and averaged as variance between each two different OTUs. The coefficients are presented in Appendix III.



(a)



(b)

Figure 4.2: Colour planes comprising all OTUs (a) spot (b) elytra

The intra-OTU variance was 0.185 ± 0.086, whereas the inter-OTU variance was 0.134 ± 0.126. The results showed that the intra-OTU variance was larger than the inter-OTU variance. With this, it can be concluded that the use of only colour information a^* and b^* is insufficient for the classifier to identify the OTUs, and more features are required. This has prompted the selection of decision tree as a feature selector and minimiser.

4.6 Learning System

In developing an automated identification system, there is a need for a platform for software development such as code developing, testing, etc. In this thesis it is called the learning system. The learning system supports the core functionality of the intelligent system, which also means the use of a reliable platform is essential. For instance, during code development there are many tasks involved and it is crucial to use a reliable software platform. This platform also serves as a starting point for further system redesign in future, if necessary. WEKA and MATLAB R2010 have been utilised for system development and testing. Both MATLAB and WEKA have made the research more explorable due to their capabilities. WEKA was used mainly for the many data mining techniques it contains, and MATLAB was used for image processing and neural network tasks. When MATLAB is not capable to perform certain algorithms or not as efficient for machine learning, WEKA has been very useful for this purpose.

4.6.1 WEKA machine learning toolkit

WEKA is a machine learning tool which stands for 'Waikato Environment for Knowledge Analysis' (Witten and Frank, 2005). It was developed by researchers in

the University of Waikato, New Zealand (Hall *et al.*, 2009). It contains a collection of machine learning programs developed in JAVA to facilitate data mining tasks such as training and testing artificial neural networks, decision trees and statistical visualisations. Classifiers included in WEKA are Bayes, RBF functions, Support Vector Machine (SVM), Multilayer Perceptron (MLP), Learning Vector Quantisation (LVQ), J48 decision tree and many more. The user interface is presented in Figure 4.3.

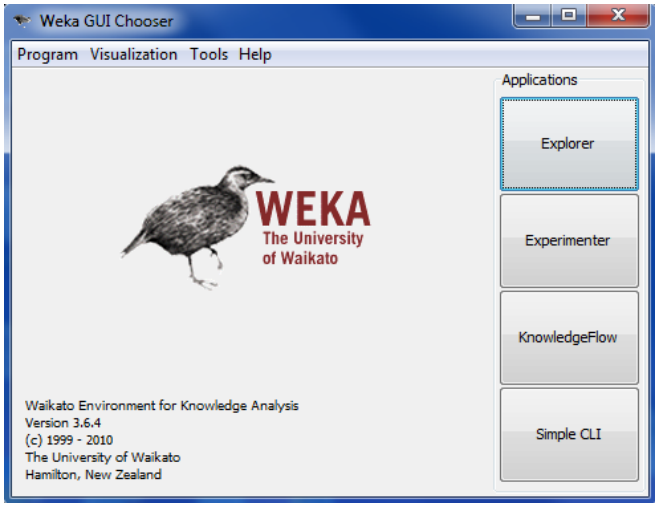


Figure 4.3: WEKA Graphical User Interface (Hall *et al.*, 2009)

A user has four choices when using the program; either as Explorer, Experimenter, KnowledgeFlow or Simple CLI. The author opted for Explorer option for ease of use and implementation.

4.6.2 Decision tree

In WEKA, there are many machine learning techniques that a user can use. The author opted J48 which is an open source Java implementation of the C4.5

algorithm. C4.5 was actually derived from ID3. Both are Ross Quinlan's algorithms for generating classification models, better known as decision trees (Quinlan, 1996; Witten and Frank, 2005; Omid, 2011). It contains a hierarchy of branches and leaves stemming from a root. When a classification is required, a decision tree uses its hierarchical and recursive nature to make decisions at each node.

An example is given in Figure 4.4. Imagine there are 10 samples each for the two dummy classes 'C5' and 'C7'. The most important is to determine which attribute or feature to place at the root (top most node). The decision tree calculates the values of entropy before and after a node. For a binary split, entropy and information gain are given as:

$$Entropy = - p(a)*log_2(p(a)) - p(b)*log_2(p(b)) \quad (4.7)$$

$$Information\ Gain = Entropy\ Before - Entropy\ After \quad (4.8)$$

Witten and Frank uses the term 'information value' instead of entropy (Witten & Frank, 2005). The information gain for each candidate attribute is evaluated at each node, and the attribute with the highest information gain is selected.

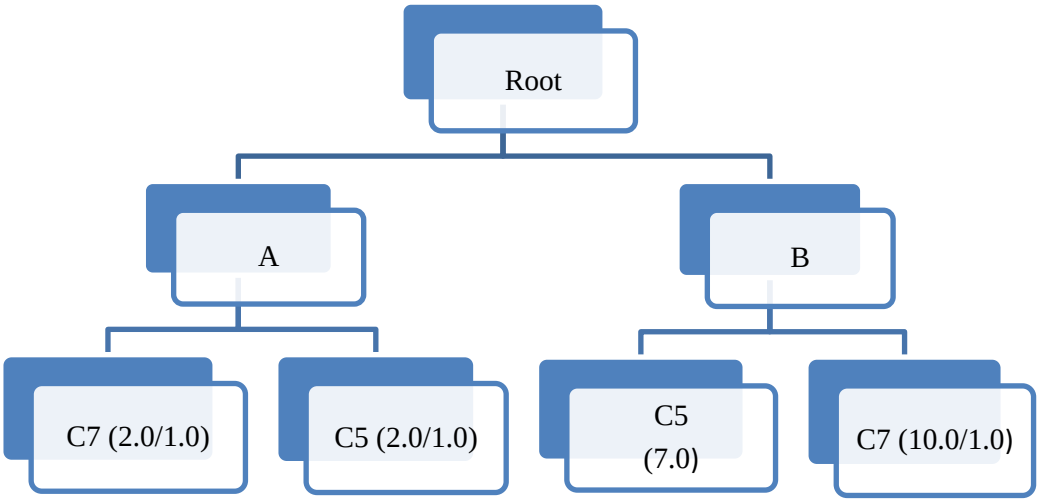
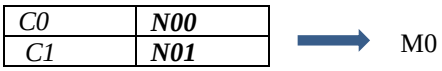


Figure 4.4: Example decision tree for the case of C5 and C7

Taking the tree in Fig 4.4 to determine Information Gain for arbitrary attributes A and B:

Before Split:



After Split:

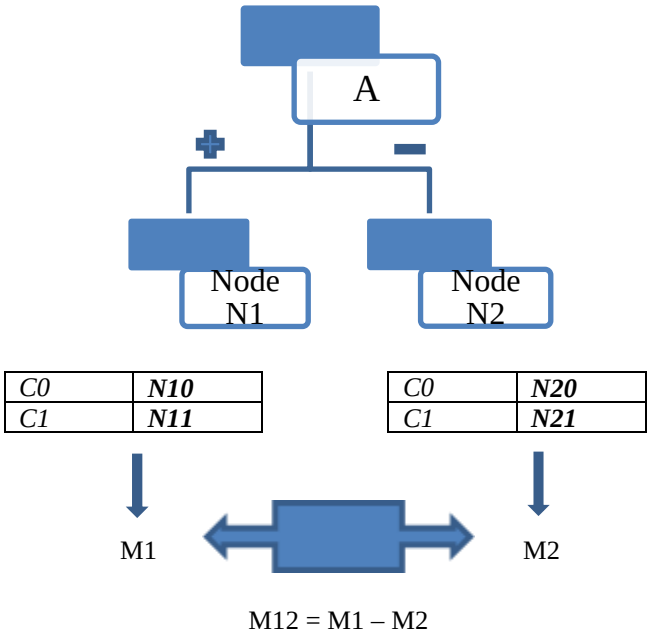


Figure 4.5: Determine entropy for attribute 'A'

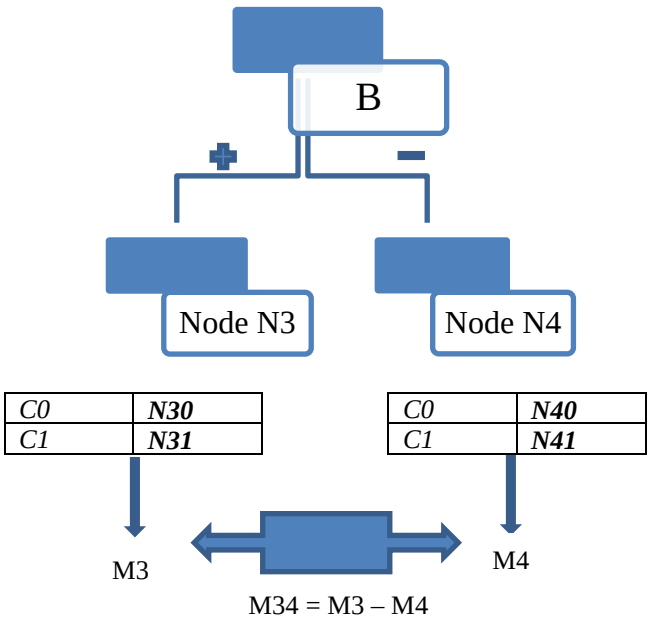


Figure 4.6: Determine entropy for attribute 'B'

The information gain (IG) will be chosen from the attribute with the highest value:

Information Gain = M0 – M12

OR

Information Gain = M0 – M34

To classify an unknown instance, the tree is traversed based on the values tested in successive nodes. If an attribute value is not nominal, the tree will form two subsets or branch. The branching depends on which subset the value lies in the decision tree. In the case of ladybird identification, the attributes are numeric. At a node, the number is checked if it is greater or smaller than a constant. This constant is the split criterion, where binary split occurs. Notice the two numbers at some leaves in Figure 4.4 (the last nodes). The first number represents the total number of instances reaching the leaf. The second is the number of those instances which are misclassified. In short, the decision tree simplifies the solution when looking for which feature to use in a particular identification. It makes automated identification easier by reducing number of features and shorten identification time (Ayob and Chesmore, 2012).

4.6.3 Comparing decision trees with neural networks

Artificial neural networks (ANNs) are a promising technology in computer aided taxonomy, as they learn from examples presented to them rather than rote learning of inputs (Boddy *et. al.*, 2000). ANNs have been used in many areas and proven to work to some extent, which will be discussed later. ANNs have been developed to mimic the human brain and consists of an interconnected set of basic information units called neurons (Ham and Kostanic, 2001; Negnevitsky, 2005). In this section,

the applicability of artificial neural networks and their mathematical model are investigated. A single neuron is connected to other neurons via weighted links, and they form a hierarchy of arranged layers. Each neuron, or node, receives input signals through this link. The weighted inputs are combined to give the internal activation level, hence producing an output signal. An output signal will only be generated once an activation level is triggered. This level is triggered by a factor of the inputs and their associated numerical weights, whereby a neuron calculates the weighted sum of the input signals and compares the result against a threshold value using the following 'sign' activation function (Negnevitsky, 2005):

$$X = \sum_{i=1}^{n} x_i w_i$$

$$Y = sign \{ \sum_{i=1}^{n} x_i w_i - \Theta \} \qquad (4.9)$$

where X is the net weighted input to the neuron, x_i is the value of input i, w_i is the weight of input i, n is the number of neuron inputs, and Y is the output of the neuron. As neurons are interconnected and form layers, they form a network. The network could have one or more hidden layers that do not have direct link to the outside world. They only accept input and generate outputs based on their activation. All neurons in one layer are connected to other layers via unidirectional links that can only transmit in the forward direction (Ham and Kostanic, 2001). In short, a node receives a few signals from its input links, computes an activation level and sends an output signal via the output links, as shown in Figure 4.7.

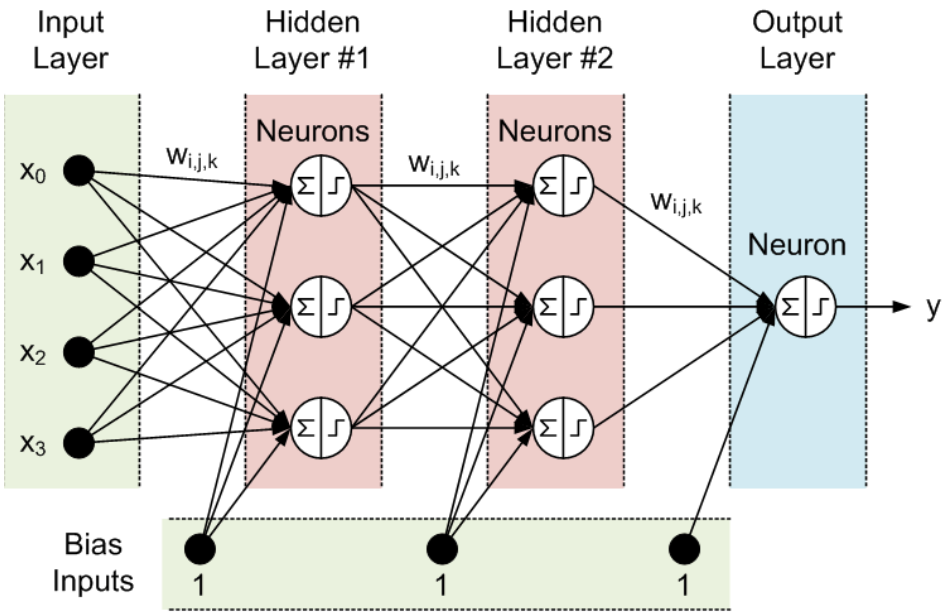


Figure 4.7: Structure of a neuron using mathematical model (Activation function, n.d)

Neurons will learn through repeated adjustments of the weights after a few iterations or after sufficient training. These adjustments of weights represent the enhanced long-term memory in ANN. Therefore, unlike expert system, ANN is not rule-based but learns from patterns presented to the input layer.

With this, there are two schemes of learning, namely 'supervised' and 'unsupervised' learning. *Supervised* learning works by presenting a number of known inputs and the corresponding target outputs to the network, as shown in Figure 4.8.

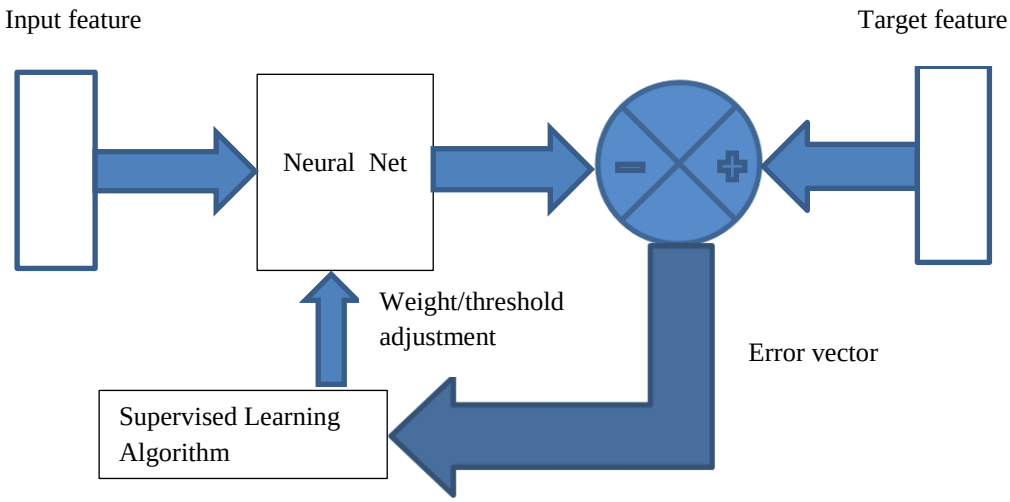


Figure 4.8: Structure of supervised learning
(redrawn from http://www.learnartificialneuralnetworks.com)

After some iterations, the network will adapt its weight based on the patterns. It will then try to classify the correct output category based on the learnt patterns, as if there is an outside teacher guiding the ANN to correctly classify for the particular pattern. Examples of supervised learning are backpropagation and its variants. In contrast, *unsupervised* learning means the ANN is presented with input patterns only, and has no teacher. It will perform 'self-discovery' in detecting similarities of the patterns and forms classification groups or 'clusters' (Kohonen 1990, 2001; Boddy, Morris and Morgan, 1998).

ANN can be useful in this work since they can cope with partially contradictory 'fuzzy' data. Implementation wise, unlike expert systems, they do not need a taxonomic expert beyond the original determinations of example patterns (species) upon which the system is to be trained. A few ANN paradigms are commonly used for identification problems; the multilayer perceptron (MLP), the learning vector

quantization network (LVQ), the radial basis function network (RBF), the asymmetric RBF network (ARBF) (Boddy, Morris & Morgan, 1998).

4.6.4 Multilayer Neural Network & Backpropagation Algorithm

A Multilayer Perceptron (MLP) Neural Network consists of numerous units of perceptrons with one or more hidden layers. A perceptron consists of a single neuron with adjustable synaptic weights and a hard limiter (Negnevitsky, 2005). The weighted sum of the inputs is applied to the hard limiter. The input signals are propagated in a forward direction on a layer-by-layer basis. Neurons in the hidden layer function to detect the features, because the weights of the neurons represent the features hidden in the input pattern. The perceptron gives out +1 if the input is positive, while giving -1 if the input is negative. Therefore, the perceptron behaves as a simple classifier. In other words *n*-dimensional space is divided by a hyper plane into two decision regions.

Central to the operations of a MLP Neural Network is the feed forward and backpropagation algorithm. Feed forward operation works by introducing input to the hidden neuron, firing up neurons, and calculating errors. This is normally done during training stage. Training is done by presenting examples of the input and output relationship to the neural network. The connection weights will be adjusted in order to minimise an error function between the historical outputs and the outputs predicted by the neural network. Backpropagation itself means adjusting weights in hidden layers by propagating errors back towards the input layer. By doing so the changes in input weight and output weight per neuron are calculated (Lang, 2007). In order to perform classification hence identification, a neural network algorithm has to discriminate taxa by constructing decision boundaries. The boundaries are

constructed between example patterns of known taxa in *n*-dimensional space. A simple two-dimensional feature space is shown in Figure 4.9.

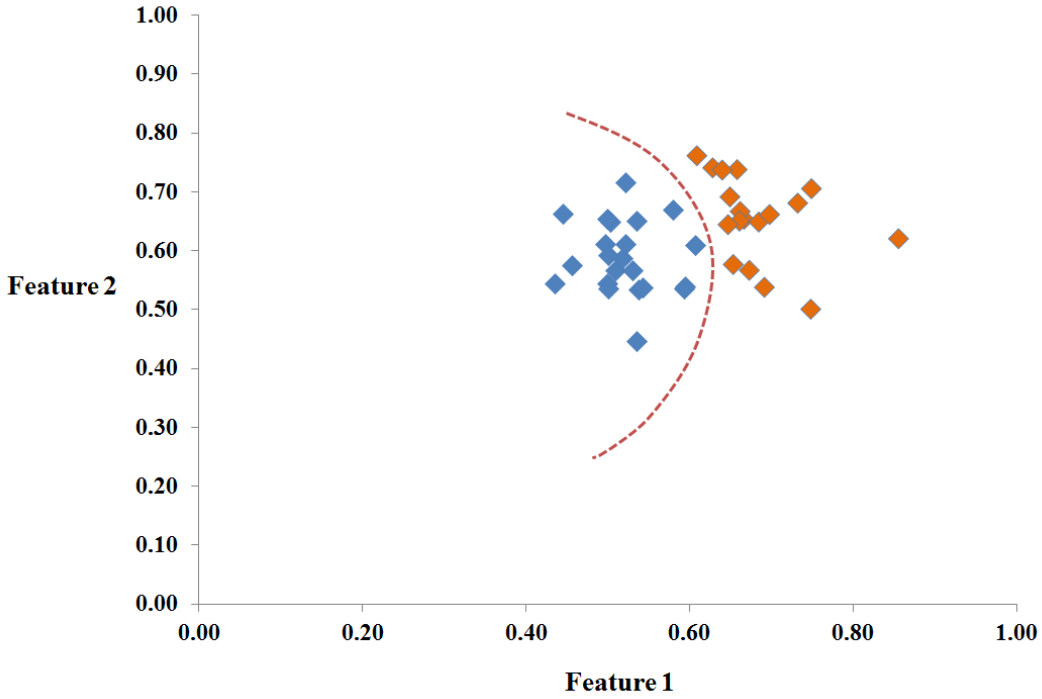


Figure 4.9: Two dimensional feature space and non-linear decision boundary

The author will show that the neural network outputs are estimates of Bayesian probabilities, hence linking the concept with outputs of a decision tree is a possibility. This relationship of decision tree with neural network shows that Bayesian probabilities are estimated using a minimum squared-error cost function. When Bayesian probabilities are correctly estimated, the errors are minimum and outputs are sum to one, hence treating it as probabilities.

The following arguments are excerpts from Richard and Lippmann's paper on 'Neural Network Classifiers Estimate Bayesian *a posteriori* Probabilities' (Richard and Lippmann, 1991). Consider assigning an input vector X $\{x_i : \ i = 1,\ldots, D\}$ to *1 of*

M classes $\{C_i: \ i = 1,\ldots,M\}$. In this case the input values are continuous numbers and the classes are ladybird species.

The network parameters are chosen to minimise the squared-error cost function:

$$\Delta = E\{\sum_{i=1}^{M}[y_i(X) - d_i]^2\} \tag{4.10}$$

where E{.} is the expectation operator.

Using the definition of expectation, and the joint probability of the input and the *i*th class by $p(X,C_i)$:

$$\Delta = \int \sum_{j=1}^{M}\{\sum_{i=1}^{M}[y_i(X) - d_i]^2\}p(X, C_j)\, dX \tag{4.11}$$

For a pair of input X and class C_i each error term in the equation is the difference between the actual output and the desired output d_i. The errors are squared , summed and weighted by $p(X,C_j)$. By definition, $p(X,C_j) = p(C_j \mid X)p(X)$. By substitution into (4.11) gives the following equation:

$$\Delta = \int\left[\sum_{j=1}^{M}\sum_{i=1}^{M}[y_i(X) - d_i]^2\, p(C_j|X)\right]p(X)\, dX \tag{4.12}$$

$$= \int \sum_{k=1}^{M}\left[\sum_{j=1}^{M}\sum_{i=1}^{M}[y_i(X) - d_i]^2\, p(C_j|X)\right] p(X, C_k)dX \tag{4.13}$$

$$= E\left\{\sum_{j=1}^{M}\sum_{i=1}^{M}[y_i(X) - d_i]^2\, p(C_j|X)\right\} \tag{4.14}$$

$$= E\left\{\sum_{j=1}^{M}\sum_{i=1}^{M}\left[y_i^2(X)p(C_j|X) - 2y_i(X)d_i p(C_j|X) + d_i^2 p(C_j|X)\right]\right\} \tag{4.15}$$

The fact that $y_i^2(X)$ is a function of X and $\sum_{j=1}^{M} p(C_j \ |X) = 1$ makes (4.15) into

$$\Delta = E\left\{\sum_{i=1}^{M}\left[y_i^2(X) - 2y_i(X)\sum_{j=1}^{M} d_i p(\, C_j, X) + \sum_{j=1}^{M} d_i^2\, p(C_j, X)\right]\right\} \tag{4.16}$$

$$= E\{\sum_{i=1}^{M}[y_i^2(X) - 2y_i(X)E\{d_i|X\} + E\{d_i^2|X\}]\} \tag{4.17}$$

where $E\{d_i|X\}$and $E\{d_i^2|X\}$ are the conditional expectations of d_i and d_i^2. Given the conditional variance of d_i is $var\{d_i|X\} = E\{d_i^2|X\} - E^2\{d_i|X\}$, then (4.17) is expanded to become:

$$\Delta = E\{\sum_{i=1}^{M}[y_i^2(X) - 2y_i(X)E\{d_i|X\} + E^2\{d_i|X\} + E\{d_i^2|X\} - E^2\{d_i|X\}]\} \tag{4.18}$$

$$= E\{\sum_{i=1}^{M}[y_i(X) - E\{d_i|X\}]^2\} + E\{\sum_{i=1}^{M} var\{d_i|X\}\} \quad (4.19)$$

The first expectation term is the mean squared error between network outputs $y_i(X)$ and the conditional expectation of the desired outputs. The target is to minimise the squared-error cost function, consequently the network parameters are chosen to minimise the first term. For a *1 of M* problem when input X belongs to class C_i , d_i will equal the value of 1, else equals zero. Consequently, the conditional expectations will be:

$$E\{d_i|X\} = \sum_{j=1}^{M} d_i p(C_j|X) \quad (4.20)$$

$$= p(C_i|X) \quad (4.21)$$

which is actually the conditional probability of class C_i given the input X.

Apart form squared-error cost function commonly used by backpropagation algorithm for choosing network parameters (eg. updating connection weights), an alternative cost function is the cross-entropy cost function.

4.6.5 Cross-entropy cross function

This cost function is motivated by the assumption that the desired outputs are independent and binary random variables. The actual network outputs will be continuous, and represent the conditional probabilities that the binary, random variables are 1 (Richard and Lippmann, 1991). When the desired outputs are 0 and 1, the cross-entropy cost function will be

$$\Delta = -E\{\sum_{i=1}^{M}[d_i log\ y_i(X) + (1 - d_i)\log(1 - y_i(X))]\} \quad (4.22)$$

The cross-entropy cost function weights errors more heavily when actual outputs are closer to 0 and 1. When desired outputs are binary, the cross-entropy cost function is

minimised when network outputs estimate Bayesian probabilities. Assuming the network outputs are binary, the following cross-entropy cost function is obtained:

$$\Delta = -E\left\{\sum_{i=1}^{M}[E\{d_i|X\}\log_2 y_i(X) + (1 - E\{d_i|X\})\log_2(1 - y_i(X))]\right\}$$

$$= -E\{\sum_{i=1}^{M}[E\{d_i|X\}\log_2 y_i\,(X) - E\{d_i|X\}\log_2 E\{d_i|X\} + E\{d_i|X\}\log_2 E\{d_i|X\} + (1 - E\{d_i|X\})\log_2(1 - y_i(X)) - (1 - E\{d_i|X\})\log_2(1 - E\{d_i|X\}) + (1 - E\{d_i|X\})\log_2(1 - E\{d_i|X\})]\} \quad (4.23)$$

$$\Delta = -E\left\{\sum_{i=1}^{M}\left[E\{d_i|X\}\log_2\frac{y_i(X)}{E\{d_i(X)\}} - (1 - E\{d_i|X\})\log_2\frac{1-y_i(X)}{1-E\{d_i|X\}}\right]\right\} -$$

$$E\{\sum_{i=1}^{M}[E\{d_i|X\}\log_2 E\{d_i|X\} + (1 - E\{d_i|X\})\log_2(1 - E\{d_i|X\})]\} \quad (4.24)$$

The first expectation term in (4.24) is minimised when $y_i(X) = E\{d_i|X\}$ for i = 1,…, M. The outputs is an estimate of the conditional expectations of the desired outputs since the target is to minimise the cross-entropy cost function by choosing network parameters. When the desired outputs are binary, the conditional expectations are the conditional probabilities of the desired outputs being 1. For the case of *1 of M* problems, the conditional expectations are Bayesian probabilities (Richard and Lippmann, 1991). It is desirable to see whether the same conditions hold for multilayer perceptrons trained using backpropagation algorithm, and radial basis functions in Probabilistic Neural Network (PNN). It is also interesting to see the relationship between the outputs of a neural network with a decision tree through WEKA and MATLAB simulation, as derived in (4.7). The simulation results are presented in the next chapter.

4.7 Summary

This chapter has elaborated on the concept of feature extraction and classifiers. There are two types of features for the ladybirds: geometrical and colour. Feature selection is important to reduce dimensions and computations. After feature selection, the features have been normalised before they were fed into classifiers. Classifiers function to perform input/output mapping. Neural networks require the minimisation of cost functions through connection weight adjustments. Neural network outputs are estimates of Bayesian probabilities, while decision trees use the concept of entropy and information gain to perform decision making at nodes. The link between decision tree and neural network permits their usage in hybrid systems.

CHAPTER 5

CLASSIFIERS

CHAPTER 5

CLASSIFIERS

The previous chapter introduced the concepts on feature extraction, neural networks and decision tree. This chapter considers classifiers in use, how the datasets are partitioned, test run setups, balanced and unbalanced datasets. The chapter aims to elaborate the components of supervised classifiers, as well as maintaining a good research practice to make the processes technically reproducible for future reference (Prechelt, 1995).

5.1 Classifiers

A classifier functions to map unlabeled instances to a label using internal data structures (Kohavi, 1995). For the automated identification of biological species involving at least two classes, where each has its associated class labels, then a classifier will be required. An exception occurs when more species are involved, which will be explained in Chapter 7. The classifiers used are multilayer perceptron artificial neural networks (MLP), Probabilistic Neural Networks (PNN), Learning Vector Quantisation (LVQ) and Support Vector Machine (SVM). They are supervised classifiers, where training is required for the classifier to learn the input

patterns. MLP has been explained in Chapter 4; the other classifiers are introduced here.

5.1.1 Probabilistic Neural Networks (PNN)

If a neural network is used the system needs training exemplars and targets. The training exemplars are fed into a network or classifier, trained to learn some learning functions and produce outputs based on decision boundaries. PNN differ from MLP in many ways; however, the most obvious is the learning function they use. In this work, the function used was a radial basis function (RBF). PNN implements kernel discrimination analysis, meaning the operation are organized into a multilayer feed forward neural network consisting of input layer, radial basis layer and competitive layer (Wu et. al., 2007; MathWorks, 2012). Figure 5.1 shows the PNN network structure.

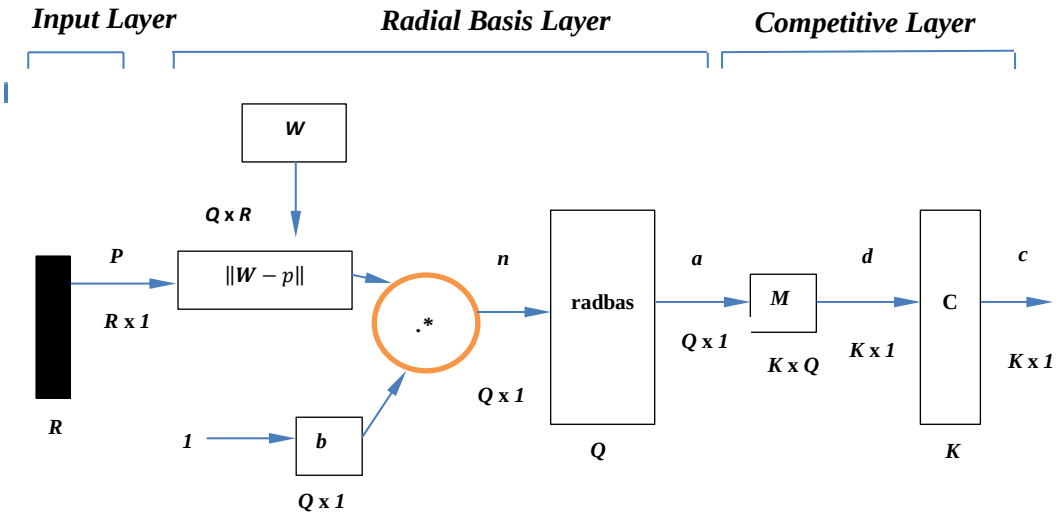


Figure 5.1: Network structure of PNN

The input layer consists of nodes that receive the data, labelled as *P* of size Rx1. The radial basis layer contains a probability density function (pdf), using a given set of data points as centres. In this layer the vector distance between the weight vector and the input vector *p* are calculated, making a vector of size QxR. This is mathematically done using the dot product notation. Next, the vector distance is multiplied with the bias vector *b* through element-by-element multiplication, shown as '.*' in Figure 5.1. This effectively produces a resultant *n* of size Qx1. The resultant is fed into the radial basis function '*radbas*' and correspondingly produces the output vector *a*. When the input *p* is identical to the i-th row of weight matrix *W*, it is assigned a value of 1. Consequently, a neuron with weight close to the input vector *p* will produce a value close to 1. In the final layer, the vector *a* is multiplied with weight matrix *M* of size KxQ, therefore producing output vector *d* of size Kx1. The competitive function selects the highest value and determines the class label.

In general, a PNN for M classes is defined as the following (Foody, 2001; X. Hong, 2009):

$$y_j(x) = \frac{1}{n_j} \sum_{i=1}^{n_j} e^{\left(-\frac{(\|x_{j,i} - x\|)^2}{2\sigma^2} \right)}$$

where j = 1,…, M and n_j is the number of data points in class *j*.

A decision boundary is found by finding the numerical solution to the above for each class. For instance, for a two-class problem this is done by equating $y_1(x)$ to $y_2(x)$ and finding solution using grid search (X. Hong, 2009).

5.1.2 Learning Vector Quantisation (LVQ)

Vector quantisation aims to find prototypes or representatives of the input data that provides a good approximation of the original input space (Bullinaria, 2012; Ham and Kostanic, 2001). These codebook vectors are used to classify unseen vectors. LVQ is a supervised version of vector quantisation. Procedure wise, initially a random set of vectors are trained to be the representatives. LVQ uses a winner-takes-all strategy, where one or more vectors similar to the given input vectors are selected and adjusted to come closer to the input vector. The error on the distance is determined by the formula:

$$D = \sum_{x} \left\| x - w_{l(x)} \right\|^2$$

where x are the input vectors and $w_{l(x)}$ are the reference or codebook vectors.

This process repeats until the distribution of codebook vectors in the input space approximates the distribution of the samples from the test dataset. This is similar to a self-organising map (SOM), in fact, LVQ embeds SOM in the operation. This is shown in Figure 5.2 (Bullinaria, 2012).

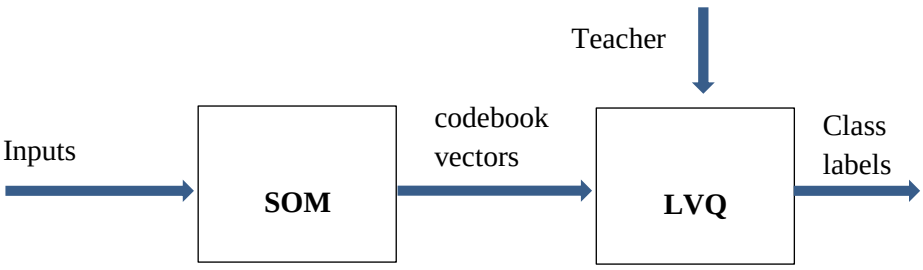


Figure 5.2: A two-stage process involving SOM and LVQ

SOM is inspired by the self-organising capability of neurons in the visual cortex, and provides a topological feature mapping of the input space to the output space. LVQ provides a way to shift cell boundaries for better classification (Bullinaria, 2012). It compares the input classes against the classification label for each weight as provided by SOM. If x and $w_{l(x)}$ have the same class label, the distance between them is shortened. This is governed by the equation

$$\Delta w_{l(x)}(t) = \beta(t)(x - w_{l(x)}(t))$$

where β is the learning rate, and should decrease with the number of iterations. However, for difference in class label they are moved apart by $\Delta w_{l(x)}(t) = -\beta(t)(x - w_{l(x)}(t))$. The weights for other input regions will remain unchanged. In this manner, the winner will eventually be reinforced while others are reduced.

5.1.3 Support Vector Machine (SVM)

SVM is a technique for data classification, where data is non-linearly mapped into a higher dimensional space and a separating hyper plane with maximal margin is found (Cortes and Vapnik, 1995; Chen and Lin, 2005).

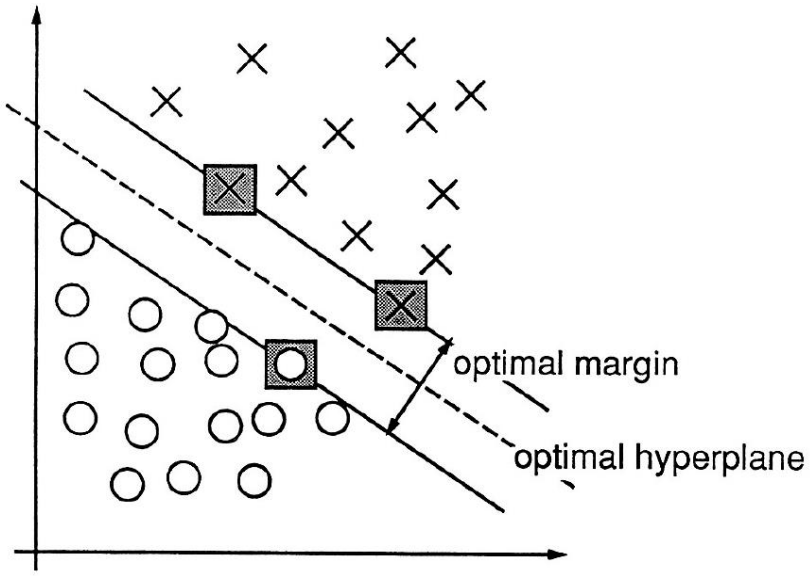


Figure 5.3: Example of separating hyper plane in a higher dimensional plane, showing support vectors on the optimal margin (Cortes and Vapnik, 1995)

The separating hyper plane is determined by an orthogonal vector *w* and bias *b* that satisfy the equation $w.x + b = 0$ and constrained by $\min_i |w.x_i + b| \geq 1$. Since a separating hyperplane in canonical form must satisfy the constraints

$$y_i(\boldsymbol{w}.x_i + b) \geq 1, \qquad i = 1, 2, .., n$$

then the hyper plane that optimally separates the data must minimise

$$\Phi(w) = \frac{1}{2}(\boldsymbol{w}.\boldsymbol{w})$$

After introducing a slack variable $\xi_i \geq 0, i = 1, 2, .., n$ the constraint becomes

$$y_i(w.x_i + b) \geq 1 - \xi_i, \qquad i = 1, 2, \ldots, n$$

and the optimisation problem becomes

$$\Phi(w) = \frac{1}{2}(\boldsymbol{w}.\boldsymbol{w}) + C\sum_{i=1}^{n}\xi_i$$

where C is a user defined positive finite constant.

An optimal hyper plane is constructed using support vectors, which is a small subset of the training vectors. The optimal hyper plane is defined as the linear decision function with maximal margin (Cortes and Vapnik, 1995). According to Kuhn-Tucker optimisation theory, the optimal solution satisfies

$$\alpha_i[y_i(w.x_i + b) - 1] = 0, \qquad i = 1, 2, \ldots, n$$

and contains non-zero Lagrange multipliers if the points x_i (support vectors) satisfy

$$y_i(w.x_i + b) = 1, \qquad i = 1, 2, .., n$$

If the training vectors are separated without error by this optimal hyper plane, the expectation value of the probability of committing an error on a test example is given by

$$E[Pr(error) \leq \frac{E[number\ of\ support\ vectors]}{number\ of\ training\ vectors}$$

This suggests that if the optimal hyper plane can be constructed from a small number of support vectors relative to the size of the training set, then the generalisation ability is high (Cortes and Vapnik, 1995). In the thesis, there are two parameters which need to be optimised before testing is done. The pair is the penalty term, C and the kernel function parameter gamma, γ. They are selected through grid search so that the classifier can predict unknown data (Hsu, Chang and Lin, 2003).

5.1.3.1 Optimisation of C and γ

Optimisation is done through selecting a range of values from graphs. One can:

- Fix the value of C, determine mean squared error (MSE) while varying γ, and
- Fix value of γ, and determine MSE while varying C

The technique was applied to *E. 4-pustulatus* and *C. 14-guttata* where spot colour was used as feature, as depicted in Figures 5.4 and 5.5. Using poly kernel, test according to the above scheme was conducted and graphs are plotted:

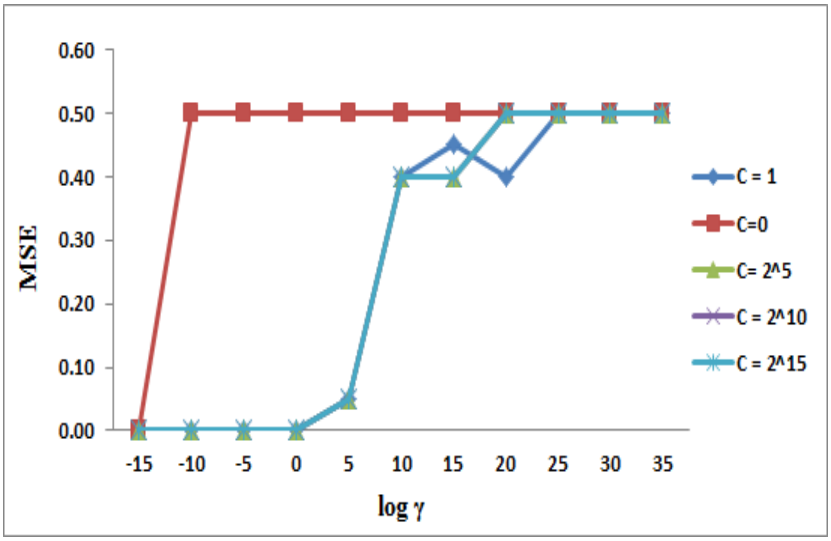


Figure 5.4: MSE vs. γ for various C (E4C14 spot colour data)

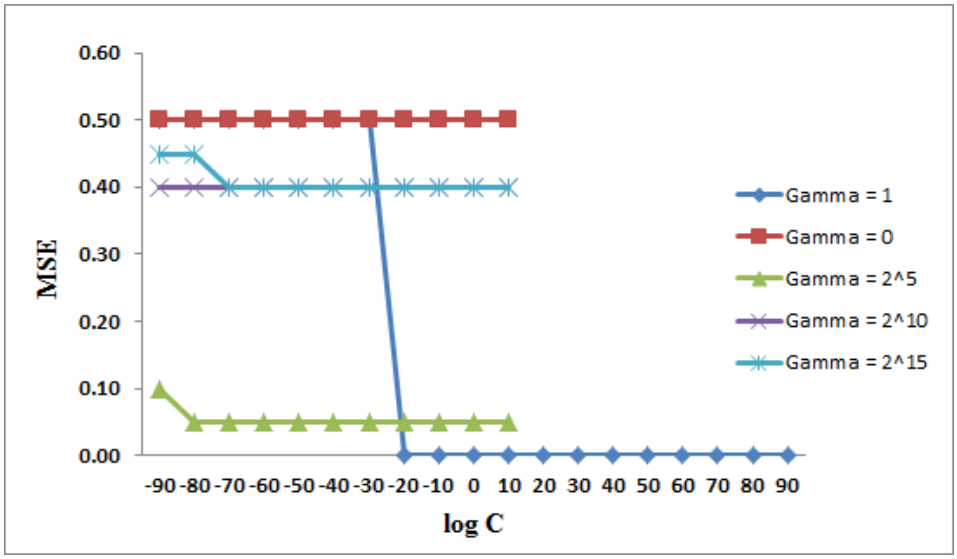


Figure 5.5: MSE vs. C for various γ (E4C14 spot colour data)

The same procedure was applied to elytra colour used as feature, as shown in Figures 5.6 and 5.7.

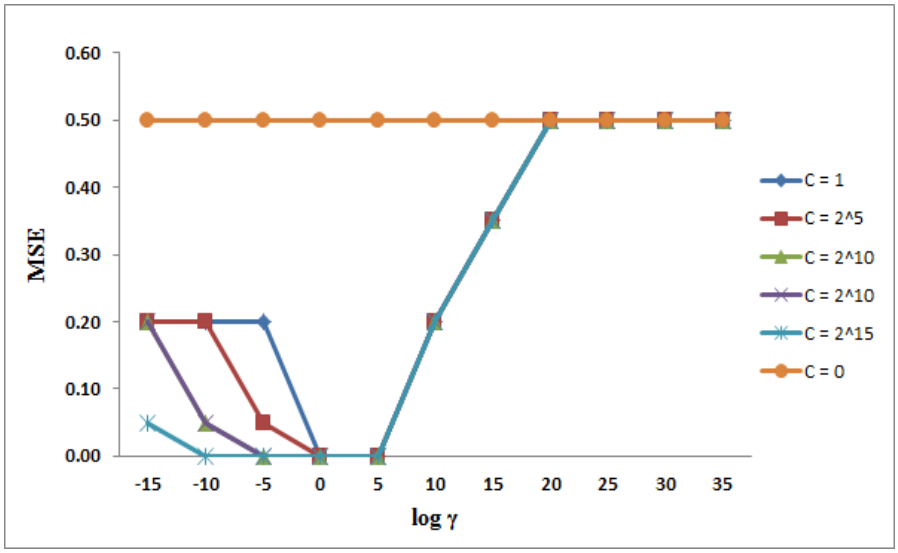


Figure 5.6: MSE vs. γ for various C (E4C14 elytra colour data)

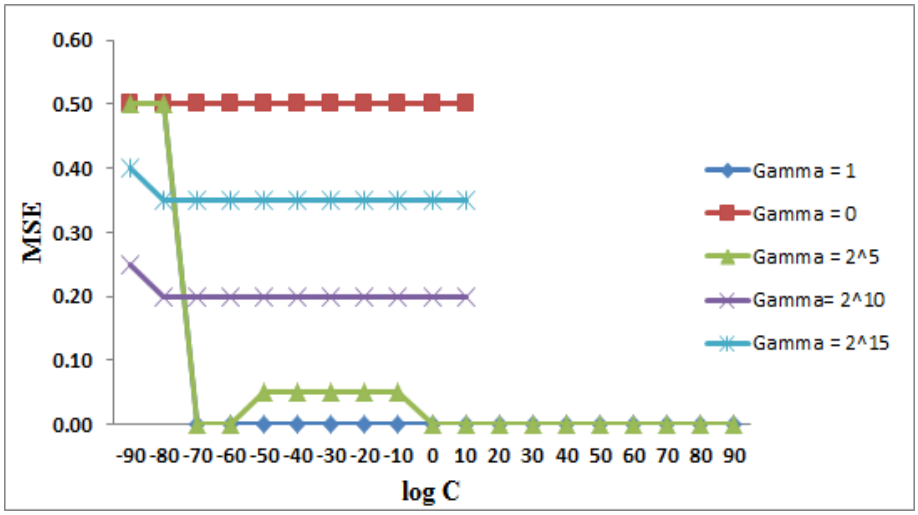


Figure 5.7: MSE vs. C for various γ (E4C14 elytra colour data)

Based on the graphs, a recommended value to use for (C,γ) is (1,1). This value gives the lowest MSE. Next, using these values SVM is tested for the rest of the ladybird datasets.

5.2 Datasets

There were 40 samples per species, and a total of 360 samples were obtained. Dataset partitioning is performed by dividing a set of data into training and test data, and data is selected at random using a MATLAB function 'dividerand()'. The objective of dataset partitioning is to measure network performance. The training data is used to produce a model for training, while the test data is used to check the network's ability to generalise on other inputs not been used during training (Prechelt, 1998). In this study, 85% of the data is used as training set and the remaining 15% used as the test set (MathWorks, 2012). The training data is further partitioned into training set and validation set. The training set is used to adjust network weights during training. The training set comprises 70% of the total dataset. The validation set is to minimise any bias during performance measurement, to check if training is completed subject to stopping criteria therefore preventing over fitting during training (Prechelt, 1998, Nikolaou, 2010; Clark, 2012). The validation set makes 15% of the total dataset. It should be highlighted that the partitioning of the datasets applies to experiments involving MLP neural networks only, whereas 10-fold cross-validation was applied for works involving other classifiers.

The MLP network was trained for each colour group using the MATLAB R2010 Neural Network Toolbox. During training, the network was fed with training input sequentially in batches, where the parameters are updated when the whole training set is completely presented. Initially a total of 12 hidden neurons were used using

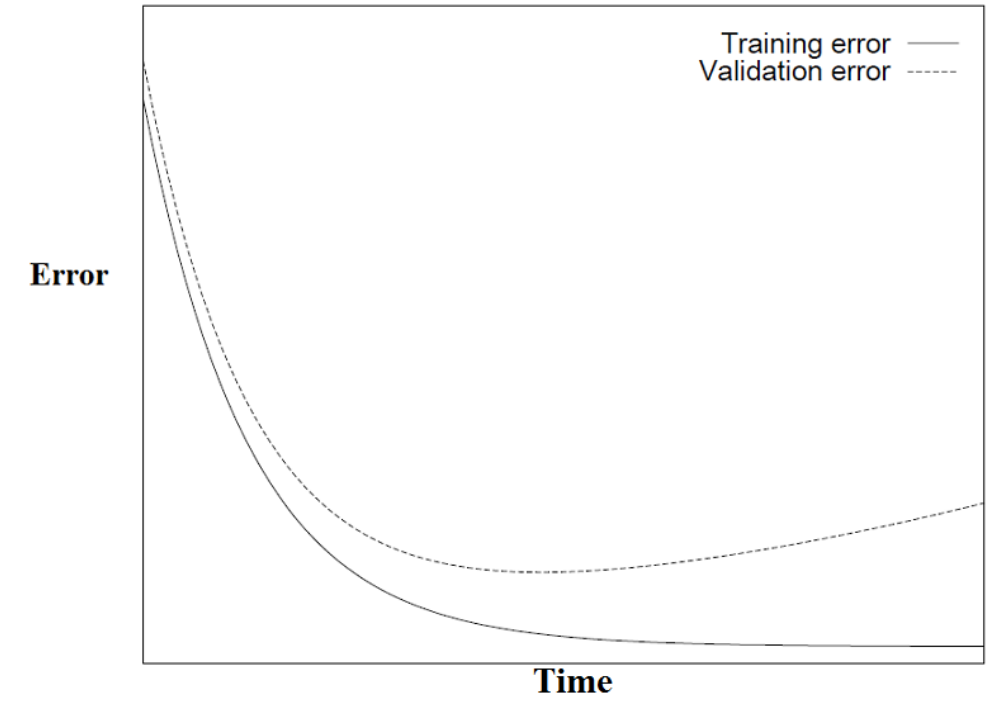
only one hidden layer. The figure was obtained by experiments, following testimonials by researchers such as Looney and Boddy et al. (Looney, 1997; Boddy, 1994). Other parameters include momentum and learning rate, which were finally set to 0.2 and 0.3 respectively. These values are those that gave the best results after a number of trials using validation set.

Momentum is the parameter used to smooth the trajectories for convergence. When back propagation is used in a MLP, it will try to converge to a solution but slowly. This is due to the change in curvature of the squared error surface over the path of the trajectory (Hagan, Demuth and Beale, 2002). To improve the speed of convergence, the learning rate will need to be increased. However, it can cause the trajectories to produce many local minima and therefore become stuck. To avoid this situation, the momentum is adjusted so as to give a smooth transition while training is taking place. This is analogous to implementing a low-pass filter to smooth out oscillations (Hagan, Demuth and Beale, 2002).

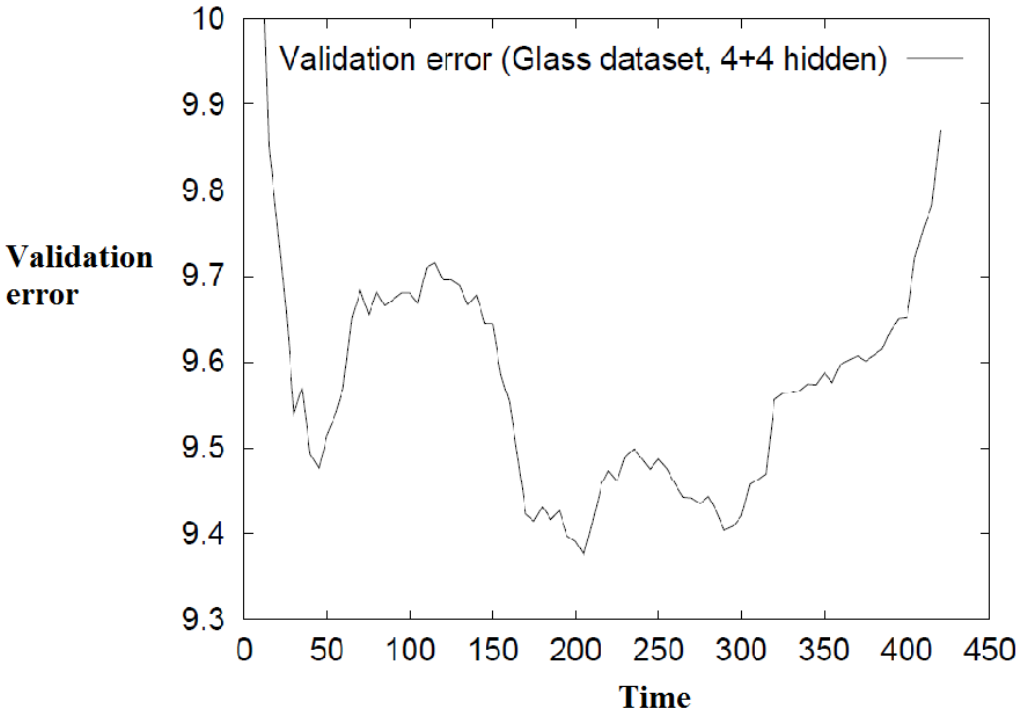
The initial weights and biases were chosen to be small random values. The neural network was trained for 1500 epochs at the maximum, or when the early stopping condition is satisfied. This is done by applying the value of 1.0 when using the Levenberg-Marquardt network training function ‘trainlm’ (MathWorks, 2012). Early stopping was introduced as it is widely used and easier to implement (Prechelt, 1998).

5.2.1 Over fitting

Over fitting is a situation which occurs during training when a classifier seems to memorise training data instead of learning from it. Over fitting needs to be avoided because the error on unseen training examples increases while the training error reduces (Geman, Bienenstock & Doursat, 1992; Prechelt, 1998). This means the generalisation curve is getting worse. Figure 5.8 illustrates the over fitting situation. It shows that both the training error and validation error curves reduce against training epochs. In short, over fitting occurs when validation error starts to increase (Prechelt, 1998).



(a)



(b)

Figure 5.8: (a) Ideal training and generalisation curves, (b) Example of validation error for glass dataset (Prechelt, 1998)

5.2.2 Cross validation

Cross validation is a technique commonly used to compare models, to estimate accuracy of a classifier and to avoid over fitting (Wolpert, 1992; Kohavi, 1995; Schneider, 1997). Cross validation is performed by partitioning into training and test sets, and the simplest version is called the holdout method (Schneider, 1997). After partitioning, a classifier is trained using the training set and then tested using the test set which contains examples which have not appeared in the training set. By doing so, the generalisation of a trained classifier is assessed against an independent dataset.

A variant of cross validation is called K-fold cross validation, where the dataset is partitioned into K equal-sized folds and the holdout method is repeated K times (Kohavi, 1995). For each run, one of the folds is used as the test set and the (K-1) remaining folds used as the training. Each of the K folds will be used once as validation data. After K runs, the average cross validation error across all runs is computed. The error is an estimate of how the classifier would perform if the data collected is an accurate representation of the real world (Weiss, 2011). In this thesis 10-fold cross validation was used for works involving J48 decision trees, SVM, LVQ and PNN so that their results can be compared (Shri and Sriraam, 2012).

5.2.3 Balanced and Unbalanced set

It will be useful to observe and analyse the effect of changing the proportion of samples used. For instance, given the identification of two arbitrary species, A and B, there is a need to check whether a bias towards the count of samples in a particular class affects the outcome of the test. Species A may use only one-third of

its total samples, whereas species B uses all of its samples, assuming initially both have equal number of samples. For this purpose, an experiment has been setup for the identification of a British ladybird species, *E. 4-pustulatus* and an invasive species, *H. axyridis* f. *spectabilis*. The analysis involves MLP applying back propagation and J48 decision tree, although the same method can also be applied to SVM, LVQ and PNN.

5.3 Summary

The chapter has introduced the concept of classifiers, by explaining how PNN, LVQ and SVM work in relation to the ladybird automated identification system. It then explains the process of partitioning the input data into three sets (training, validation and test sets) to avoid over fitting, the proportions and the use of cross validation. Finally, the use of balanced and unbalanced sets in the identification of *E. 4-pustulatus* and *H. axyridis* f. *spectabilis* has been investigated. The results of MLP, LVQ, PNN and SVM are presented in Chapter 6.

CHAPTER 6

IDENTIFICATION RESULTS

CHAPTER 6

IDENTIFICATION RESULTS

This chapter covers identification methods covered in literature reviewed in section 2.2, and will explicitly show the identification results of UK ladybirds, including *H. axyridis*. It will also show a novel method of a hybrid system consisting of ANN and decision tree.

6.1 Classifiers and Confusion Matrix

For the identification of biological species, results are typically presented in terms of contingency table, or better known as confusion matrix. The confusion matrix shows the dispositions of the set of instances in a matrix form. Suppose an identification system involves only two classes, where each has its associated class labels. If a neural network is used the system needs training exemplars and binary targets. The training exemplars are fed into a network or classifier, trained to learn some learning functions and come out with some outputs based on decision boundaries. There will be four possible outcomes in this case because it is a binary case. Figure 6.1 shows a typical confusion matrix for a binary case.

Table 6.1: Confusion matrix

	y	*n*
Y	TP	FP
N	FN	TN

The counts of true positive outcomes are labeled as TP, and counts of true negative outcomes are labeled as TN as shown in the diagonals of the confusion matrix. There could also be instances where false positives and false negatives are obtained; these are labeled as FP and FN respectively. A false positive is a negative instance that is classified as a positive (false alarm), whereas a false negative is a positive instance that is counted as a negative. For a perfect confusion matrix, these off-diagonal values FP and FN need to be zero. The total numbers of positives are in column *y*, and total numbers of negatives are given in column *n*. Hence, true positive rates (also called sensitivity, or recall) are calculated as the ratio between the number of true positives and the total number of positives, (TP/*y*). True negative rates are calculated as the ratio between the number of true negatives and the total number of negatives, (TN/*n*). The reader is referred to some extension of the metrics derived from the confusion matrix in Bradley's and Fawcett's work (Bradley, 1997; Fawcett, 2006; Omid, 2011). The metrics are:

$$Sensitivity = Recall = TP\ rate = \frac{TP}{TP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Specificity = \frac{TN}{FP + TN}$$

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

6.1.2 Methodology of Classifier

The 12 features extracted from an image containing an unknown ladybird, as explained in previous chapter, will be fed into a classifier. These feature instances are joined together with feature instances owned by, for example, a group of 99 Harlequin ladybirds. A scheme is shown in Figure 6.1 with an aim to perform pre-sorting between Harlequins and non-Harlequins.

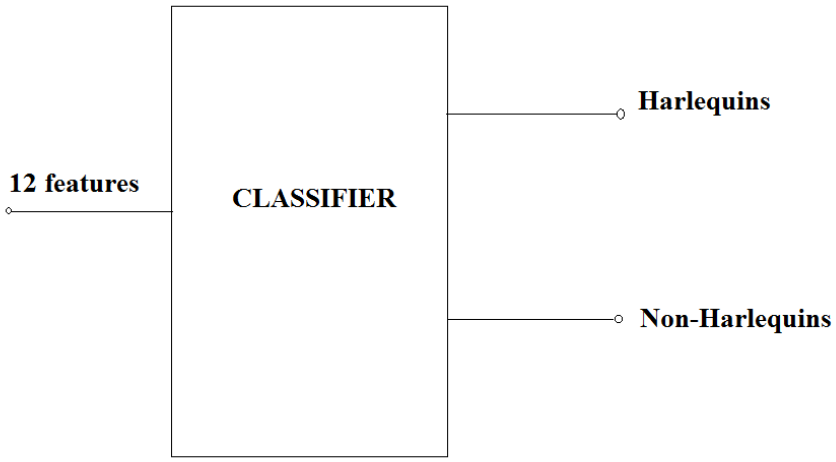


Figure 6.1: Training scheme for sorting Harlequins from non-Harlequins

Harlequins are labeled H, while non-Harlequins as N. Ideally, test is done after training and the confusion matrix will show up as in Table 6.2, where the diagonal values sum up to equal the total number of instances, which means no misclassification occur.

Table 6.2: Perfect accuracy confusion matrix

	H	**N**
H	99	0
N	0	1

Misclassification arises, for instance, when N is misclassified as H, as shown in Figures 6.3-6.5. Table 6.3 shows that N is misclassified as H. Table 6.4 shows another example where some non-diagonal values occurring (labeled as X and Y),

whilst there is no N been classified correctly. Table 6.5 shows another confusion matrix, though unlikely to occur due to higher sample numbers, where H is not correctly identified. Note that X and Y may not be proportionately balanced.

Table 6.3: Confusion matrix showing misclassification (example 1)

	H	**N**
H	100	0
N	0	0

Table 6.4: Confusion matrix showing misclassification (example 2)

	H	**N**
H	100-(X+Y)	X
N	Y	0

Table 6.5: Confusion matrix showing misclassification (example 3)

	H	**N**
H	0	X
N	Y	1

The confusion matrices above show how much accuracy, therefore indicating the confidence level of correct classification. Confidence level will be revisited in chapter 7, when analysis of the integrated system is performed. For now, it suffices to mention that confidence level helps during identification.

6.2 Classifier training and test

6.2.1 Training and test setup

The tests are categorized into three different groups; namely white, black and red groups. The groups have been named based on the ladybirds' spot colour. Readers can refer to Table 6.6 for the groups and the corresponding acronyms.

Table 6.6: Ladybird acronyms in bold, arranged in groups according to their spot colours

	Groups		
	White	**Red**	**Black**
Species	• ***C14*** • ***H16***	• ***E4*** • ***H1*** • ***H2***	• ***A2*** • ***C5*** • ***C7*** • ***H3***

Results are presented in the form of confusion matrix to show level of accuracy for the tests. Detailed identification metrics such as TP rates, FP rates are provided in Appendix IV.

6.2.2 MLP training and test groups

The objectives were:

- To train a MLP neural network using backpropagation algorithm.
- To determine the identification results.
- To evaluate the contributions of feature sets on the identification accuracy.

6.2.2.1 Test 1: White set

Tables 6.7 a, b & c show the resultant confusion matrix for test on all features, colour features and geometrical features respectively. Table 6.8 lists relevant metrics obtained from all the tests.

Table 6.7a: Confusion matrix for test on White set (all features)

	C14	**H16**
C14	8	0
H16	0	4

Table 6.7b: Confusion matrix for test on White set (colour features)

	C14	**H16**
C14	6	0
H16	2	4

Table 6.7c: Confusion matrix for test on White set (geometrical features)

	C14	H16
C14	8	0
H16	0	4

Table 6.8: Summary of test results for the three feature sets

Test set	True positive rate (%)	False positive rate (%)	Sensitivity (%)
White All	100	0	100
White Colour	75	0	75
White Geo	100	0	100
Test	**Precision (%)**	**Specificity (%)**	**Accuracy (%)**
White All	100	100	100
White Colour	100	100	83.3
White Geo	100	100	100

In this work, the author is more concerned about the true positives and false positives rather than the other detailed metrics, mainly because they show how much accuracy is obtained for the identification system (Ayob and Chesmore, 2012). True positive rates and false positive rates are columnar ratio and independent of class distributions (Fawcett, 2006). For instance, consider the identification results of *C. 14-guttata* and *H. 16-guttata* in Tables 6.7 and 6.8, which show test result for using backpropagation algorithm in a MLP neural network. Using colour features only made TP rate dropped to 75% from 100%, hence accuracy was reduced to 83.3% due to the two false negatives. When geometrical features were used next all results were 100%, with the exception of false positive rate. This may suggest that colour features have not been useful for the identification of *C. 14-guttata* and *H. 16-guttata* using MLP with backpropagation algorithm.

6.2.2.2 Test 2: Red set

The test covers the identification of ladybird species known to have reddish spots. Following the same training process as those in white-spotted group, test results were obtained as given in Table 6.9a, Table 6.9b and Table 6.9c.

Table 6.9a: Confusion matrix for test on Red set (all features)

	E4	H1	H2
E4	5	0	0
H1	0	7	0
H2	0	6	0

Table 6.9b Confusion matrix for test on Red set (colour features)

	E4	H1	H2
E4	0	0	0
H1	0	1	0
H2	3	14	0

Table 6.9c: Confusion matrix for test on Red set (geometrical features)

	E4	H1	H2
E4	3	1	0
H1	0	13	0
H2	0	1	0

For the test on Red group, there were 6 instances where *H. axyridis* f. *conspicua* misclassified as *H. axyridis* f. *spectabilis* when all features were used, giving 66.7% accuracy only. When colour features were used, there was only 1 instance where *H. axyridis* f. *spectabilis* was identified correctly giving only 5.6% accuracy. Three instances were misidentified as *E. 4-pustulatus*, and 14 instances misidentified as *H. axyridis* f. *spectabilis*. From Table 6.9c where geometrical features were used, 3 correct identifications for *E. 4-pustulatus* and an instance when it was incorrectly identified as *H. axyridis* f. *spectabilis*. An abundance of 13 correct identifications were registered for *H. axyridis* f. *spectabilis*, while there was an instance where *H. axyridis* f. *conspicua* was misidentified as *H. axyridis* f. *spectabilis*. The accuracy was 88.9%.

6.2.2.3 Test 3: Black set

Test results were obtained as given in Table 6.10a, b and c.

Table 6.10a: Confusion matrix for test on Black set (all features)

	A2	C5	C7	H3
A2	7	0	0	1
C5	0	4	0	2
C7	0	0	6	0
H3	0	1	0	3

Table 6.10b: Confusion matrix for test on Black set (colour features)

	A2	C5	C7	H3
A2	7	1	0	1
C5	0	8	0	0
C7	0	0	2	1
H3	1	0	0	3

Table 6.10c: Confusion matrix for test on Black set (geometrical features)

	A2	C5	C7	H3
A2	8	0	0	1
C5	0	9	1	3
C7	0	0	0	1
H3	0	0	1	0

For the test on Black group, there was one instance where *A. 2-punctata* was misidentified as *H. axyridis* f. *succinea* when all features were used. There were 4 correct identification for *C. 5-punctata*, while 2 instances when it was misidentified as *H. axyridis* f. *succinea*. 6 instances of correct identification was registered for *C. 7-punctata*, while 3 correct identification for *H. axyridis* f. *succinea* and an instance of misidentification as *C. 5-punctata*. Accuracy was 83.3%.

Table 6.10b shows *A. 2-punctata* was correctly identified for 7 instances, and 1 misidentification as *C. 5-punctata* and *H. axyridis* f. *succinea* respectively. There was an instance of *C. 7-punctata* misidentified as *H. axyridis* f. *succinea*. For *H. axyridis* f. *succinea* there was an instance where it has been misidentified as *A. 2-punctata*. Accuracy was 83.3%.

When geometrical features were used, there was an instance where *A. 2-punctata* misidentified as *H. axyridis* f. *succinea*. *C. 5-punctata* was misidentified as *C. 7-punctata* for 1 instance, and 3 misidentification as *H. axyridis* f. *succinea*. There was no correct identification for *C. 7-punctata* and there was an instance where it was misidentified as *H. axyridis* f. *succinea*. Similar situation for *H. axyridis* f. *succinea* where it was misidentified as *C. 7-punctata* once. Accuracy for the test on geometrical feature was 70.8%.

6.2.3 Tests using SVM

SVM using Sequential Minimal Optimisation (SMO) algorithm has been tested on the three colour groups as previously listed in Table 5.1, using both balanced class and unbalanced class. The followings are the datasets and results of identification for unbalanced class:

White-spotted group

The results for unbalanced dataset in the white-spotted group are given in Tables 6.11a, 6.11b and 6.11c.

Table 6.11a: Confusion matrix for SVM using SMO (C14H16 white group, unbalanced class, all features)

	C14	**H16**
C14	9	0
H16	1	40

Table 6.11b: Confusion matrix for SVM using SMO (C14H16 white group, unbalanced class, colour features)

	C14	H16
C14	6	0
H16	4	40

Table 6.11c: Confusion matrix for SVM using SMO (C14H16 white group, unbalanced class, geometrical features)

	C14	H16
C14	10	0
H16	0	40

Balanced class

Table 6.12a: Confusion matrix for SVM using SMO (C14H16 white group, balanced class, all features)

	C14	H16
C14	40	0
H16	0	40

Table 6.12b: Confusion matrix for SVM using SMO (C14H16 white group, balanced class, colour features)

	C14	H16
C14	32	0
H16	8	40

Table 6.12c: Confusion matrix for SVM using SMO (C14H16 white group, balanced class, geometrical features)

	C14	H16
C14	40	0
H16	0	40

The unbalanced dataset result in Table 6.11a has 98% accuracy. There was 1 miss where *C. 14-guttata* was misidentified as *H. 16-guttata*. Using colour features only shows a reduced accuracy of 92% compared to using all features. Using geometrical features only shows perfect accuracy.

For the balanced dataset, obvious improvements on all identification metrics can be seen on Table 6.12a whereby using all features and geometrical features gave perfect class match. For the set which used colour features only, a slight reduction from 92% when using unbalanced set to 90% accuracy when using a balanced set was obtained.

Red-spotted group: Balanced class

Table 6.13a: Confusion matrix for SVM using SMO (E4H1H2 red group, balanced class, all features)

	E4	H1	H2
E4	40	2	2
H1	0	25	11
H2	0	13	27

Table 6.13b: Confusion matrix for SVM using SMO (E4H1H2 red group, balanced class, colour features)

	E4	H1	H2
E4	28	10	2
H1	12	16	6
H2	0	14	32

Table 6.13c: Confusion matrix for SVM using SMO (E4H1H2 red group, balanced class, geometrical features)

	E4	H1	H2
E4	40	5	4
H1	0	28	25
H2	0	7	11

For the results of Table 6.13a, the accuracy is 76.7%, where *E. 4-pustulatus* was correctly identified. Only 25 instances of *H. axyridis* f. *spectabilis* correctly identified, while 11 instances misidentified as *H. axyridis* f. *conspicua.* Similarly for *H. axyridis* f. *conspicua*, there were 27 instances correctly identified but 13 misidentified as *H. axyridis* f. *spectabilis*. For results using colour features in Table 6.13b, misidentification occurs on all species and accuracy dropped to only 63.3%. Prime suspects are elytra colour and spot colour components. This can only be confirmed with tests involving MLP and J48, as results obtained through SVM using SMO algorithm does not visually reveal features to be inspected as good as a MLP and decision tree. For Tables 6.13a and 6.13c, suspicion arises on why misidentification occurs within only the two *H. axyridis* forms, and not on *E. 4-pustulatus*.

Table 6.14a: Confusion matrix for SVM using SMO (A2C5C7H3 black group, balanced class, all features)

	A2	C5	C7	H3
A2	40	4	0	10
C5	0	32	4	17
C7	0	4	36	2
H3	0	0	0	11

Table 6.14b: Confusion matrix for SVM using SMO (A2C5C7H3 black group, balanced class, colour features)

	A2	C5	C7	H3
A2	24	0	0	11
C5	8	32	0	10
C7	8	8	40	8
H3	0	0	0	11

Table 6.14c: Confusion matrix for SVM using SMO (A2C5C7H3 black group, balanced class, geometrical features)

	A2	C5	C7	H3
A2	39	5	0	11
C5	0	12	12	13
C7	0	19	24	10
H3	1	4	4	6

The identification accuracies for results in Tables 6.14a, 6.14b and 6.14c are 74.4%, 66.9% and 50.6% respectively. The figures are lower than White and Red groups, which is to be expected as the numbers of classes in the dataset grow. Interestingly observations on the result show that the majority of misidentifications revolve around the two species, *C. 5-punctata* and *H. axyridis* f. *succinea*. This has been a striking observation; however, no clear conclusion can be drawn on their relationship as more tests are needed.

6.2.4 Tests using Learning Vector Quantisation (LVQ)

Using 10-fold cross validation, the following results were obtained for each group:

White-spotted group

The results for unbalanced class distribution are presented first in Tables 6.15a-6.15c.

Table 6.15a: Confusion matrix for test using LVQ (C14H16 white group, unbalanced, all features)

	C14	H16
C14	9	0
H16	1	40

Table 6.15b: Confusion matrix for test using LVQ (C14H16 white group, unbalanced, colour features)

	C14	H16
C14	7	0
H16	3	40

Table 6.15c: Confusion matrix for test using LVQ (C14H16 white group, unbalanced, geometrical features)

	C14	H16
C14	10	0
H16	0	40

White-spotted group: Balanced class

Table 6.16a: Confusion matrix for test using LVQ (C14H16 white group, balanced, all features)

	C14	H16
C14	40	0
H16	0	40

Table 6.16b: Confusion matrix for test using LVQ (C14H16 white group, balanced, colour features)

	C14	H16
C14	34	2
H16	6	38

Table 6.16c: Confusion matrix for test using LVQ (C14H16 white group, balanced, geometrical features)

	C14	H16
C14	40	0
H16	0	40

Red-spotted group: Balanced class

Table 6.17a: Confusion matrix for test using LVQ (E4H1H2 red group, balanced, all features)

	E4	H1	H2
E4	40	3	0
H1	0	25	12
H2	0	12	28

Table 6.17b: Confusion matrix for test using LVQ (E4H1H2 red group, balanced, colour features)

	E4	H1	H2
E4	39	7	3
H1	0	25	10
H2	1	8	27

Table 6.17c: Confusion matrix for test using LVQ (E4H1H2 red group, balanced, geometrical features)

	E4	H1	H2
E4	40	3	1
H1	0	23	20
H2	0	14	19

Black-spotted group

Table 6.18a: Confusion matrix for test using LVQ (A2C5C7H3 black group, balanced, all features)

	A2	C5	C7	H3
A2	40	0	0	6
C5	0	32	7	11
C7	0	2	32	8
H3	0	6	1	15

Table 6.18b: Confusion matrix for test using LVQ (A2C5C7H3 black group, balanced, colour features)

	A2	C5	C7	H3
A2	23	3	0	7
C5	8	32	0	9
C7	5	4	36	3
H3	4	1	4	21

Table 6.18c: Confusion matrix for test using LVQ (A2C5C7H3 black group, balanced, geometrical features)

	A2	C5	C7	H3
A2	40	1	0	5
C5	0	24	8	10
C7	0	7	27	11
H3	0	8	5	14

Accuracies for the unbalanced class distribution dropped to 94% when only colour features have been used. For the balanced group, this observation happened for the White group only. It is interesting to see that accuracies reduced almost linearly for the Red and Black group when identification were performed using all features,

colour and geometrical features one after another. There is no explanation for this phenomenon.

6.2.5 Tests using Probabilistic Neural Network (PNN)

PNN uses normalised Gaussian radial basis functions as a network (Hagan, Demuth and Beale, 2002). Using 10-fold cross validation, the following results were obtained for each group:

White-spotted group

The results using unbalanced class distribution are shown first.

Table 6.19a: Confusion matrix for test using PNN (C14H16 white group, unbalanced, all features)

	C14	H16
C14	9	0
H16	1	40

Table 6.19b: Confusion matrix for test using PNN (C14H16 white group, unbalanced, colour features)

	C14	H16
C14	9	0
H16	1	40

Table 6.19c: Confusion matrix for test using PNN (C14H16 unbalanced, geometrical features)

	C14	H16
C14	9	0
H16	1	40

White-spotted group: Balanced

Table 6.20a: Confusion matrix for test using PNN (C14H16 white group, balanced, all features)

	C14	H16
C14	40	0
H16	0	40

Table 6.20b: Confusion matrix for test using PNN (C14H16 balanced, colour features)

	C14	H16
C14	37	0
H16	3	40

Table 6.20c: Confusion matrix for test using PNN (C14H16 balanced, geometrical features)

	C14	H16
C14	40	0
H16	0	40

Red-spotted group

Table 6.21a: Confusion matrix for test using PNN (E4H1H2 red group, balanced, all features)

	E4	H1	H2
E4	40	1	0
H1	0	28	17
H2	0	11	23

Table 6.21b: Confusion matrix for test using PNN (E4H1H2 red group, balanced, colour features)

	E4	H1	H2
E4	34	7	4
H1	6	25	9
H2	0	8	27

Table 6.21c: Confusion matrix for test using PNN (E4H1H2 red group, balanced, geometrical features)

	E4	H1	H2
E4	40	2	1
H1	0	25	27
H2	0	13	12

Black-spotted group

Table 6.22a: Confusion matrix for test using PNN (A2C5C7H3 black group, balanced, all features)

	A2	C5	C7	H3
A2	39	0	0	0
C5	0	34	0	10
C7	0	0	39	0
H3	1	6	1	30

Table 6.22b: Confusion matrix for test using PNN (A2C5C7H3 black group, balanced, colour features)

	A2	C5	C7	H3
A2	30	0	0	5
C5	0	38	0	6
C7	2	0	38	0
H3	8	2	2	29

Table 6.22c: Confusion matrix for test using PNN (A2C5C7H3 black group, balanced, geometrical features)

	A2	C5	C7	H3
A2	40	0	0	0
C5	0	27	4	12
C7	0	5	32	7
H3	0	8	4	21

6.3 Analysis

For a classification system to be useful, the results need to be benchmarked with statistical analysis techniques to signify improvements.

6.3.1 Parameter analysis

Figures 6.2 a-b graphically show the variation in average accuracy and model time for White-spotted group when the Minimum Standard Deviation constant (MinStdDev) is adjusted from 0 to 1.

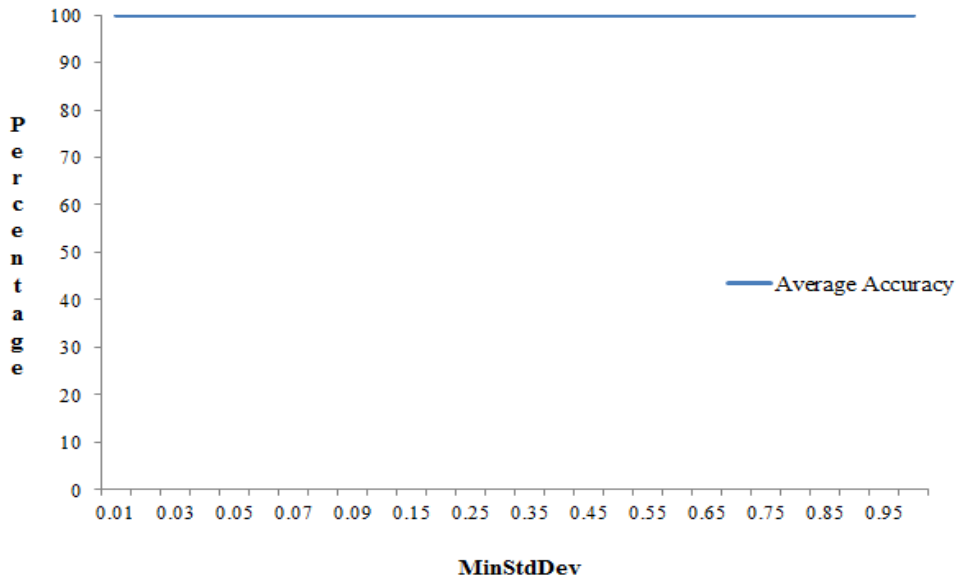


Figure 6.2a: Average Accuracy vs. MinStdDev for White-spotted group

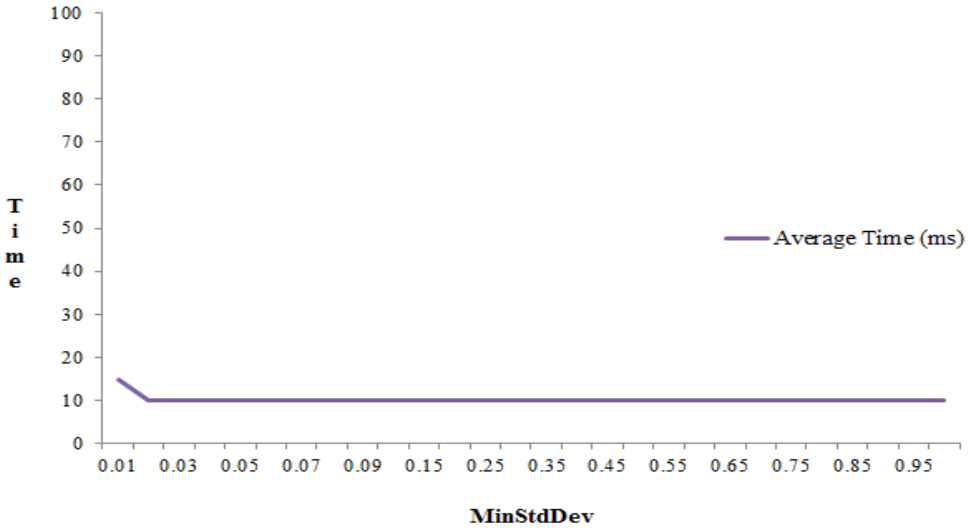


Figure 6.2b: Average Time vs. MinStdDev for White -spotted group

Similarly, Figures 6-3 a-b and Figures 6.4 a-b show the variations for the same quantities in concern for Red-spotted and Black-spotted groups.

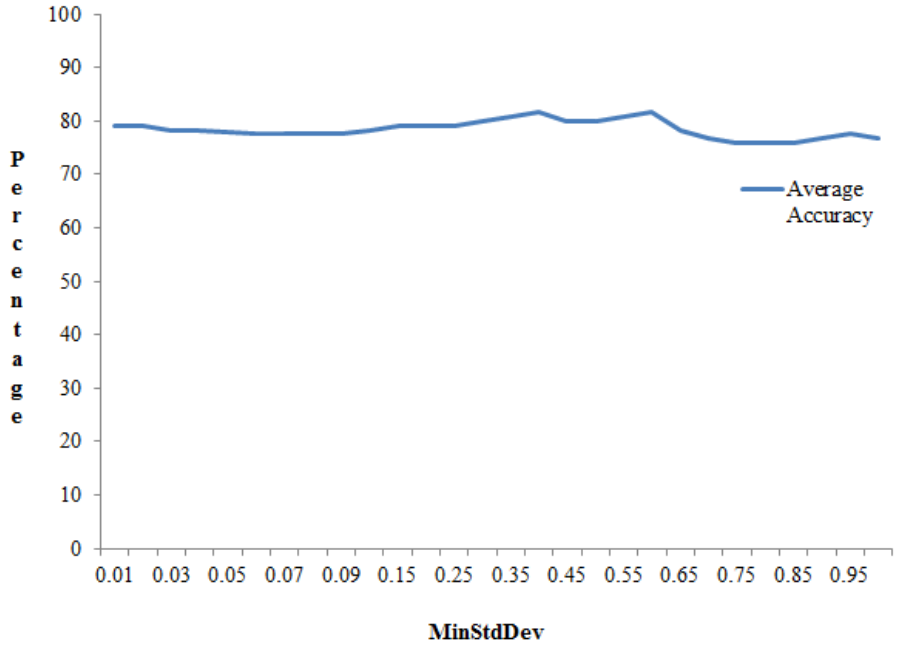


Figure 6.3a: Average Accuracy vs. MinStdDev for Red-spotted group

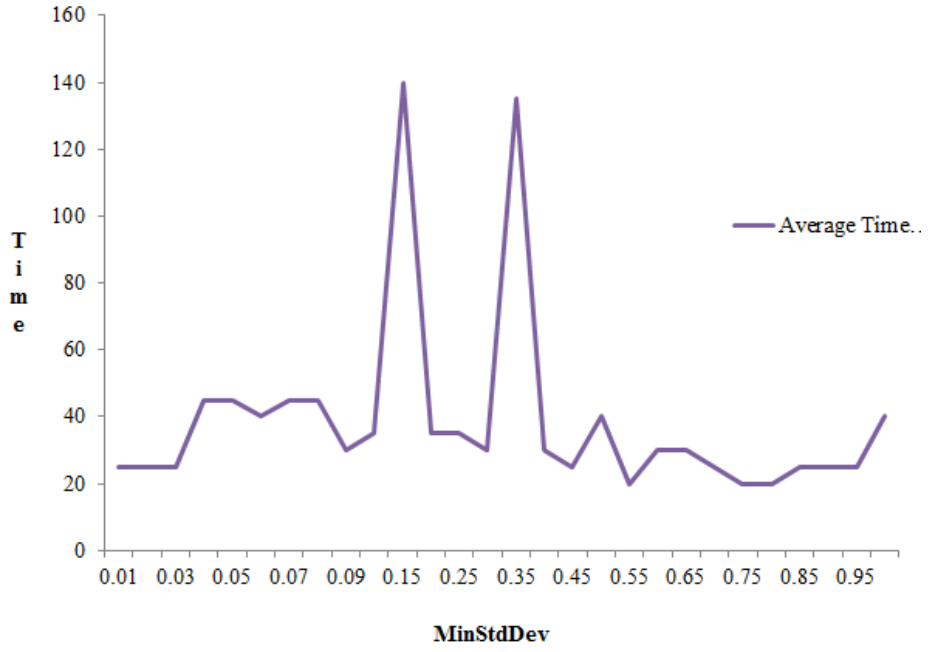


Figure 6.3b: Average Time vs. MinStdDev for Red-spotted group

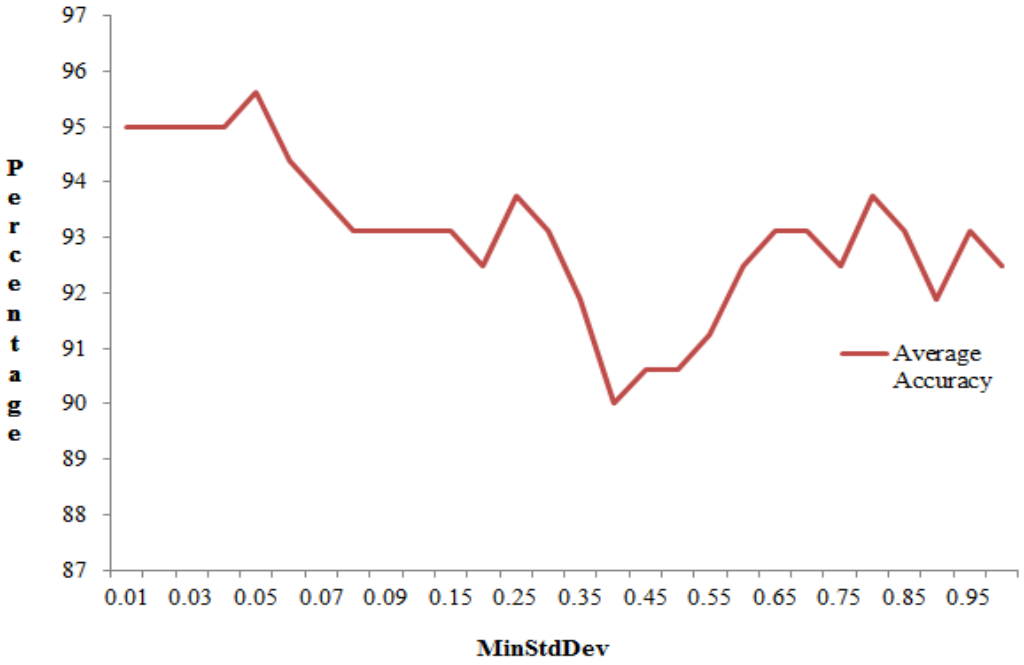


Figure 6.4a: Average Accuracy vs. MinStdDev for Black-spotted group

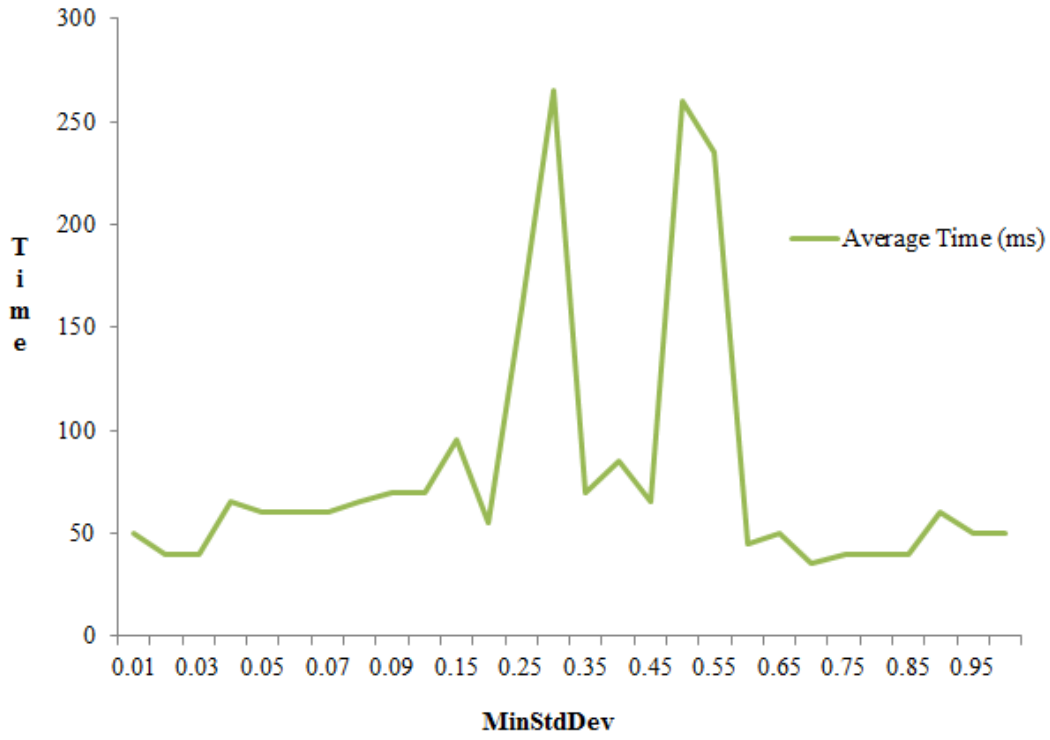


Figure 6.4b: Average Time vs. MinStdDev for Black-spotted group

6.3.2 Test of significance

In order to measure improvements and validate the results, the author used z-test as the test statistics. Here the author assumed the population distribution was a standard normal distribution. In term of the features, selected single-feature has been obtained from J48 decision tree test. For others, more than one feature is obtained. Some examples are given in Table 6.23.

Table 6.23: Features obtained after J48 operations for four species

Features	2-spot	5-spot	7-spot	Pine
Spot area				X
Spot perimeter				
Spot max axis length				X
Spot min axis length				
Spot area ratio				X
Spot aspect ratio				
*Spot colour a**				
*Spot colour b**	X			
Spot hue angle				
*Elytra colour a**		X	X	
*Elytra colour b**				
Elytra hue angle				

The test of significance will first consider the null hypothesis and the alternative hypothesis (Graham, 2010). Null hypothesis is denoted as Ho, while the alternative hypothesis is called H_1. The procedure to carry out the hypothesis test is outlined below.

Step 1: Set up the hypothesis.

Step 2: Calculate test statistic, S.

Step 3: Determine the critical value, C.

Step 4: Check if S is less than, or equal to C.

If this condition is satisfied, reject the alternative hypothesis.

An example calculation on the procedure is explained using the ladybird scenario. Suppose data is obtained from a population consisting of two ladybird species; the

two-spot ladybird (*C. 2-punctata*) and *h. axyridis f. spectabilis*, each of which containing 50 samples. From earlier test using J48 decision tree, it is agreeable that the useful feature is spot colour (b*). The z-test used the mean value of some samples from the population of the feature, which in this case the mean of spot colour (b*) is used. Following the above procedure:

Step 1: Set up the hypothesis.

The alternative hypothesis H_1 states that the mean value of spot colour (b*) is significantly different from the population mean. The null hypothesis Ho will assume otherwise, meaning that there is no significant difference between the spot colour of the two species therefore they are the same species.

Step 2: Calculate test statistic, S.

This figure shows how much standard deviation units the samples are from the mean. 20 random samples are taken from the population. In this case standard deviation, σ, of the population is 0.02376.

Standard Error, SE = σ / sqrt(n) = 0.005313

Test statistic, S = (mean(2-spot) – mean(other)) / SE = mod(-38.3719)

Step 3: Determine the critical value, C.

Use Normal distribution table, a two-tailed test and a 5% level of significance will give approximately C = 2.0.

Step 4: Check if S is less than, or equal to C.

Since S is larger than C, the alternative hypothesis is accepted and the null hypothesis is rejected. On this basis, the samples of ladybirds are significantly different from the expected value i.e. they are not the same ladybird species.

6.4 Comparison of Classifiers Performances

Comparison between classifiers is shown in this section to analyse classifiers performances.

6.4.1 Balanced class distribution

The accuracies for identification of C14H16 between each classifier are shown as a bar graph format in Figure 6.5:

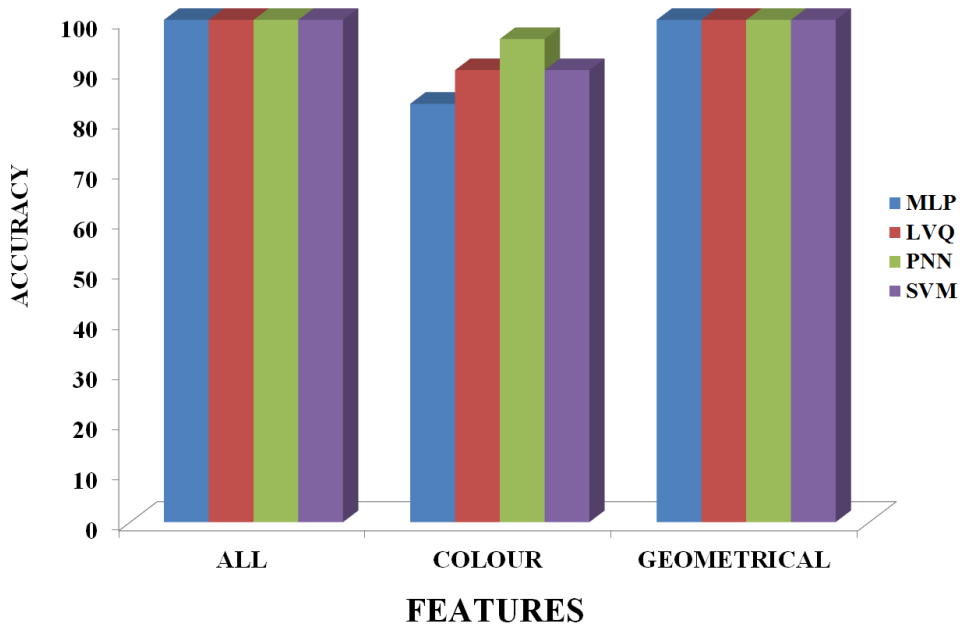


Figure 6.5: Comparison of accuracies between classifiers to identify C14H16 (balanced class distribution)

When using all features and geometrical features, perfect identification was recorded. An obvious observation is the identification rates go lower when colour features were used. This happened to all classifiers. It shows that the use of colour features are insufficient to completely identify the two species correctly, which suggests more useful features to be extracted and utilised by classifiers in addition to colour features. The graph also shows PNN is the best classifier for the identification between C14 and H16 in a balanced class distribution.

6.4.2 Tests using J48 decision tree

The objectives of the tests were:

- To investigate the effect of varying sets of features to the identification results of both balanced and unbalanced datasets
- To determine the best feature sets
- To investigate intra-species variations in *H.axyridis*

The tests involve the following groups of species:

- *E. 4-pustulatus* and *H. axyridis* f. *spectabilis* (E4H1)
- *H. axyridis:* f. *spectabilis,* f. *conspicua* and f. *succinea* (H1H2H3)

10-fold cross validation was implemented for J48 tests.

6.4.2.1 Unbalanced class 1:4

There was a suspicion on the role of the feature set towards the outcome, that it could give way to better results. A simple test was conducted to check this possibility. The decision tree obtained through J48 for the test on *E. 4-pustulatus* and *H. axyridis* f. *spectabilis* was inspected and revealed three features, as shown in Figure 6.6.

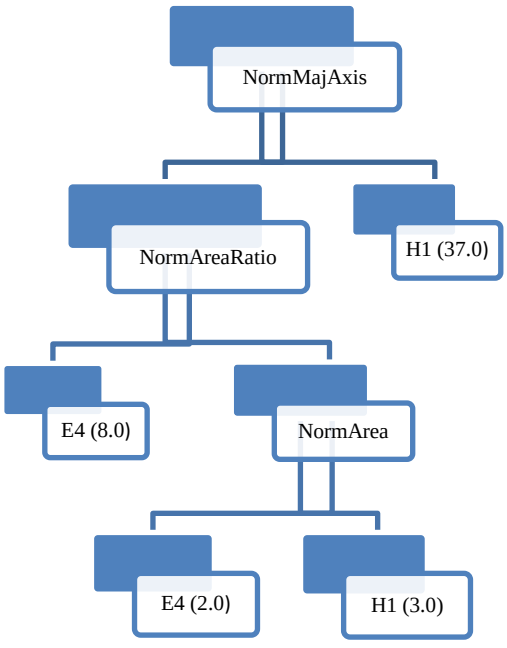


Figure 6.6: Decision tree for the test on *E. 4-pustulatus* and *H. axyridis* f. *spectabilis*

The decision tree indicates that three most important features for the identification of *E. 4-pustulatus* and *H. axyridis* f. *spectabilis* were Major Axis, Area Ratio and Area. These were part of geometrical feature set. A MLP neural network was trained for the unbalanced class, and training results are shown in Table 6.24.

Table 6.24: Training of MLP for unbalanced class of *E. 4-pustulatus* and *H. axyridis* f. *spectabilis*

Run	Q	N	M	J	I	RMS error		% Accuracy	
						MLP	With J48	MLP	With J48
1	50	12	24	2	500	0.45	0.2	80	80
					1000	0.45	0.29	80	92
					1500	0.45	0.29	80	92
2	50	12	22	2	500	0.25	0.45	94	80
					1000	0.2	0.29	96	92
					1500	0.2	0.29	96	92
3	50	12	12	2	500	0.2	0.29	96	92
					1000	0.2	0.32	96	90
					1500	0.2	0.29	96	92
4	50	12	8	2	500	0.2	0.29	96	92
					1000	0.2	0.29	96	92
					1500	0.2	0.29	96	92
5	50	12	4	2	500	0.2	0.29	96	92
					1000	0.2	0.29	96	92
					1500	0.2	0.32	96	90

Note:

Q = no. of exemplar vectors
N = input features
M = no. of hidden neurons
J = no. of output neurons
I = no. of iterations

Based on the training, 24 hidden neurons were selected as it gave the largest margin of improvement. Table 6.25a shows the confusion matrix obtained for the unbalanced class, firstly using MLP with backpropagation algorithm. Initially the outcome was not favourable for *E. 4-pustulatus*, where no true positive was obtained and striking 100% false negative.

Table 6.25a: Confusion matrix for unbalanced class using MLP (all features)

	E4	H1
E4	0	0
H1	10	40

Table 6.25b shows the confusion matrix for the decision tree. It is interesting to note that the weighted average was based on the number of samples per class. This works to a disadvantage for the outnumbered class, in this case it was *E. 4-pustulatus*.

Table 6.25b : Confusion matrix for J48 decision tree (all features)

	E4	**H1**
E4	4	2
H1	6	38

The results for the combination of J48 decision tree and MLP are given in Table 6.25c.

Table 6.25c: Confusion matrix for combination of J48 and MLP (3 features)

	E4	**H1**
E4	9	3
H1	1	37

It is clear that J48 decision tree gave better accuracy i.e. an improvement of 5%, than using MLP with backpropagation algorithm. The two techniques were merged together i.e. the decision tree provided the optimum features and MLP reached a minimised MSE through training it's network using a 'reduced' feature set. This is shown in Figure 6.7, where this technique has greatly improved accuracy to about 12% as compared to using MLP alone.

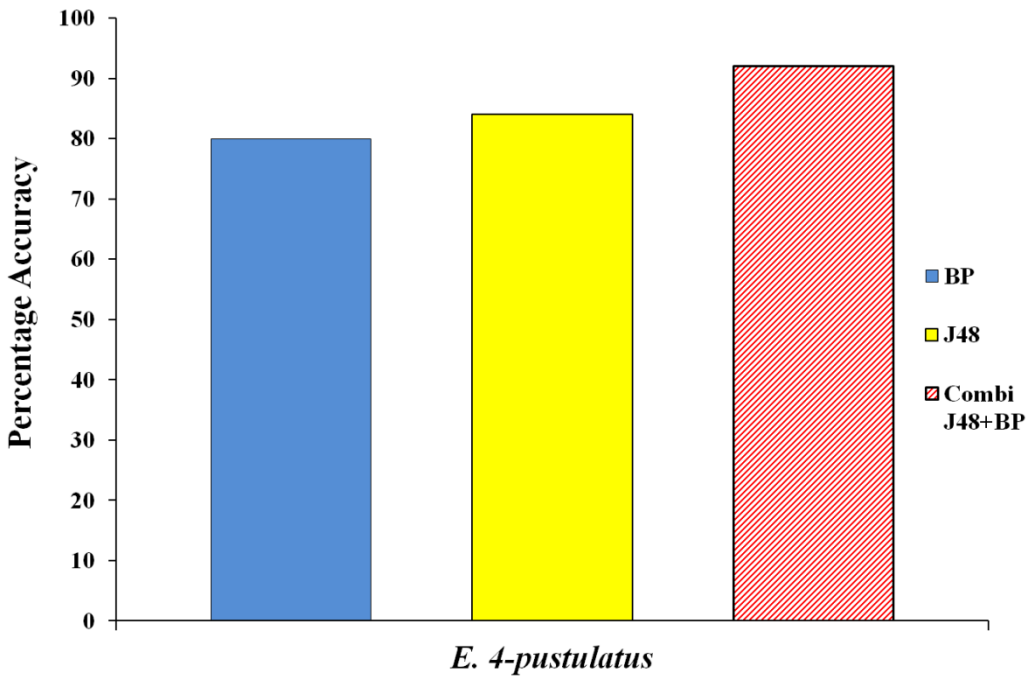


Figure 6.7: Cross-validation accuracies for *H. axyridis* f. *spectabilis* against *E.4-pustulatus* using BP, J48 and a combination of the two classifiers

On the test of *H. axyridis* f. *spectabilis* against other species, the results were improved as shown in Figure 6.8. *A. 2-punctata* scores an improvement around 3%, slightly well better than using J48 alone; *C. 5-punctata* and *C. 7-punctata* improves very little. The test revealed a colour feature '*Spot colour b**' a suitable feature that minimise the decision tree and simplifies the solution for the test on *A. 2-punctata*, while the colour feature '*BG colour a**' is the right feature for the test on both *C. 5-punctata* and *C. 7-punctata*.

The same test has been conducted on *H. axyridis* (H1H2H3) to investigate intra-species identification. Results are shown in Figure 6.9. It shows *H. axyridis* f. *spectabilis* can be correctly identified against *H. axyridis* f. *conspicua* to 72.5% accuracy, and 97.5% correctly identified against *H. axyridis* f. *succinea*.

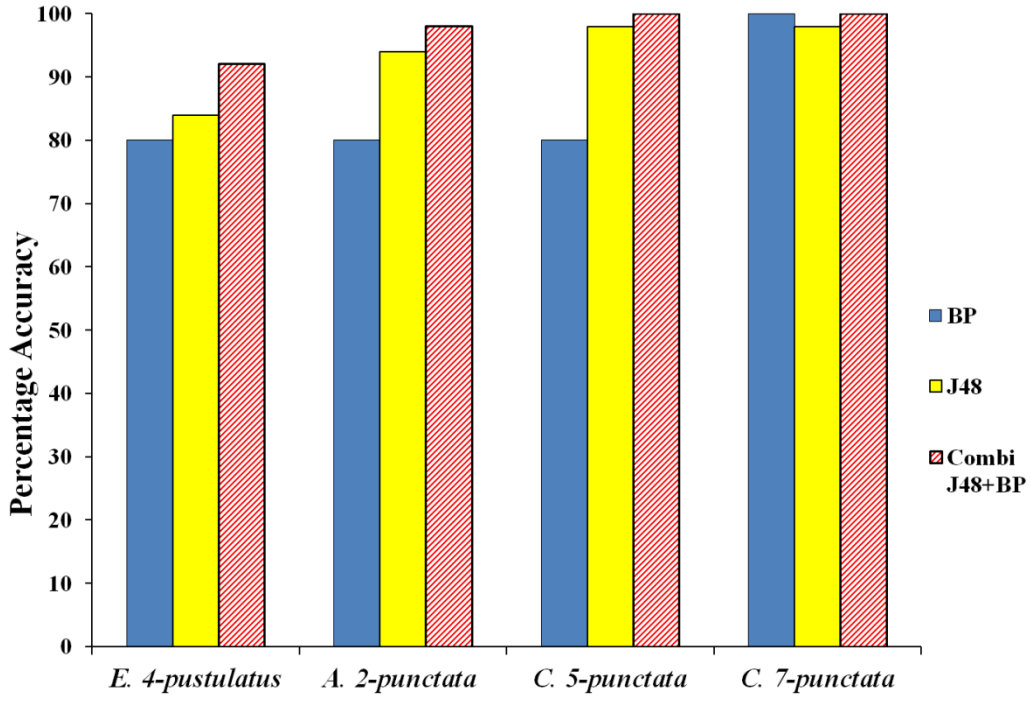


Figure 6.8: Cross-validation accuracies for *H. axyridis* f. *spectabilis* against *E.4-pustulatus* and other species using BP, J48 and a combination of the two classifiers

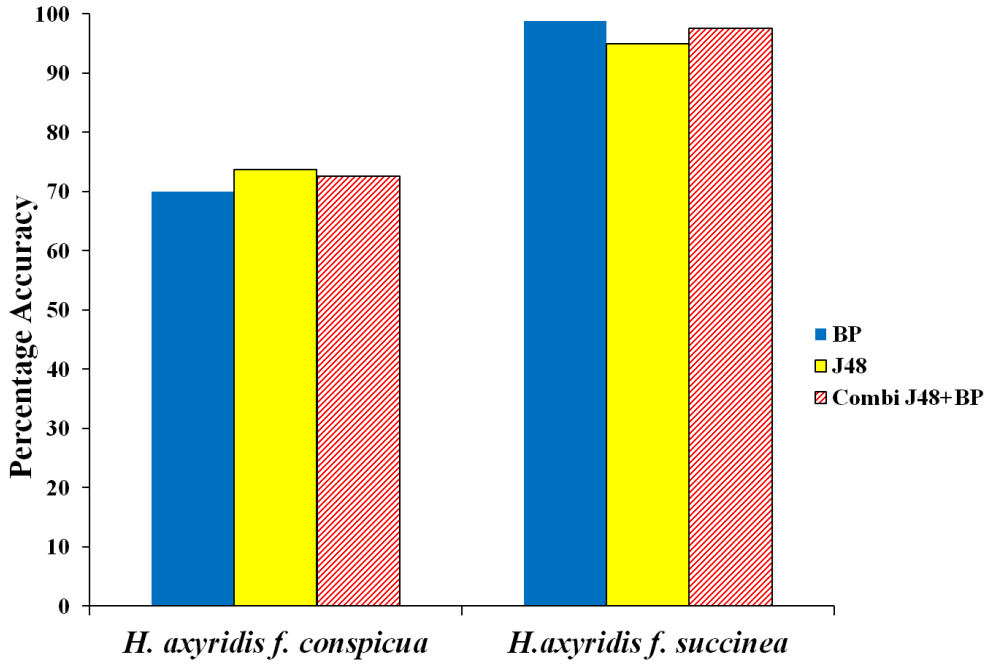


Figure 6.9: Intra-species cross-validation accuracies for *H. axyridis* f. *spectabilis* against *H. axyridis* f. *spectabilis* and *H. axyridis* f. *succinea* using BP, J48 and a combination of the two classifiers

On average, the area under curve (acronym 'AUC') showed significant improvement from 0.5 when using MLP alone, to 0.825 using J48 alone, and finally 0.913 using a combination of both classifiers. AUC represents "..the probability that a randomly chosen positive example is correctly ranked with greater suspicion than a randomly chosen negative example" (Bradley, 1997). The use of ROC curves and AUC does not depend on class skews, hence they are important metrics to consider for evaluating identification systems. The fact that AUC improves show that for the identification of *E. 4-pustulatus* and *H. axyridis* f. *spectabilis*, the use of combination classifiers can improve their identification.

6.4.2.2 Balanced class

Next, a balanced class distribution was used, where 40 samples per species were trained and tested. The same processes performed on *E. 4-pustulatus* and *H. axyridis* f. *spectabilis* in the unbalanced set were repeated for this balanced dataset. Test results and the resulting identification metrics for the balanced dataset are given in Tables 6.26a -6.26c. Training outcomes are provided in Table 6.27.

Table 6.26a: Confusion matrix for MLP (balanced class)

	E4	**H1**
E4	40	2
H1	0	38

Table 6.26b: Confusion matrix for J48 decision tree (balanced class)

	E4	**H1**
E4	37	3
H1	3	37

Table 6.26c: Confusion matrix for combination of J48 and MLP (balanced class)

	E4	**H1**
E4	40	3
H1	0	37

Table 6.27: Training of MLP for balanced class of *E. 4-pustulatus* and *H. axyridis* f. *spectabilis*

Run	Q	N	M	J	I	RMS error		% Accuracy	
						MLP	With J48	MLP	With J48
1	80	12	36	2	500	0.46	0.61	78.75	62.5
					1000	0.16	0.49	97.5	76.25
					1500	0.16	0.30	97.5	91.25
2	80	12	24	2	500	0.25	0.19	93.75	96.25
					1000	0.16	0.19	97.5	96.25
					1500	0.16	0.19	97.5	96.25
3	80	12	12	2	500	0.19	0.19	96.25	96.25
					1000	0.16	0.19	97.5	96.25
					1500	0.16	0.19	97.5	96.25
4	80	12	8	2	500	0.19	0.19	96.25	96.25
					1000	0.16	0.19	97.5	96.25
					1500	0.16	0.19	97.5	96.25
5	80	12	4	2	500	0.16	0.19	97.5	96.25
					1000	0.16	0.19	97.5	96.25
					1500	0.16	0.19	97.5	96.25

Note:

Q = no. of exemplar vectors N = input features I = no. of iterations
M = no. of hidden neurons J = no. of output neurons

It is interesting to note the difference in performances between an unbalanced class and a balanced class distribution of the same test dataset, in this case the identification of *E. 4-pustulatus* and *H. axyridis* f. *spectabilis*. The metrics in concern are accuracy, RMS error and AUC. For 24 hidden neurons, the accuracy of the balanced class improved from 93.8% using MLP, to 96.3% when MLP was combined with J48. When the number of hidden neurons was reduced, there was slight reduction in accuracies for the combination system. This means there were not enough neurons to train the network.

The RMS error values were better for the balanced class. This is observed based on the steady minimum values throughout the runs which converge to 0.26 for MLP,

and 0.19 for the combination system. In contrast, the unbalanced class gives RMS error of 0.2 for MLP, and 0.29 for the combination system.

It is interesting to note that the results of this test complies with the two-sample Kolmogorov's test, which shows that if the class distribution is not balanced a huge value is required for the dimensionality (Evangelista, 2006). This is shown in Figure 5.14, where N_1 is the number of samples in the minority class and N_2 is the number of samples in the majority class.

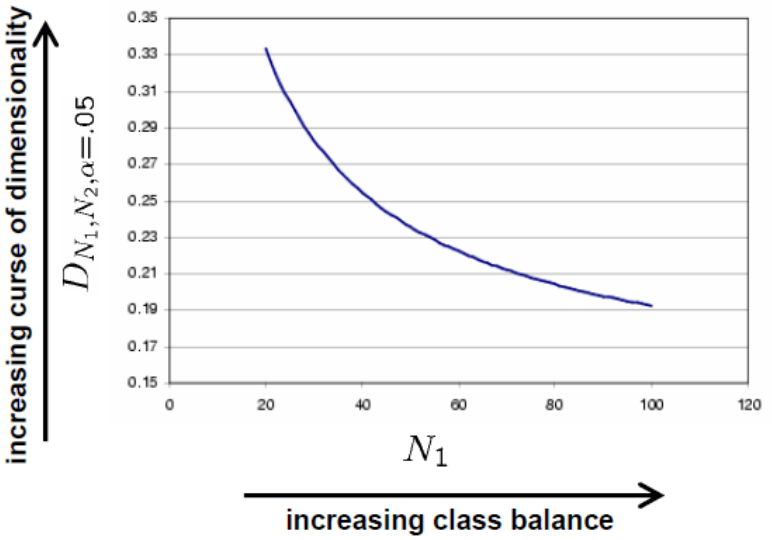


Fig 6.10: Two-sample Kolmogorov test with fixed N_2 =100, α = 0.05 (Evangelista, 2006)

Here Kolmogorov showed that if the classes were more severely imbalanced, the curse of dimensionality will grow exponentially. Similar situation occurs for the project in hand.

6.5 Summary

Sorting of Harlequins and non-Harlequins using a classifier is proposed. WEKA using J48 pruned tree is a useful reference when creating a rule-based expert system because it provides the decision paths which create the rules. Inter-species separation can be performed using geometrical features or colour features. This selection depends on the visual appearance or popular query. For instance, a better scheme may ask user for body colour to narrow down query and this can be reconfirmed with CIELAB colour check. For some species, intra-species cases may be separated using colour features. This is because specimen from same species but different forms may have minimal variations in geometrical measurements, and geometrical features are linearly related. In this case colour will be an excellent choice, and this can be confirmed from statistical data.

Comparisons of identification using MLP, SVM, LVQ and PNN classifiers for ladybird species placed under White, Red and Black spot colour groups have been investigated. Analysis based on these colour groups will be useful as a guide when the system is tested as a whole with other system components. For the identification of *E. 4-pustulatus* and *H. axyridis* f. *spectabilis*, results are discussed on the basis of balanced and unbalanced class distributions. This is summarised in Tables 6.28 and 6.29. It is shown that a balanced class distribution is an excellent choice for automated ladybird identification, and has been used as a basis for testing the classifiers.

Table 6.28: Summary of results (Unbalanced class distribution)

No	Test Group	Classes	Classifier(s)	Accuracy based on features (%)		
1	**Inter-species**			**Geometrical**	**Colour**	**All**
	White	C14H16	MLP			
			SVM	100	92	98
			LVQ	100	94	98
			PNN	98	98	98
		E4H1	MLP	80		
			J48	84		
			MLP + J48	92 (3 features)		
		A2H1	MLP	80		
			J48	96		
			MLP + J48	86		
		C5H1	MLP	80		
			J48	98		
			MLP + J48	96		
		C7H1	MLP	80		
			J48	98		
			MLP + J48	100		
2	**Intra-species**					
		H1H2	MLP	70		
			J48	73.8		
			MLP + J48	72.5		
		H1H3	MLP	98.8		
			J48	95		
			MLP + J48	97.5		

Table 6.29: Summary of results (Balanced class distribution)

No	Test Group	Classes	Classifier(s)	Accuracy based on features (%)		
1	**Inter-species**			**Geometrical**	**Colour**	**All**
	White	C14H16	MLP	100	83.3	100
			SVM	100	90	100
			LVQ	100	90	100
			PNN	100	96.3	100
	Red	E4H1H2	MLP	88.9	5.6	66.7
			SVM	65.8	63.3	76.7
			LVQ	68.3	75.8	77.5
			PNN	64.2	71.7	75.8
	Black	A2C5C7H3	MLP	70.8	83.3	83.3
			SVM	50.6	66.9	74.4
			LVQ	65.6	70	74.4
			PNN	75	84.4	88.8
		E4H1	MLP	97.5		
			J48	92.5		
			MLP + J48	96.3		
		A2H1	MLP	80		
			J48	94		
			MLP + J48	98		
		C5H1	MLP	80		
			J48	98		
			MLP + J48	100		
		C7H1	MLP	100		
			J48	98		
			MLP + J48	100		

CHAPTER 7

SYSTEM INTEGRATION

CHAPTER 7

SYSTEM INTEGRATION

The previous chapter shows the importance of neural networks; they can be a learning module inside an automated identification system. Decision trees have been working well with neural networks. This chapter will show a use of decision trees other than for extracting the most important features. The resultant decision tree can be used for creating meaningful rules for a rule-based expert system. This idea is elaborated next with a proposed solution for building the overall automated ladybird identification system.

7.1 Introduction

Imagine an expert system query for specific characters or 'features' related to a ladybird. Typically it will start off with a query on colour, which would reduce the problem to a smaller number of species. For typical query, background colour can be the primary interrogator followed by spot colour. It may be useful to get extra inputs from query (spot count, pronotum patterns and pronotum colour) whenever image evident is scarce. Other than characters and geographic information, an expert normally requires time of day, season of the year, habitat (grass, trees, wetlands,

conifers, coastal, generalist), and context (Atkinson and Gammerman, 1987; Clark, 2007).

Whenever both geometric and colour features are required, as in the case of *E. 4-pustulatus vs. H. axyridis* f. *spectabilis*, this means not enough information is obtained yet. It could be that the boundary of separation is too small, or training samples are insufficient. It may be useful to get other extra inputs (pronotum patterns/colour). Hence this suggests the use of an expert system as a knowledge base. The neural network may become a source for the database checker in the inference system. Expert system implementation can start with using colour as primary interrogator. The tests in previous chapter show that spot colour and elytra colour are the primary interrogator for the following species: 2-spot, 5-spot and 7-spot ladybirds. The system will also need extra information like 'count of spots' and 'pronotum pattern', which can only be reliably obtained from the image itself, or from user's observation.

7.2 Proposed overall ASI system

This thesis is proposing a framework of working and individually tested components of a prototype ASI system as shown in Figure 1.5 in Chapter 1 earlier on. It emphasizes user interaction and explanation facility. There is a decision tree and a rule-based inference engine. The idea on applying rule-based inference engine has already been applied by researchers in the field of mechatronics, where Sugumaran (2007) has applied the concept in a research on the fault diagnosis of roller bearing, and further extended by Saravanan et al. for an application in vibration-based fault diagnosis of spur bevel gear box (Sugumaran, 2007; Saravanan, 2009). In the field

of agricultural industry, Omid designed an expert system for sorting pistachio nuts using decision tree and fuzzy logic classifier (Omid, 2011).

The decision tree uses input from 12 possible features obtained through image processing operations, as explained in previous chapters. Due to the 'curse of dimensionality', there is inherent limitation in the number of features to arrive at a solution. The use of decision tree algorithm has indicated which feature is best for classification for the given training set, rather than using all features. This saves resources (time and labour). The root node on top of the tree shows the best feature and other nodes show features which are arranged in descending order of importance. The values appearing between the nodes show the level of contribution, and they are useful for generating rules. The rule based part of the system aims at embedding structured human expertise into algorithmic form (Kecman, 2001). The block diagram is shown in Figure 7.1.

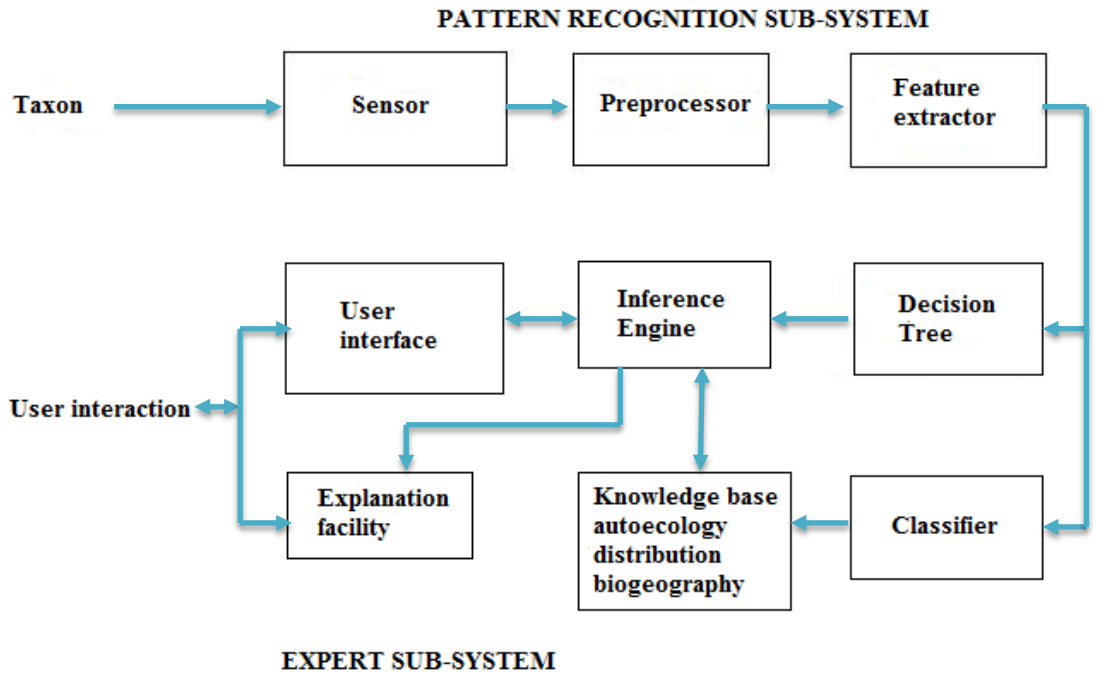


Figure 7.1: Block diagram of proposed hybrid system

Knowledge-based systems have the ability to handle non-linear, fuzzy and incomplete data; therefore, they are more suitable as the core for any computer-aided taxonomy (CAT) system (Chesmore, 2007). The proposed system is a dual-action system. The pattern recognition sub-system deals with input acquisition from images, where inputs are physical features and colours of the taxon. The extracted features are then fed to a classifier, while also being channeled to the decision trees. The output of the classifier will be stored in the knowledge base. The expert sub-system contains a rule-based inference engine based on decision trees. This part is slightly different from the original ATI system proposed by Chesmore (Chesmore, 2007). It receives the features and the resultant from the classifier. It also interacts with users through user interface. A neural knowledge base about the taxon interacts with the inference engine to supply approximate reasoning. While this neural network is able to learn, the inference engine should be able to provide reasoning. All these forms a recipe for an explanation facility to develop as the system evolves. This makes the system unique and differ from existing automated taxon identification systems.

7.2.1 Implementation based on MATLAB and WEKA

Fuzzy-inspired logic is used in the proposed system due to its ability to deal with uncertainty (Saravanan, 2009; Negnevitsky, 2005). The knowledge base may get information from human interactions, which can be inconsistent hence fuzzy (Zadeh, 1983). For instance, in the ladybird identification domain, typical characters used by experts include the length of body, spot count, elytra colour, spot colour, etc. Whilst a couple of them are precisely measured through image processing techniques, some of these are fuzzy in nature. In fact, the interpretation of colour itself varies between

individuals, for instance, different users may interpret redness level differently. Fuzzy-inspired logic maps the input space to the output space through a list of rules in the inference engine. The rules use *'if-then'* statements which are evaluated in parallel. Membership functions are defined based on the decision tree condition at a particular node. The curves define the mapping to a degree of membership, normally between 0 and 1. Choices for membership function include trapezoidal, Gaussian, log, etc. The selection of which membership function to use is arbitrary. In the ladybird identification system, trapezoidal membership function is proposed. For each trapezoidal function there is a threshold. This threshold value is given by the decision tree based on the training dataset. Once the threshold is known, other parameters of the trapezoidal function can be determined. In this system, the threshold is set to lie in the mid-point along the sloping line formed by the interconnection of the points of inflection. This is elaborated in the next sub-section.

7.2.2 Estimating the parameters of membership function

An arbitrary trapezoidal membership function is shown in Figure 7.2.

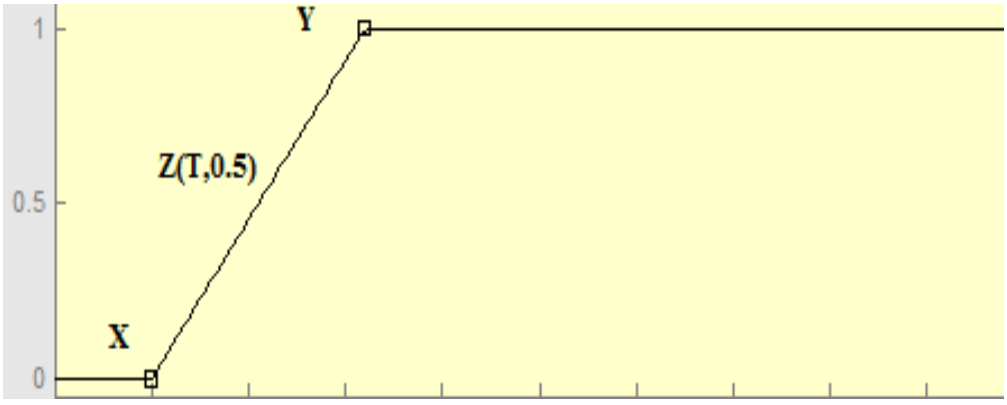


Figure 7.2: An arbitrary trapezoidal function

Let the coordinates X (*P*, *0*) and Y (*Δx, 1*) be the points of inflection. These points are where the gradient starts to change. The aim is to estimate the slope and

inflection points after a threshold Z (T, 0.5), which is also the mid-point, is obtained using decision tree. The formulation which follows is defined for an error of ± 0.05 or 5% due to normalization. The lower and upper limit are set to m = 0 and infinity. Since the limit for Δy is 1, slope is given by:

$$Slope = \frac{1}{\Delta x} \qquad (6.1)$$

Since T is known, X is obtained from:

$$X = \left(T - \frac{\Delta x}{2}, 0\right)$$

7.3 Fuzzy system test results

In the proposed system WEKA is used for generating decision tree, while the rest of the system is designed in MATLAB. The fuzzy logic toolbox is readily available in MATLAB, and the 'if-then' rules for the inference engine can be entered. In short, the sequence of operations is described as:

1. Ensuring data is normalised, get decision tree.
2. Define membership functions for all branches.
3. Set 'if-then' rules.
4. Test.

For the purpose of showing the usefulness of the proposed system, the same data which was used in the previous tests are applied in the proposed system. The proposed system has been tested on the following sets of test data:

- White-spotted ladybird group
- Red-spotted ladybird group
- Black-spotted ladybird group

For each group, an inference engine was built based on the rules. Membership functions have been defined after obtaining the decision tree structure which shows threshold values.

7.3.1 White-spotted ladybird group

The group consists of two white-spotted ladybird species in the training set, namely *C. 14-guttata* (C14) and *Halyzia 16-guttata* (H16). There were 80 samples in total, equally divided between the two species making 40 samples per species. The test data is given in Table A2 in Appendix IV. All 12 features have been fed into the system and the decision tree generated as shown in Figure 7.3:

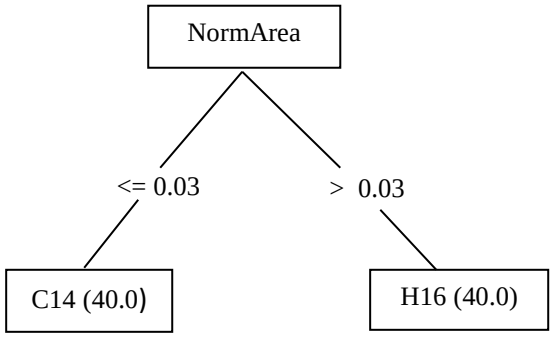


Figure 7.3: Decision tree for White-spotted ladybird group

The rules are:

1. IF (*NormalisedArea is NormArea*) THEN (*WhitespotLadybird is Orange*)
2. IF (*NormalisedArea is not NormArea*) THEN (*WhitespotLadybird is Creamspot*)

The rules stated above apply for all related White-spotted ladybirds. Membership functions generated from this operation is given in Figure 7.4.

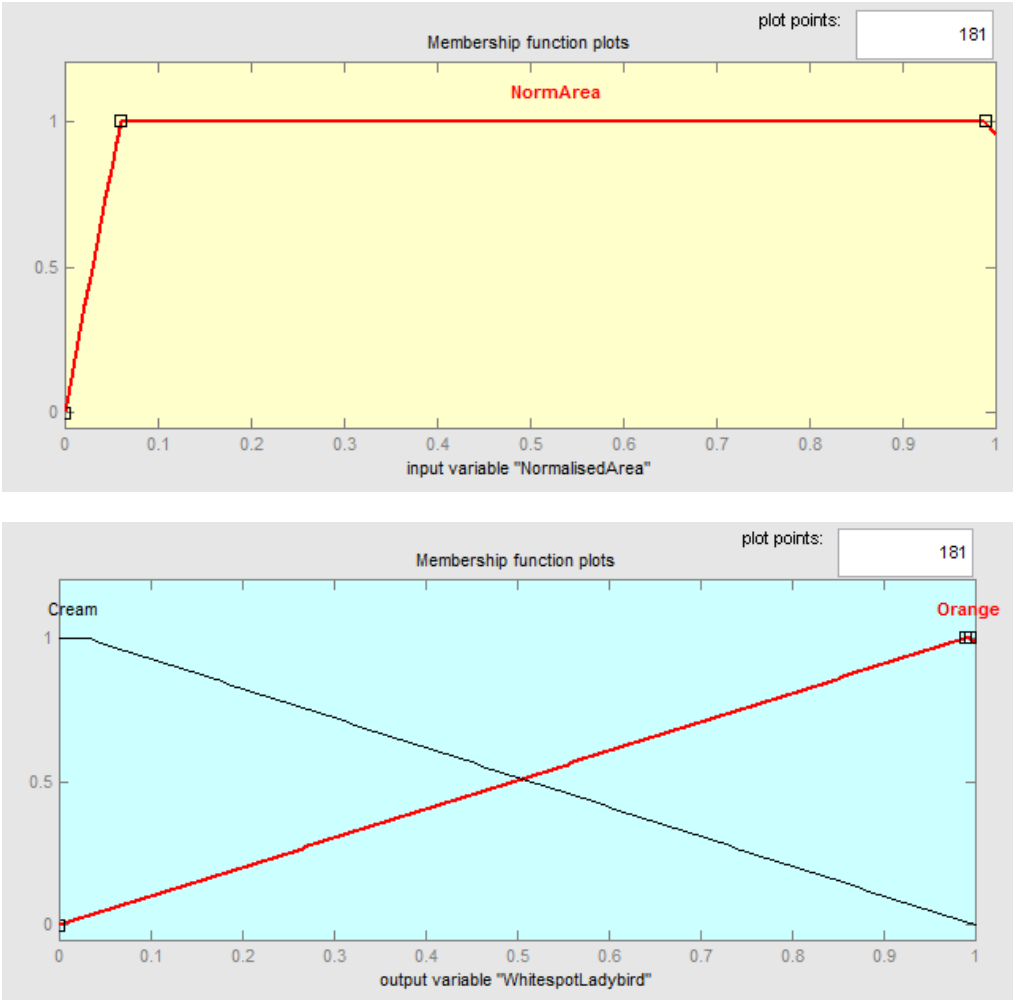


Figure 7.4: Membership functions for White-spotted ladybird

The rule viewer for one test data (NormalisedArea = 0.0213) is shown in Figure 7.5.

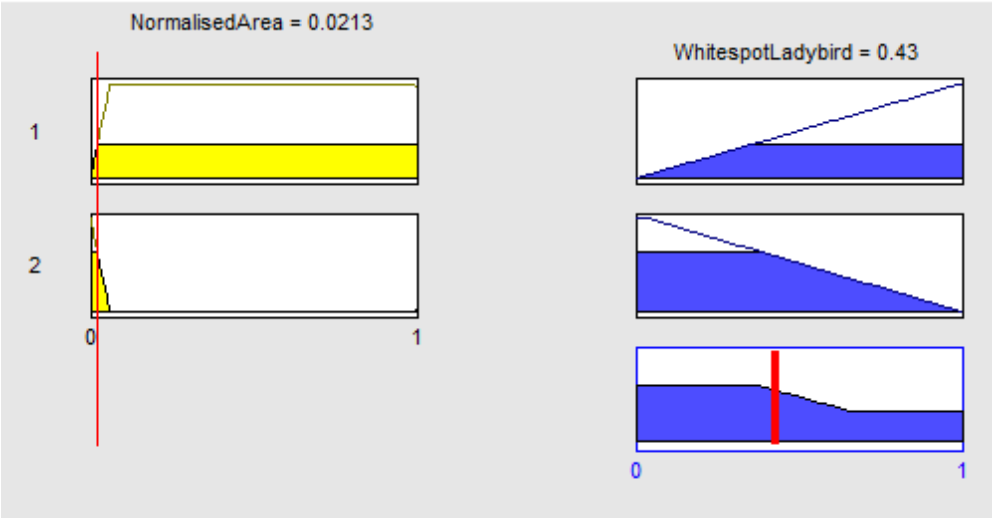


Figure 7.5: Rule viewer for test on White-spotted group

The output value for 'WhitespotLadybird' is 0.43 indicating the resultant identification is C14. To avoid results for the fuzzy-inspired system been obtained by chance only, random test data is created. Taking random samples of test values and making them a test set data; the test set now contains about 20% of the overall samples. The membership functions are shown in Appendix V. The test is repeated for this test set, and the resultant confusion matrix is shown in Table 7.1.

Table 7.1: Confusion matrix for White-spotted group

	C14	**H16**
C14	8	0
H16	0	8

Based on the confusion matrix, perfect classification was obtained for both species. This is calculated as:

$$\%\ accuracy = \frac{(TP + TN)}{(TP + FP + FN + TN)}$$

C14 and H16 can be identified using the feature 'NormArea'. This is evident from results of decision trees and the rule-based inference engine. Since accuracy is 100%, user input for further investigations is not required.

7.3.2 Red-spotted ladybird group

The group consists of three red-spotted ladybird species in the training set, namely *E. 4-pustulatus* (Pine ladybird) and two forms of Harlequin ladybirds (*H. axyridis* f. *spectabilis* and *H. axyridis* f. *conspicua*). There were 120 samples in total, equally divided between the three making 40 samples per species. The twelve features have been fed into the system and the decision tree as shown in Figure 7.6 is generated:

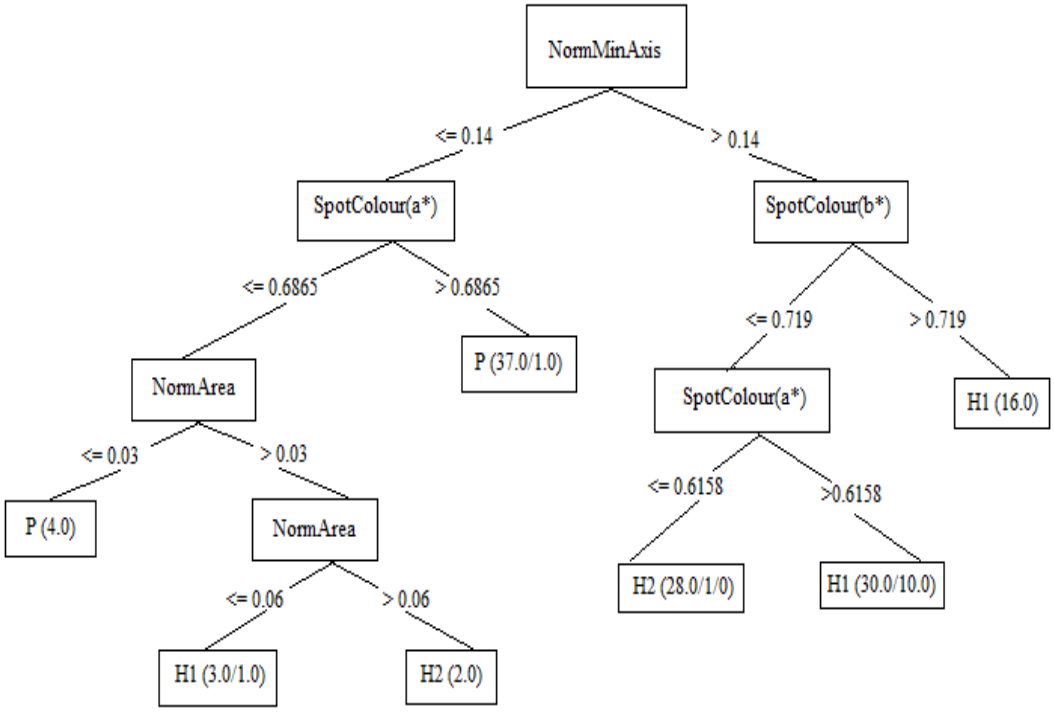
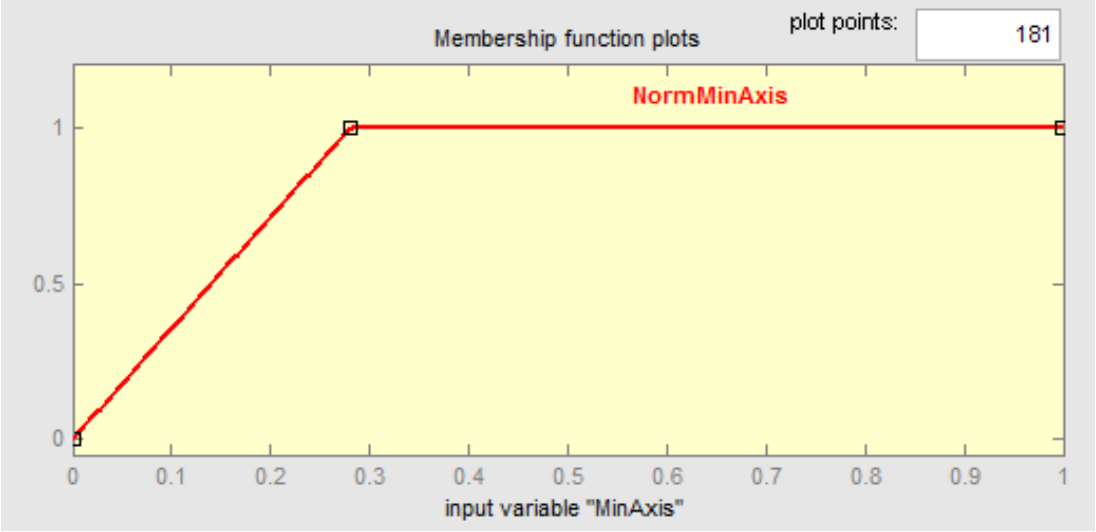
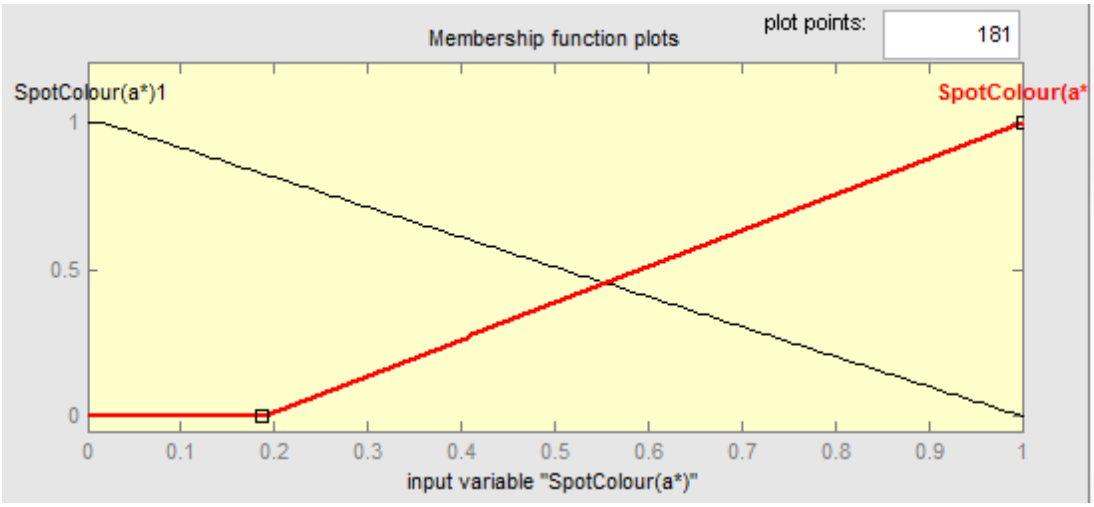
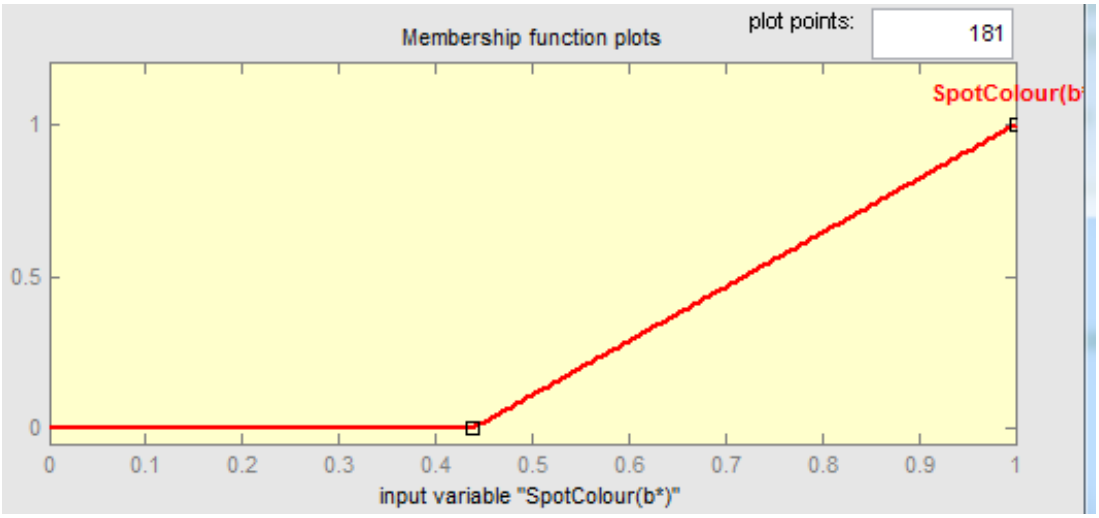


Figure 7.6: Decision tree for Red-spotted ladybird group

The membership functions for input variables are given in Figure 7.7, and membership functions for output variables are provided in Figure 7.8.







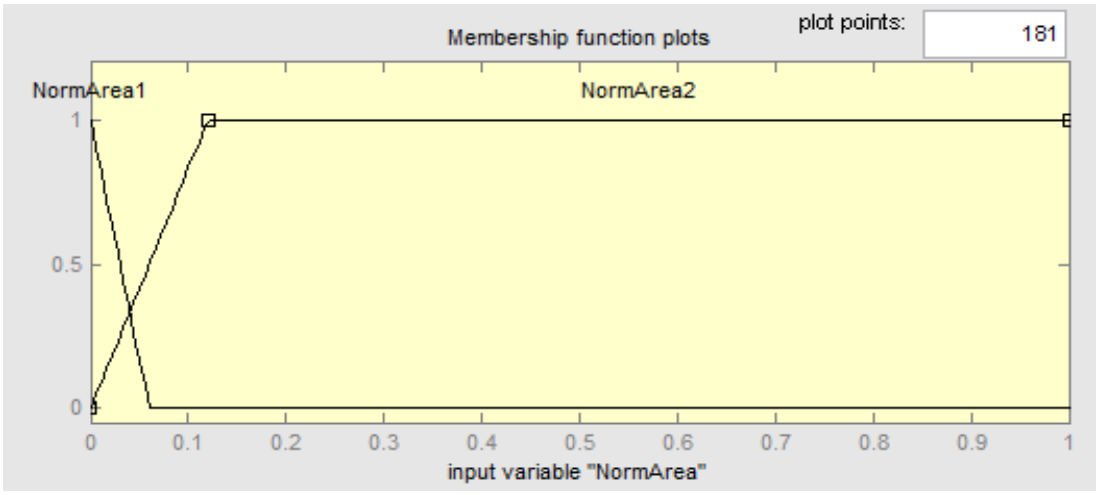


Figure 7.7: Membership functions for input variables (Red-spotted ladybird group)

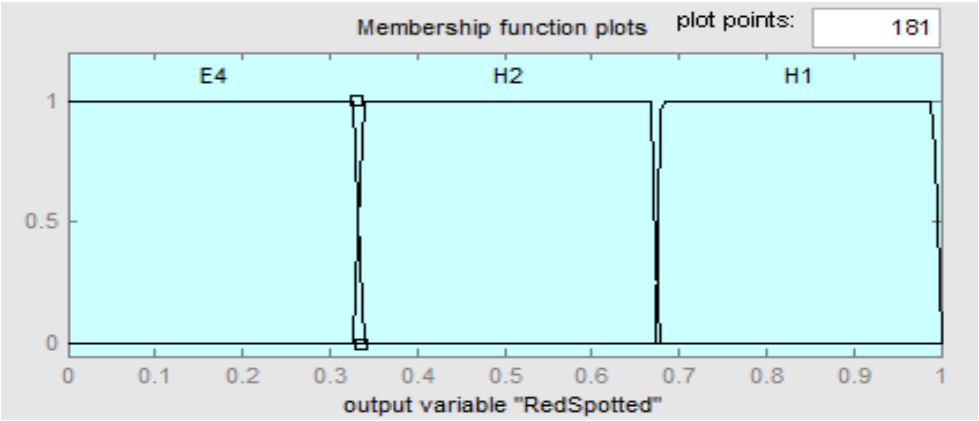


Figure 7.8: Membership functions for output variables (Red-spotted ladybird group)

The rules are:

1. IF (*MinAxis is NormMinAxis*) and (*SpotColour(b*) is SpotColour (b*)*) THEN (*RedSpotted is H1*)
2. IF (*MinAxis is NormMinAxis*) and (*SpotColour(a*) is not SpotColour (a*)1*) and (*SpotColour(b*) is not SpotColour (b*)*) THEN (*RedSpotted is H1*)
3. If (*MinAxis is NormMinAxis*) and (*SpotColour(a*) is SpotColour (a*)1*) and (*SpotColour(b*) is not SpotColour (b*)*) THEN (*RedSpotted is H2*)
4. IF (*MinAxis is not NormMinAxis*) and (*SpotColour(a*) is SpotColour (a*)2*) THEN (*RedSpotted is E4*)
5. IF (*MinAxis is not NormMinAxis*) and (*SpotColour(a*) is not SpotColour (a*)2*) and (*NormArea is NormArea1*) THEN (*RedSpotted is E4*)
6. IF (*MinAxis is not NormMinAxis*) and (*SpotColour(a*) is not SpotColour (a*)2*) and (*NormArea is not NormArea1*) THEN (*RedSpotted is H2*)

The rule viewers for each species are shown in Appendix V. Similar to the previous test, to avoid results for the fuzzy-inspired system been obtained by chance only random test data is created for this group. Taking random samples of test values and

making them a test set data; the test set now contains about 20% of the overall samples. This test data is shown in Table A3 (Appendix IV). The test is repeated for this test set, and the resultant confusion matrix is shown in Table 7.2.

Table 7.2: Confusion matrix for Red-spotted group

	E4	**H1**	**H2**
E4	8	0	0
H1	0	8	0
H2	0	7	1

7.3.3 Black-spotted ladybird group

The group consists of four black-spotted ladybird species in the training set, namely *A. 2-punctata*, *C. 5-punctata*, *C. 7-punctata* and one form of *H. axyridis* f. *succinea*. There were 160 samples in total, equally divided between the four making 40 samples per species. The twelve features have been fed into the system and the decision tree generated as shown in Figure 7.9. The membership functions, if-else rules and rule viewers are given in Appendix V. Similar to the previous tests, random test data is created for this group. Taking random samples of test values and making them test set data, it contains about 20% of the overall samples. This test data is shown in Table A4 (Appendix IV). The test is repeated for this test set, and the resultant confusion matrix is shown in Table 7.3.

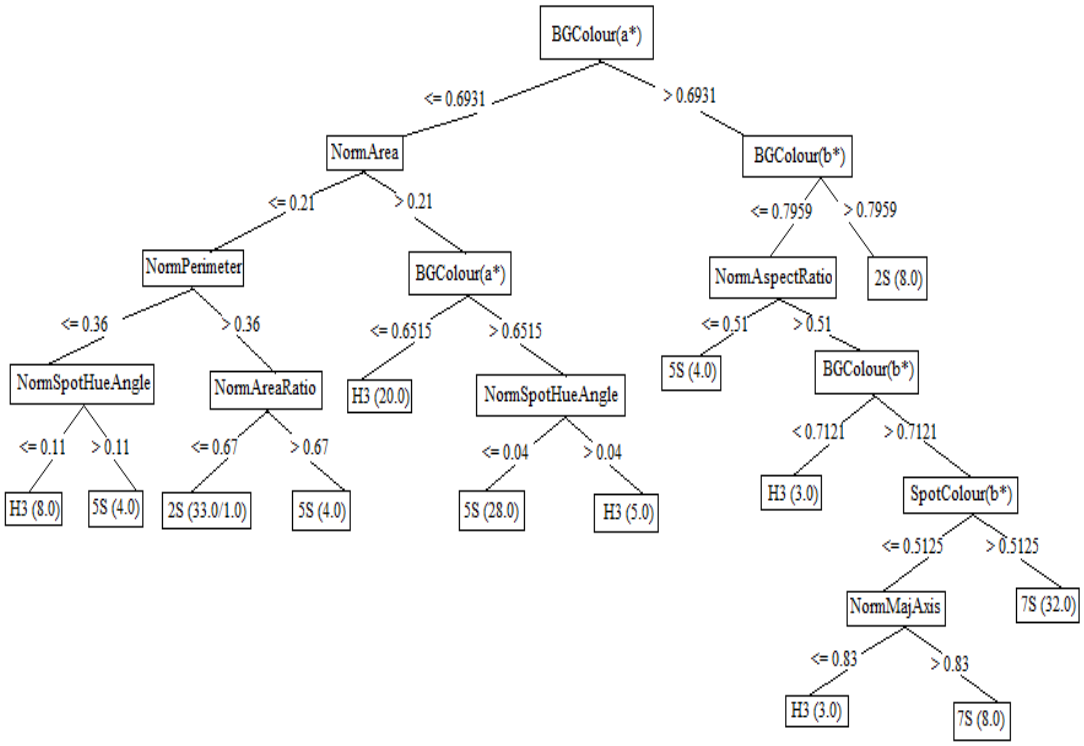


Figure 7.9: Decision tree for Black-spotted ladybird group

Table 7.3: Confusion matrix for Black-spotted group

	A2	C5	C7	H3
A2	0	8	0	0
C5	0	1	7	0
C7	0	3	5	0
H3	0	0	8	0

7.4 Overall Analysis

Table 7.1 shows perfect identification results for both species. However, for Table 6.2, the percentage of correct identification is lower, just about 70.8%. This is due to confusion between *H. axyridis* f. *spectabilis* and *H. axyridis* f. *conspicua*. Only one instance of *H. axyridis* f. *conspicua* is correctly identified. It is interesting to note that identification accuracy is 100% if separation between *E. 4-pustulatus* and *H. axyridis* is only required.

For the confusion matrix given in Table 7.3, the percentage of correct identification is only 18.8%. These results are poor due to a few factors. First, there was an increased number of classes required to be identified, unlike White and Red groups. Secondly, most of the misclassification of A2 during testing showed many confusions between A2 and C5. Looking at the decision tree A2 should have been easily discriminated by its area ratio, where 33 instances have reached the NormAreaRatio leaf and shows only one misclassification. Similar observation is profound between H3 and C7. The background colour has been the most prevalent of the characters; however, it is also highly variant for H3. Checking through individual images for background colour shows the pronotum colour of H3 is actually highly variable from pale yellow-orange to orange-red. This is supported by CIELAB colour distribution showing positive correlation. In other words, to discriminate H3 and C7 two primary characters are needed: the background colour and the spot colour.

From a different perspective, it can be interpreted in a positive way by looking at the identification of pairs of species. Looking at the identification of the pair *C. 5-punctata* and *C. 7-punctata* shows there are 7 incorrect identifications in row 2 and 3 in the third row giving a total of 10 incorrect identifications. This is more than half

of the number of samples used to test between the two species. One way would be to join the instances together, as in Table 7.4.

Table 7.4: Adjusted confusion matrix for Black-spotted group

	A2	C5/C7	H3
A2	0	8	0
C5/C7	0	16	0
H3	0	8	0

The accuracy is now 50%, which is an improvement by 31.2%. This scheme will work fine because, in reality, the two species are always confused due to their similarities in term of colours and physical measurements and there is a need to get the best feature to identify them. The number of spots could be the best feature for separating them in the feature space, as their species names suggest. Unfortunately, the number of spots is not one of the features been used in this work due to potential occlusion in images. One way around this is to get user's confirmation of the number of spots, and any extra inputs the system can get to assist identification, such as biogeography and distribution data. The knowledge base will need updating. If A2 and C5/C7 be joined together, the confusion matrix reduces to two 'groups'; one with H3, and another group is non-Harlequin. The revised confusion matrix is shown in Table 7.5.

Table 7.5: Revised confusion matrix for Black-spotted group

	A2/C5/C7	H3
A2/C5/C7	24	0
H3	8	0

The sorting system shown here has an improvement, as now three-quarter of the samples is non-Harlequins meaning the accuracy is 75%. However, notice that all H3s are still confused as non-Harlequins. The system will need to either query the user for extra information, perhaps the location where the unknown taxon was found,

the date and time, or check with the recently updated knowledge base. Once supplied by the user, the inference system will need to check this piece of information with the knowledge base, which by now should gain updated knowledge. If the supplied location is not in the areas where Harlequins are known to exist, it may not be a Harlequin at all. This technique also applies to the Red-spotted group, as per confusion matrix in Table 7.2.

Readers need to be aware that the proposed system has been proposed using a fuzzy-inspired expert system and neural knowledge base. In a more realistic way of sorting out whether a ladybird is a Harlequin or non-Harlequin, the expert system section can be achievable using a rule-based inference engine utilising decision trees. The decision trees produce production rules, hence generating pseudocodes for implementation. The threshold values at each node, as the name implies, become the deciding factor for the identification steps. Users, however, will be prompted for some 'key' questions. The questions will start by asking user the body length (in mm), pronotum pattern, location, time of year, one at a time (Southampton Natural History Society, 2005). The system will need to flag either 1 or 0 for each answer, where an array of these bits should finally make a justified and reliable identification.

An instance of identification is shown below:

>> *Do you have the body size?* *Y / N*

>> *Is the size less than 6 mm?* *Y / N*

>> *Are there markings on the pronotum?* *Y / N*

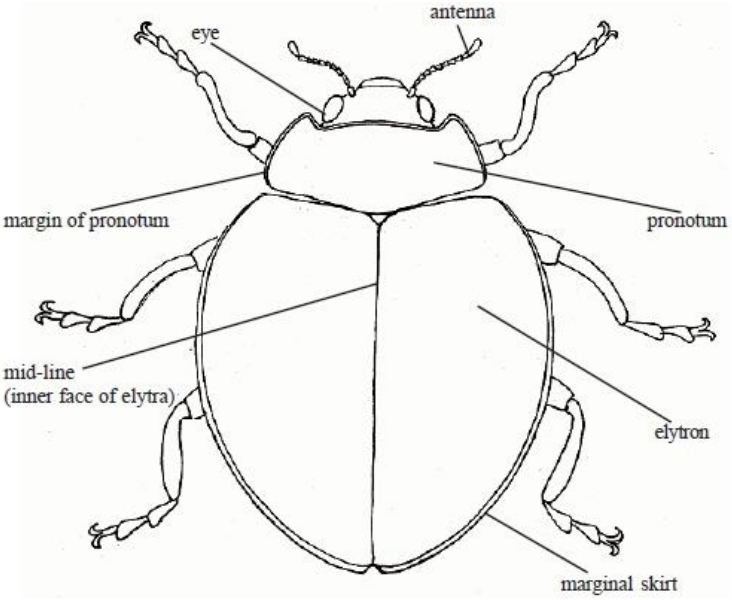


Figure 7.10: Ladybird anatomy (Southampton Natural History Society, 2005)

>> *Habitat last found: Grassland* *Y / N*

Garden *Y / N*

Heathland *Y / N*

Conifers *Y / N*

>> *Where did you find the insect (postcode if known):*

>> *Estimated date and time found:*

These questions, as one would have now realised, are definitely absent in an image-based identification system known to the author. It plays the role of the human expert part of the system, much like using dichotomous key, and is actually vital towards final identification. This thesis has shown that without denying human interaction and expert inputs, reliable identification can be obtained and without doubt this is a novel approach.

In general, the following steps may help users to follow the identification process:

1. Prepare input image based on requirements.
2. Feed image of unknown taxa into the system. User captures spot colour and elytron colour.
3. System extracts colour and geometrical features.
4. Features of an image passed to decision tree, while neural network as classifier receive features for training.
5. The knowledge base supplies the rest of the training data to neural classifier.
6. The test results of neural classifier become a confidence factor to aid user.
7. Rule-based inference engine deduce inference based on the rules derived from decision trees, and using historical data obtained from knowledge base.
8. An estimate of identification output is produced. User interactions are required to justify uncertainties, for instance, location, number of spots, etc.
9. Final identification is produced when error goes to minimum.

7.5 Summary

The techniques proposed in the chapter emphasised fuzzy-inspired expert system and user interaction to improve ladybird identification. This is a novel aspect of this work as it has not been attempted in any automated identification system, making this approach in itself a contribution to knowledge in the field. Before applying expert system, vital features have been selected through the use of decision tree, which makes the subsequent operations more efficient due to the reduced number of features in use. In terms of confusion matrix the technique is able to identify species of ladybirds, including Harlequins. Species with similar spot colours, such as *C. 14-guttata* and *H. 16-guttata* can be identified correctly. Identification between *E. 4-pustulatus* and Harlequins, *H. axyridis* f. *spectabilis* and *H. axyridis* f. *conspicua*, produces 70.8% accuracy. In delicate situations where similarities between Harlequins and non-Harlequins exist, the confidence level can be too low for the system to identify a species. Therefore, the identification is aided by additional user inputs, which has been shown to improve identification to 75% accuracy.

CHAPTER 8

CONCLUSION AND FUTURE STUDY

CHAPTER 8

CONCLUSION AND FUTURE STUDY

8.1 Conclusion

The automated identification of UK ladybirds has been investigated and implemented. At the start, a literature survey of peer reviewed journal papers and conference proceeding has been conducted which has been restricted to image-based, semi- and fully automated identification systems. In addition to image processing techniques, the classification methods and their identification accuracies have been compared. Some early works used greyscale images, such as the analysis of quarantine fungal pests by Chesmore, Bernard, Inman and Bowyer (Chesmore, Bernard, Inman and Bowyer, 2003). Some systems such as DAISY, VeSTIS and Moth ID, work with 2D-colour images of the specimen. There are researchers who work with plant identification systems such as MORPHIDAS by Clark et al., and work by Stephen Gang Wu; both using 2D-images of leaves (Clark *et al.*, 2007; Wu *et al.*, 2007). Based on the literature survey, so far there has not been an attempt to produce an automated ladybird identification system as proposed in this thesis.

A few key improvements identified from the comparisons require the proposed system to be specific rather than holistic, use morphometric features, able to generate

reasoning and future online implementation. After the needs are identified, the system structure has been formulated which involves colour image processing, neural networks, decision trees, fuzzy-inspired inference engine, knowledge-base and user interactions. Upon viewing the system as a whole, it is not 100% fully automated as it requires user interaction for improved species identification. Pre-processing of images using standard image processing techniques has been performed on all available ladybird images, and is not fully automated. Greyscale information in terms of geometrical features was extracted, where they are made rotation and scale invariant. Colour information has been manually extracted from both elytra and spots via CIELAB colour space. The application of CIELAB colour space in this area is novel, as shown in Chapter 3, where colour distributions for both spot and elytra have been plotted on CIELAB colour planes. These coordinates are meaningful to a designer as feature vectors to use with classifiers.

The use of J48 decision trees has simplified the feature maps by providing a decision path in the form of a tree diagram. It has also revealed threshold points for the use of a rule-based inference engine, and become a rule extractor. The rules are used in the fuzzy inference engine which controls information flow and initiates inference over the neural classifier, which acts as neural knowledge base. Both the decision tree and fuzzy inference engine make a fuzzy-inspired expert system. In short, user involvements are useful in the feature extraction process and during transitional stage between processes.

Experiments on the following classifiers have been conducted:

- MLP using back propagation algorithm
- J48 decision tree
- PNN

- LVQ
- SVM

They are subjected to both a balanced and unbalanced class distribution, where it has been shown in Chapter 6 that using a balanced class distribution gives better accuracies. With that, a summary of results is shown in Table 8.1. It is also noted that the preliminary work on image processing, feature extraction and classifier results are a success, considering the difficulties faced in tackling the 3D nature of the images where spots are obscured.

Table 8.1: Summary of results (Balanced class distribution)

No	Test Group	Classes	Classifier(s)	Accuracy based on features (%)		
1	**Inter-species**			**Geometrical**	**Colour**	**All**
	White	C14H16	MLP	100	83.3	100
			SVM	100	90	100
			LVQ	100	90	100
			PNN	100	96.3	100
2	Red	E4H1H2	MLP	88.9	5.6	66.7
			SVM	65.8	63.3	76.7
			LVQ	68.3	75.8	77.5
			PNN	64.2	71.7	75.8
3	Black	A2C5C7H3	MLP	70.8	83.3	83.3
			SVM	50.6	66.9	74.4
			LVQ	65.6	70	74.4
			PNN	75	84.4	88.8
4		E4H1	MLP	97.5		
			J48	92.5		
			MLP + J48	96.3		
5		A2H1	MLP	80		
			J48	94		
			MLP + J48	98		
6		C5H1	MLP	80		
			J48	98		
			MLP + J48	100		
7		C7H1	MLP	100		
			J48	98		
			MLP + J48	100		

Evaluation of the results reveals that:

- PNN is the best classifier for the identification between C14 and H16 in a balanced class distribution.
- Using colour features provided significant improvement in identification accuracy for the Black group only, whereas the identification accuracies in the White and Red groups did not improve.
- For a balanced class, identification rates are reduced when only colour features are used, hence prompting the use of other features to be used in combination.
- Combination of classifiers improves identification rates for some species, for instance, *E. 4-pustulatus* and *H. axyridis* f. *spectabilis*.

The overall ASI system is an improvement over existing automated identification systems, as it emphasises the use of fuzzy-inspired expert system and neural knowledge base. This is shown in Table 8.2.

Table 8.2: Summary of results after applying fuzzy expert system

Test Group	Classes	Accuracy (%)
White	C14H16	100
Red	E4H1H2	70.8
Black	A2C5C7H3	75

System integration tests show that *C. 14-guttata* and *H. 16-guttata* can be identified 100% correct, and *E. 4-pustulatus* can be correctly identified against the Harlequins (*H. axyridis* f. *spectabilis* and *H. axyridis* f. *conspicua*) to 70.8% accuracy. Initially the result of identification between black-spotted ladybirds shows 18.8% accuracy. Through user interaction and re-grouping into Harlequins and non-Harlequins, the identification has been improved to 75% accuracy and a pre-sorting mechanism is established. It also shows that user inputs can be digested and reused, therefore

making use of the explanation ability of the fuzzy inference engine. User input helps ladybird identification where ambiguity exists, especially when the important character needed does not exist in the 12 extracted features. User input exchanges information with the system, and such information may include the location of where the ladybird was found, time of year, etc. To the best of the author's knowledge, this feature is non-existent in any automated system for identifying ladybirds. In term of software development, the input from entomologists or more specifically, ladybird experts is vital to initially determine the classes to which the ladybird samples belong. Once the classes are established, non-experts can then explore the system.

This thesis has shown that the use of image processing, neural networks and expert systems can be used to perform automated identification of ladybirds, even though it cannot include all 26 UK ladybird species due to shortage of samples for some species. It requires maximal user involvement in the image processing stage, even though minimal user inputs are needed in the classifier stage. Although the components of the system have not been fully integrated, for instance, some components are written in MATLAB, whereas others are JAVA libraries, a useful prototype has clearly been developed here. With this, the research aims and hypothesis have been fulfilled.

8.2 Future work

There are many ways to improve the existing project. In terms of application, the architecture of this project can be slightly modified for the identification of other invasive beetles in UK, for instance, rosemary beetle (*Chrysolina americana*) and lily beetle (*Lilioceris lilii*). These pests have become a threat to rosemary and lily growers in UK (Royal Horticultural Society, 2012). For the image processing part,

texture analysis of the elytra using Fast Fourier Transform (FFT), fractals or wavelet technique can be used to represent the texture even though both frequency-domain based technique and morphological measurements are suitable for identification. The location of spots can be represented using polar or log-polar coordinates. This representation has been applied by Payne (2001) to the analysis of Hawaiian Happy Face spider images and might be a useful feature for identification in addition to existing features (Payne, 2001).

Even so, there is always the need to iron out more important issues such as tackling the 3D nature of the ladybird images. With an automated system in place and the depth of field is unknown, the ladybird size is difficult to estimate. One way around this issue, which has been applied to face recognition research, is to have multiple cameras to capture images at horizontal angles and reconstruct the image by stitching them to prepare for training. The computational loads, however, will be much higher to complete per colour images.

In terms of development software, for applications involving MATLAB there are decision trees toolkits currently available in MATLAB. It means concentrating on MATLAB as the only development platform rather than multi-platform. For mobile applications, developers may use JAVA instead of MATLAB. To developers who require more sophistication, decision trees and fuzzy logic inference engine may be replaced by rough set, which deals with vagueness and ambiguity in human thinking and perception (Pawlak, 1982; Dubois and Prade, 1990; An and Hu, 2012). The derivative of rough set theory, called fuzzy rough decision trees (FRDT), may be useful in simplifying the system. In contrast to decision trees which require selecting nodes and pruning trees, FRDT is generated using fuzzy rough sets in dealing with real valued or fuzzy data sets. This is based on fuzzy lower approximation operator

and done at the part where there is a need to select nodes and splitting branches, rather than using Information Gain (An and Hu, 2012).

APPENDICES

APPENDIX I

LIST OF LADYBIRD SPECIES AND ACRONYM

Table A1: List of 26 UK ladybird species, including Harlequins

Sub-family	Species	Common name	Acronym
Coccinellidae	*Adalia bipunctata* Linnaeus	2-spot ladybird	A2
Coccinellidae	*Coccinella quinquepunctata* Linnaeus	5-spot ladybird	C5
Coccinellidae	*Coccinella septempunctata* Linnaeus	7-spot ladybird	C7
Coccinellidae	*Coccinella magnifica* Redtenbacher	Scarce 7-spot ladybird	SC7
Coccinellidae	*Calvia quattuordecimguttata* Linnaeus	Cream-spot ladybird	C14
Coccinellidae	*Halyzia sedecimguttata* Linnaeus	Orange ladybird	H16
Coccinellidae	*Harmonia axyridis* f. *spectabilis* Pallas	Harlequin ladybird	H1
Coccinellidae	*Harmonia axyridis* f. *conspicua* Pallas	Harlequin ladybird	H2
Coccinellidae	*Harmonia axyridis* f. *succinea* Pallas	Harlequin ladybird	H3
Coccinellidae	*Adalia decempunctata* Linnaeus	10-spot ladybird	A10
Coccinellidae	*Hippodamia variegate* Goeze	Adonis ladybrid	AD1
Coccinellidae	*Anatis ocellata* Linnaeus	Eyed ladybird	E1
Coccinellidae	*Anisosticta novemdecimpunctata* Linnaeus	Water ladybird	W1
Coccinellidae	*Aphidecta obliterate* Linnaeus	Larch ladybird	L1
Coccinellidae	*Coccinella hieroglyphica* Linnaeus	Hieroglyphic ladybird	HY1
Coccinellidae	*Harmonia quadripunctata* Pontoppidan	Cream-streaked ladybird	H4
Coccinellidae	*Myzia oblongoguttata* Linnaeus	Striped ladybird	S1
Coccinellidae	*Coccinella undecimpunctata* Linnaeus	11-spot ladybird	C11
Coccinellidae	*Hippodamia tredecimpunctata* Linnaeus	13-spot ladybird	C13
Coccinellidae	*Propylea quattuordecimpunctata* Linnaeus	14-spot ladybird	P14
Coccinellidae	*Tytthaspis sedecimpunctata* Linnaeus	16-spot ladybird	C16
Coccinellidae	*Myrrha octodecimguttata* Linnaeus	18-spot ladybird	C18
Coccinellidae	*Psyllobora vigintiduopunctata* Linnaeus	22-spot ladybird	C22
Epilachninae	*Subcoccinella vigintiquattuorpunctata* Linnaeus	24-spot ladybird	C24
Epilachninae	*Henosepilachna argus* Geoffory in Fourcroy	Bryony ladybird	B1
Chilocorinae	*Chilocorus bipustulatus* Linnaeus	Heather ladybird	HE1
Chilocorinae	*Chilocorus renipustulatus* Scriba	Kidney-spot ladybird	K1
Chilocorinae	*Exochomus quadripustulatus* Linnaeus	Pine ladybird	E4

APPENDIX II

COMPARISON OF COLOUR HISTOGRAMS FOR STANDARD IMAGES

Objective: To determine the range of usability of CIEL*a*b for colour segmentation

Test images:

- mandril
- pepper
- *Coccinella magnifica Redtenbacher* (scarce 7-spot ladybird)
- Fabricated image of *Adalia 2-punctata* and *Harmonia axyridis form spectabilis*
- *Halyzia 16-guttata* (orange ladybird/H16)

Test 1: Mandril

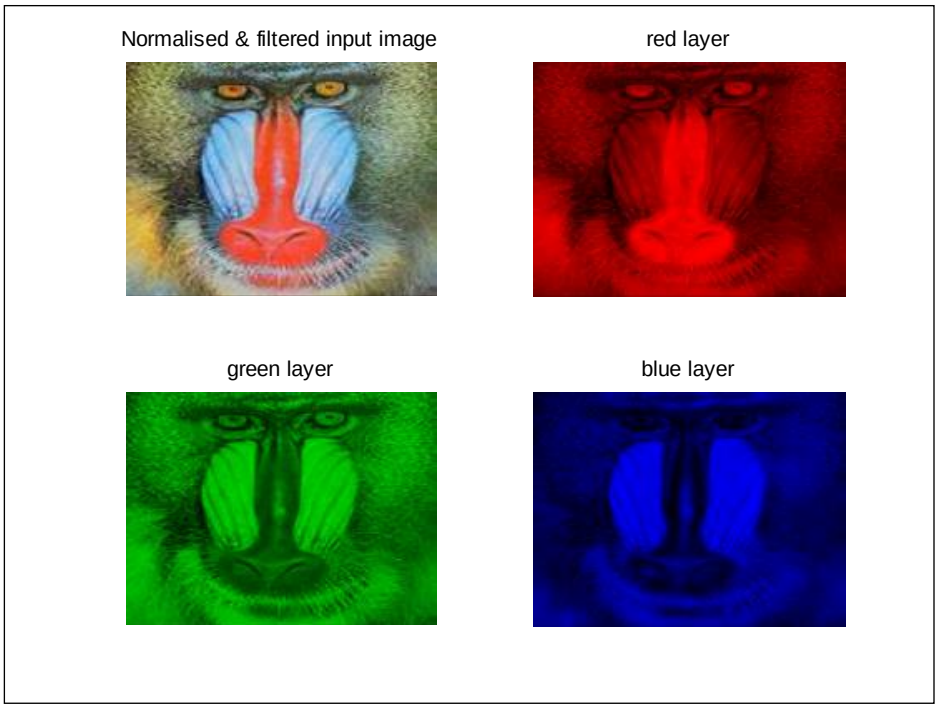


Figure A1: Mandril image in normalised RGB

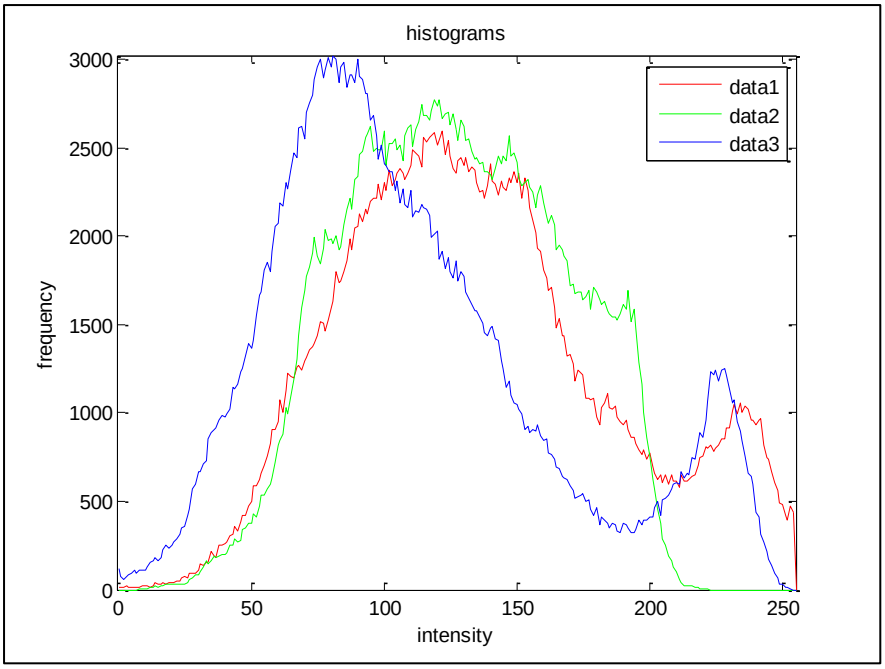


Figure A2: RGB histogram of mandril image showing range of usable intensity values

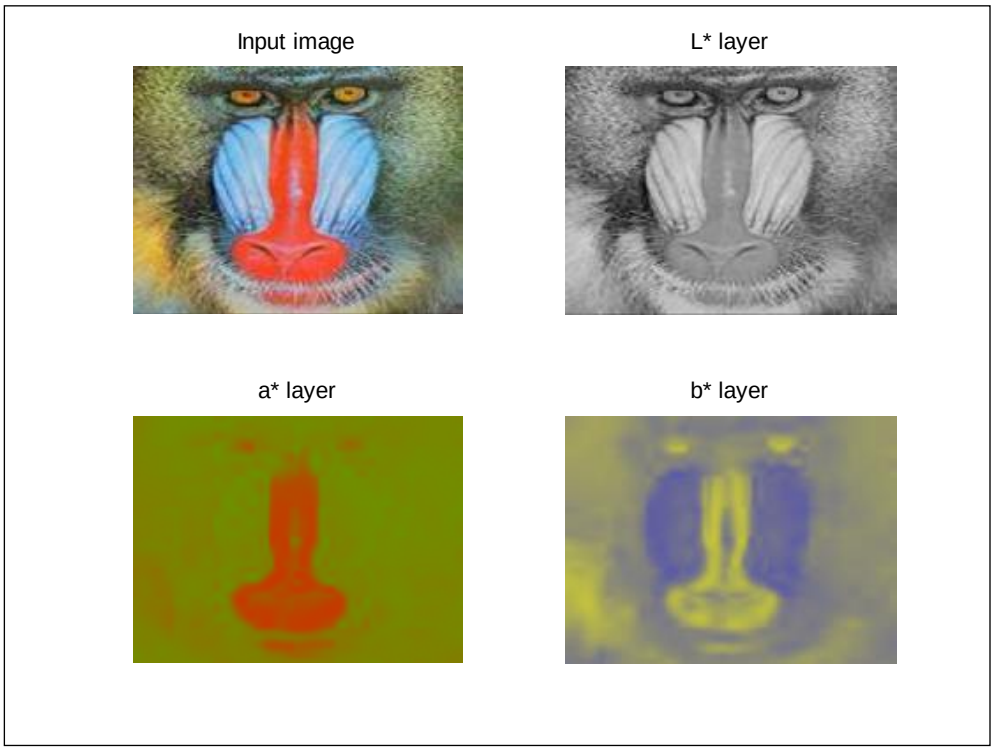


Figure A3: Colour space conversion to CIEL*a*b*

Original Image

Segmented Image

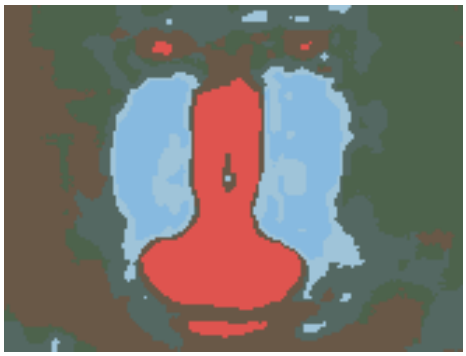


Figure A4: Segmentation via *a channel**

Figure A5: GretagMacbeth colour checker as reference

Figure A6: Binary mandril image after segmentation on channel a* to detect red and green colours only

Test 2: Pepper

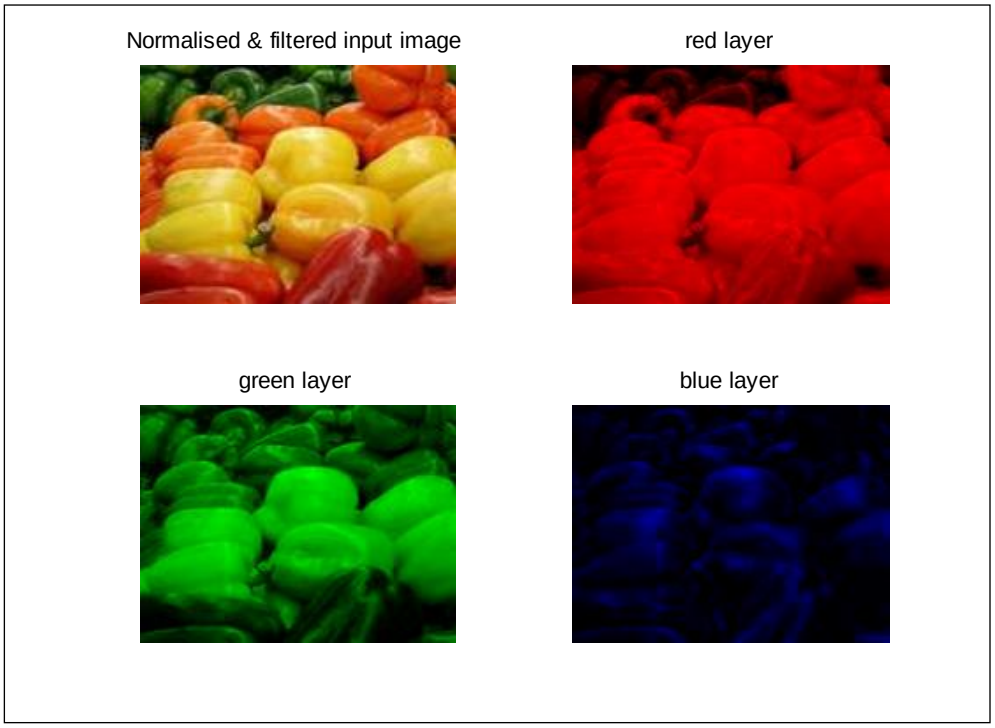


Figure A7: Pepper image in normalised RGB

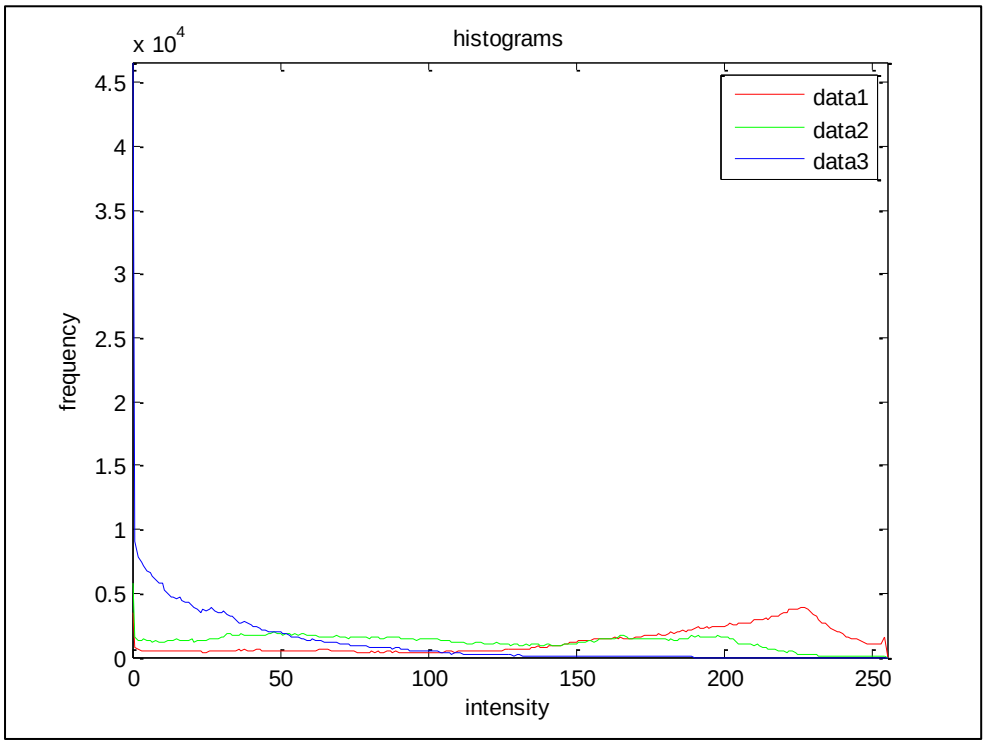


Figure A8: RGB histogram of pepper image showing range of usable intensity values

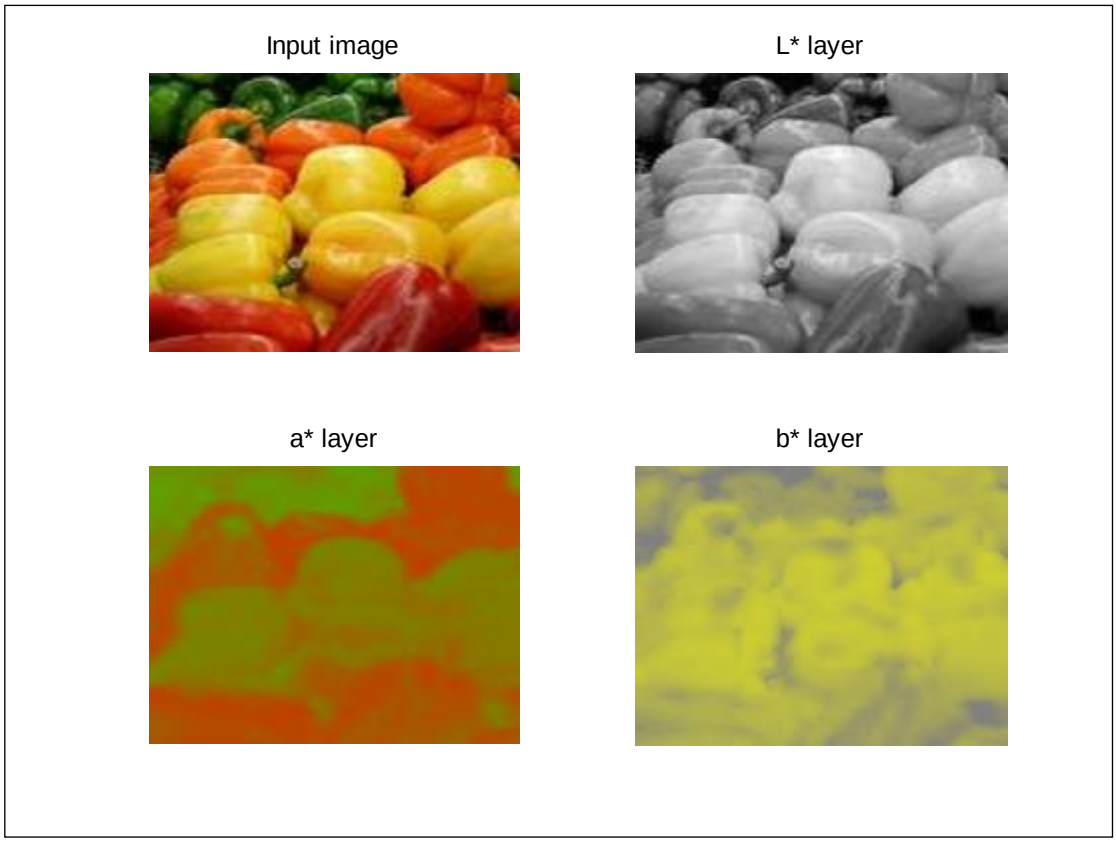


Figure A9: After conversion to CIEL*a*b*

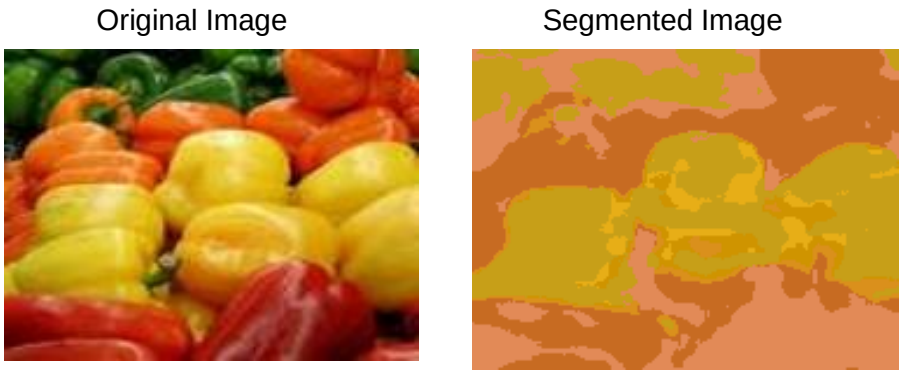


Figure A10: Segmentation via *a channel**

Test 3: **Comparison of colour histograms for scarce 7-spot ladybird**

Scarce 7-spot (with background)

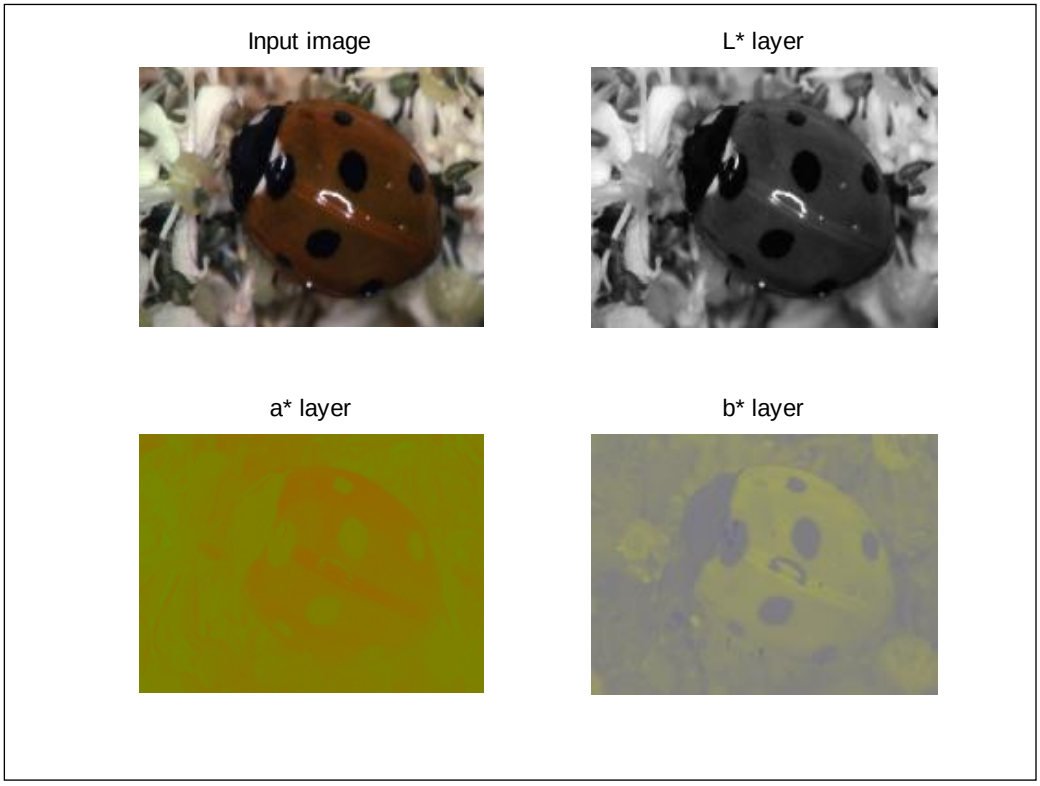


Figure A11: Colour space conversion from RGB to CIEL*a*b*

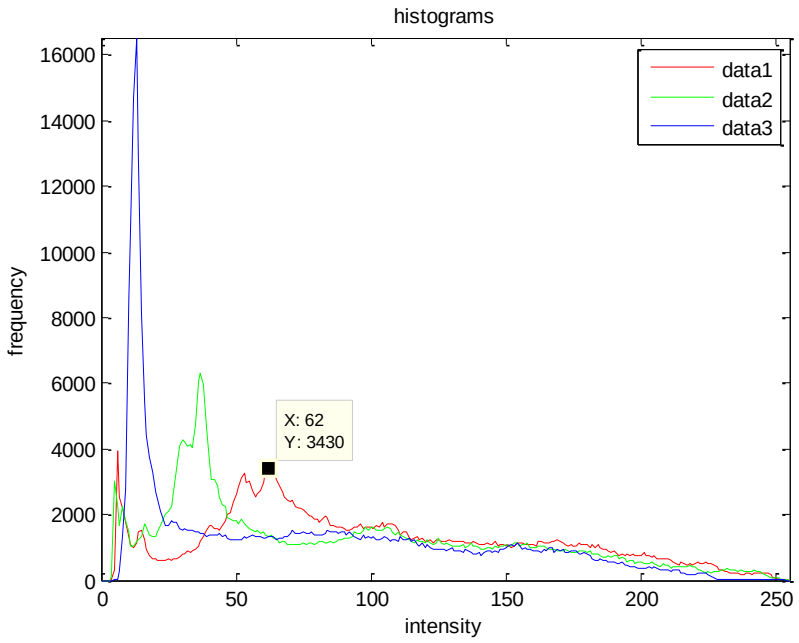


Figure A12: RGB histogram for scarce 7-spot (with background)

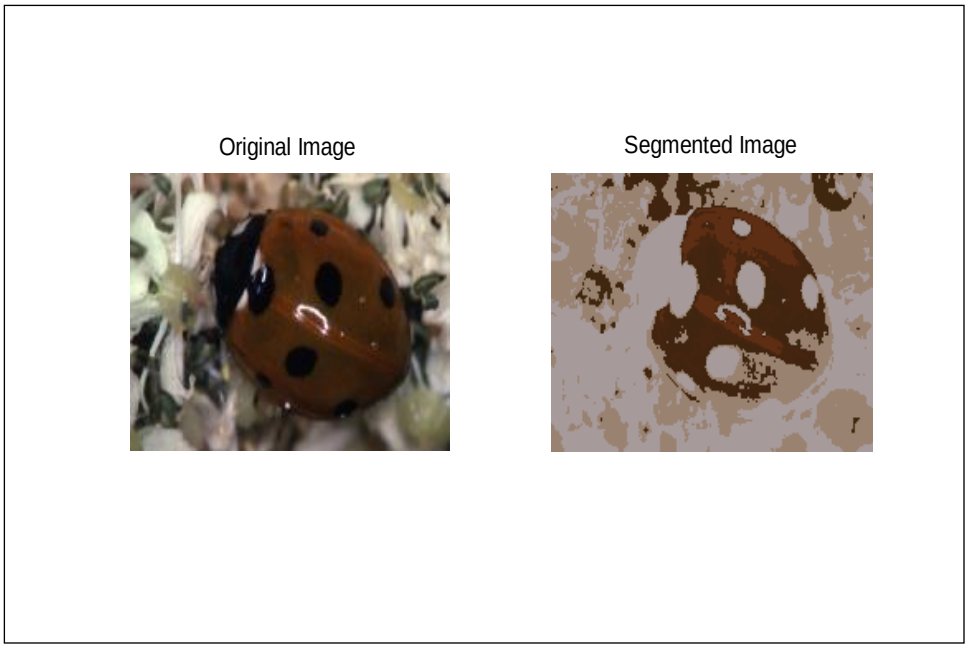


Figure A13: Image of scarce 7-spot (with background) after colour segmentation

Figure A14: Resultant binary image showing complicated background

Scarce 7-spot (without background)

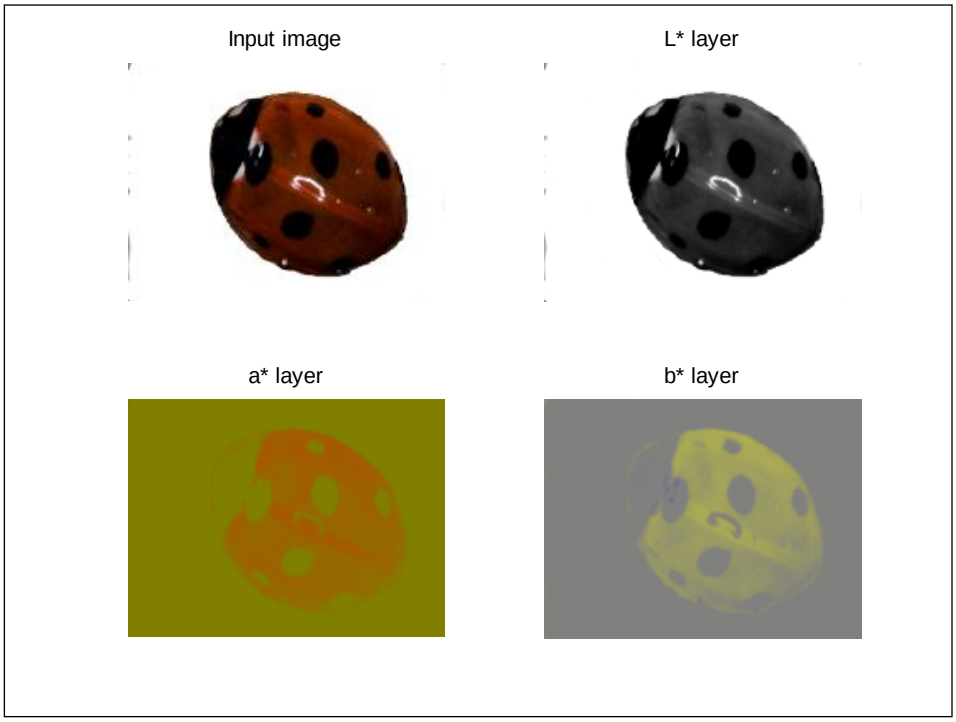


Figure A15: Colour space conversion from RGB to CIEL*a*b*

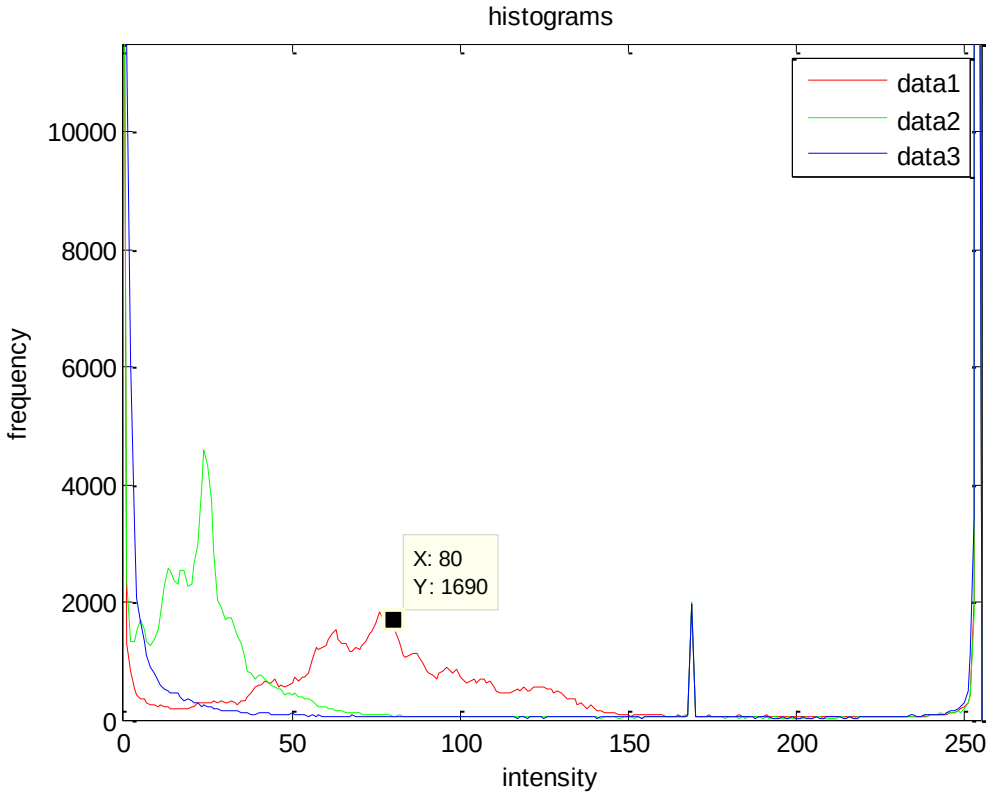


Figure A16: Colour histogram for scarce 7-spot (without background)

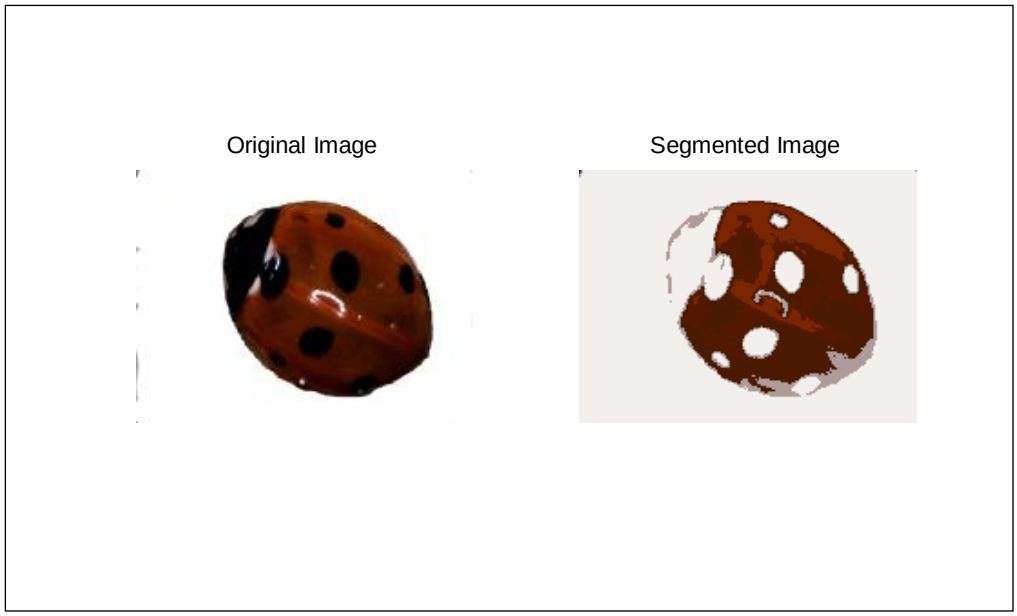


Figure A17: After colour segmentation showing rough segments of reddish colour

Test on ellipsoids

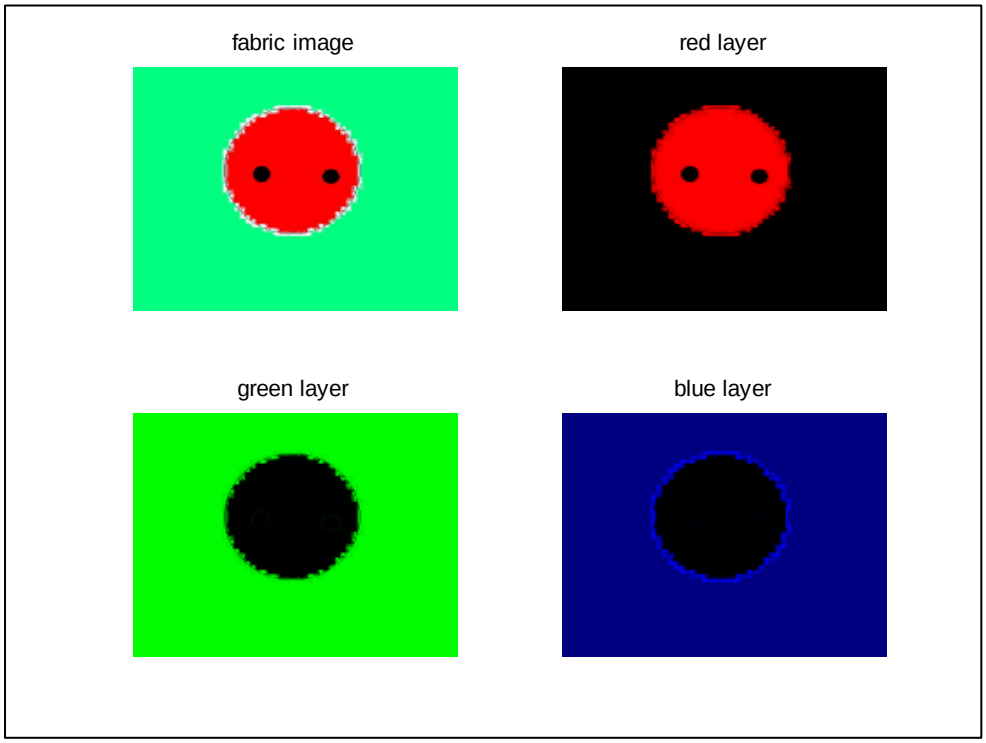


Figure A18: Fabricated 2-spot image in normalised RGB

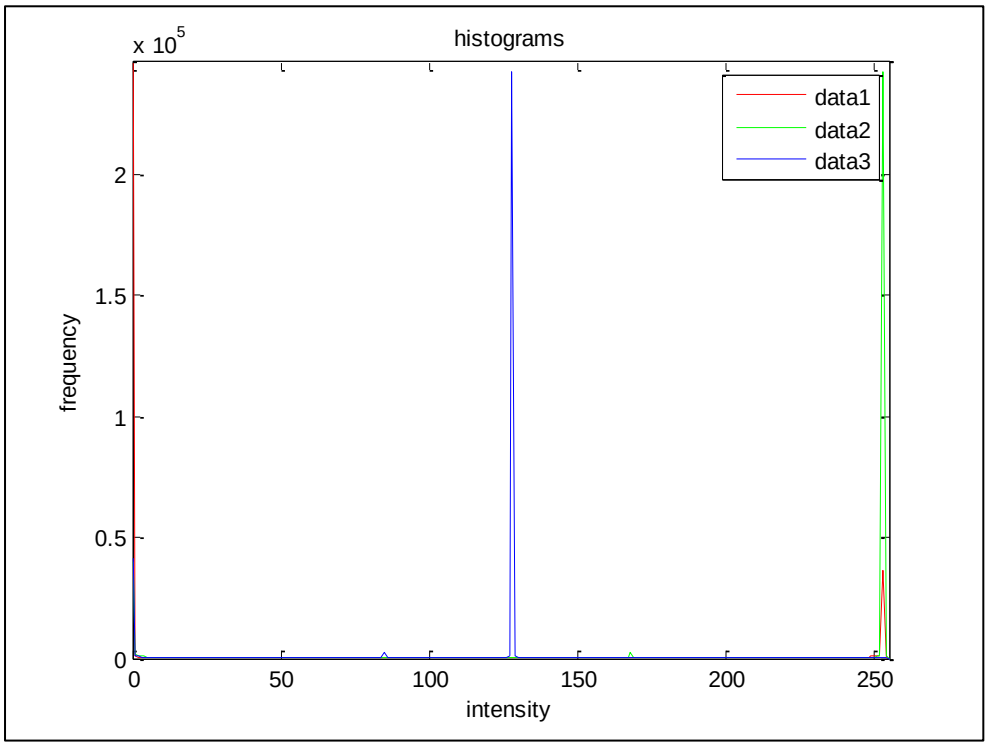


Figure A19: Colour histogram of fabricated 2-spot image (with background)

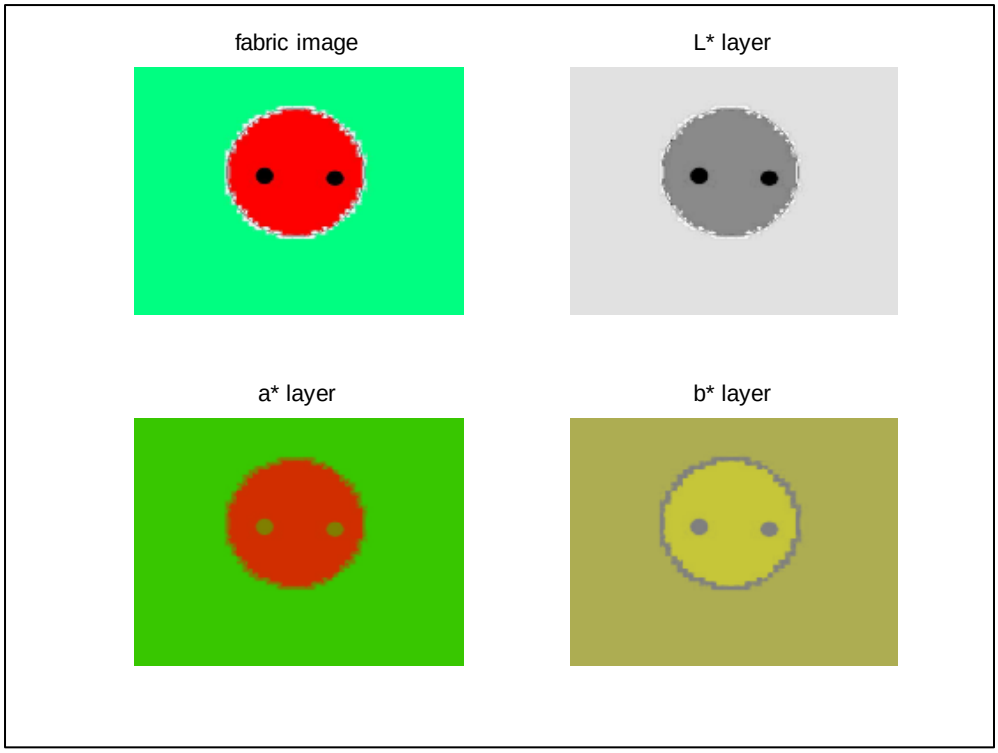


Figure A20: Fabricated 2-spot image in CIELAB

Without background:

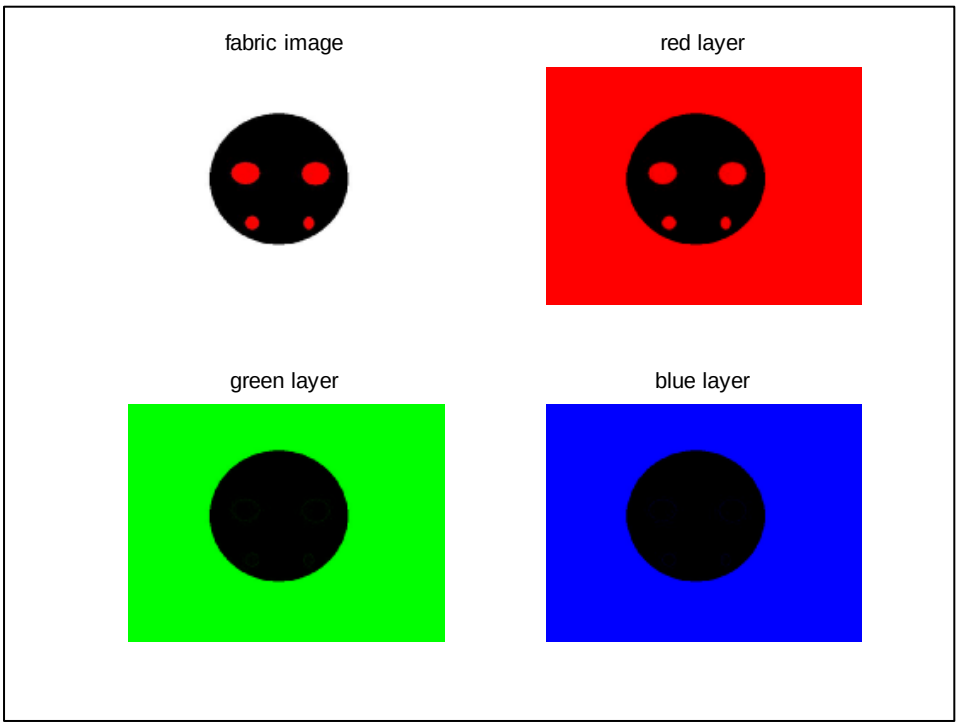


Figure A21: Fabricated H1 image in normalised RGB

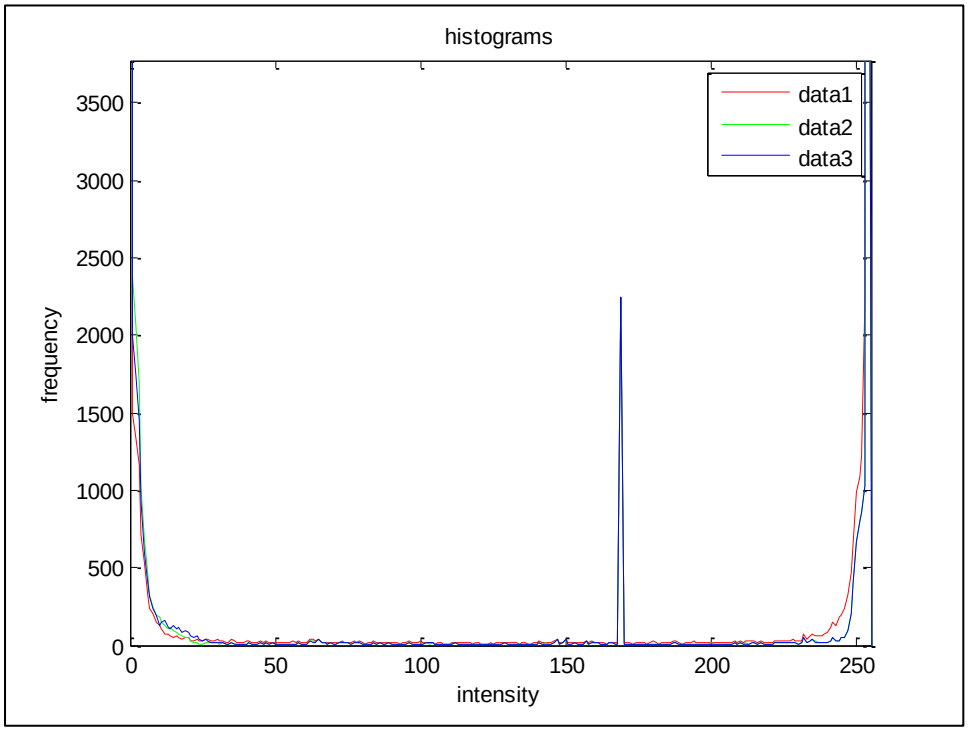


Figure A22: Colour histogram of fabricated 2-spot image (without background)

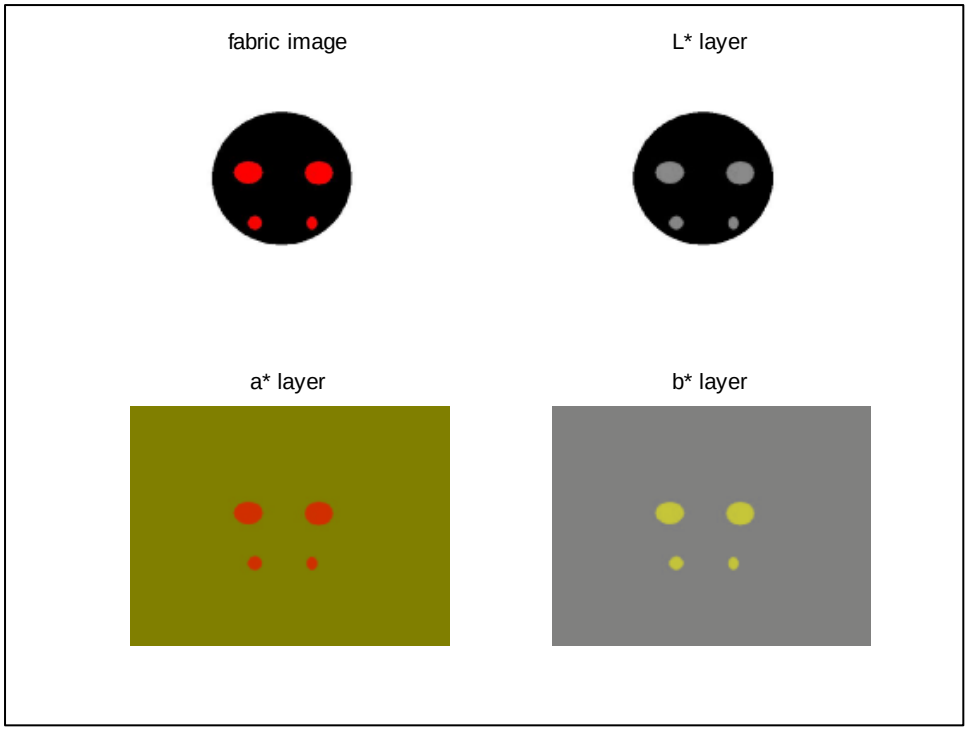


Figure A23: Fabricated H1 image in CIELAB

Test on Orange ladybird (H16)

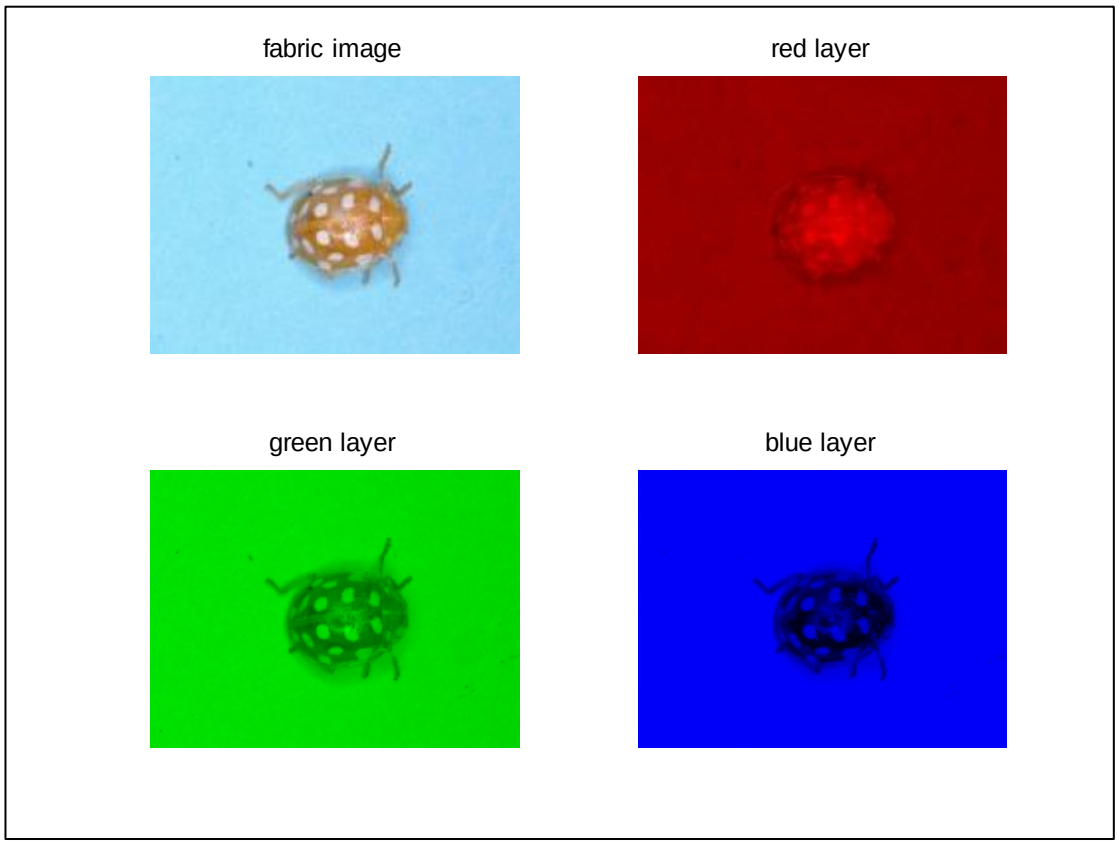


Figure A24: H16 image in normalised RGB

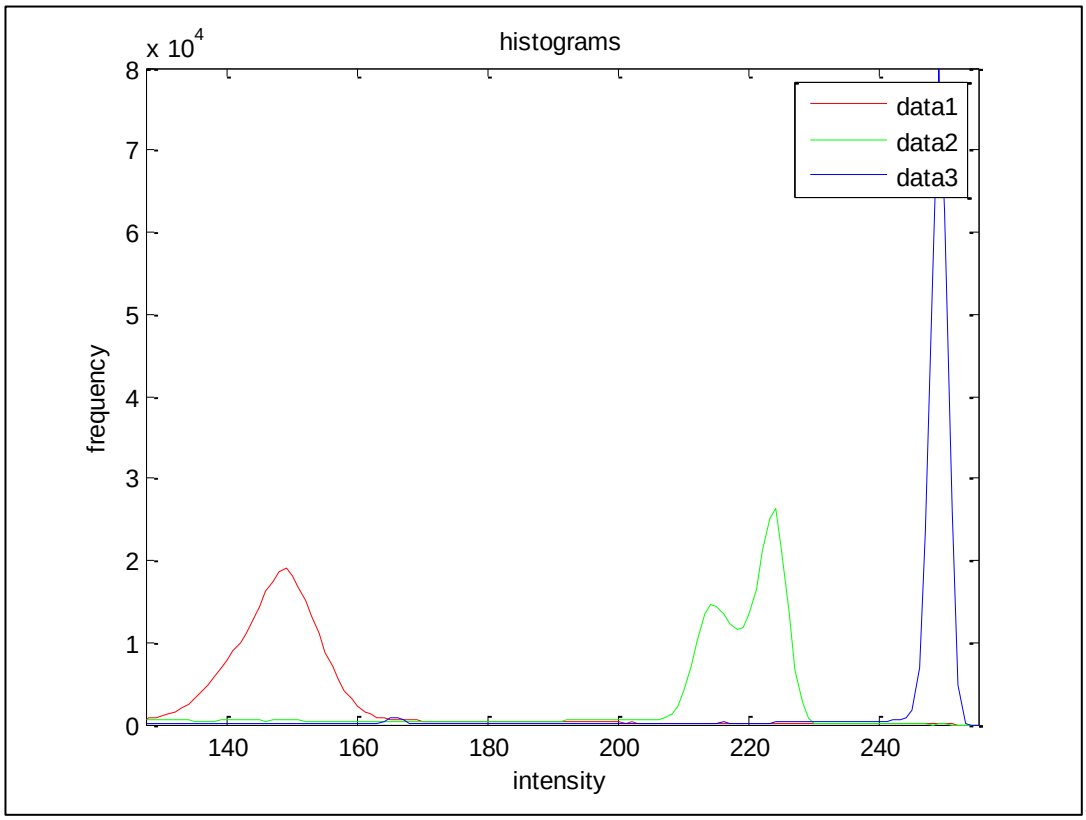


Figure A25: Histogram of H16 image in RGB

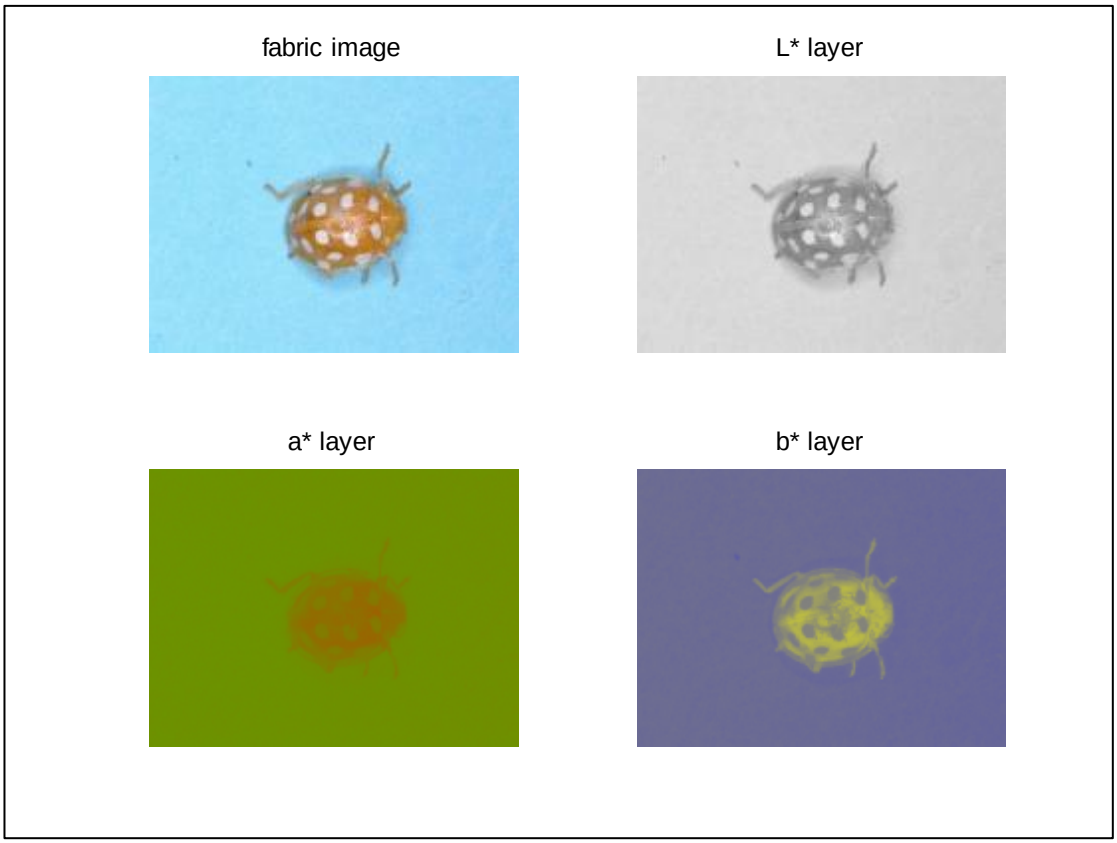


Figure A26: H16 image in CIELAB

H16 with elytra cutout

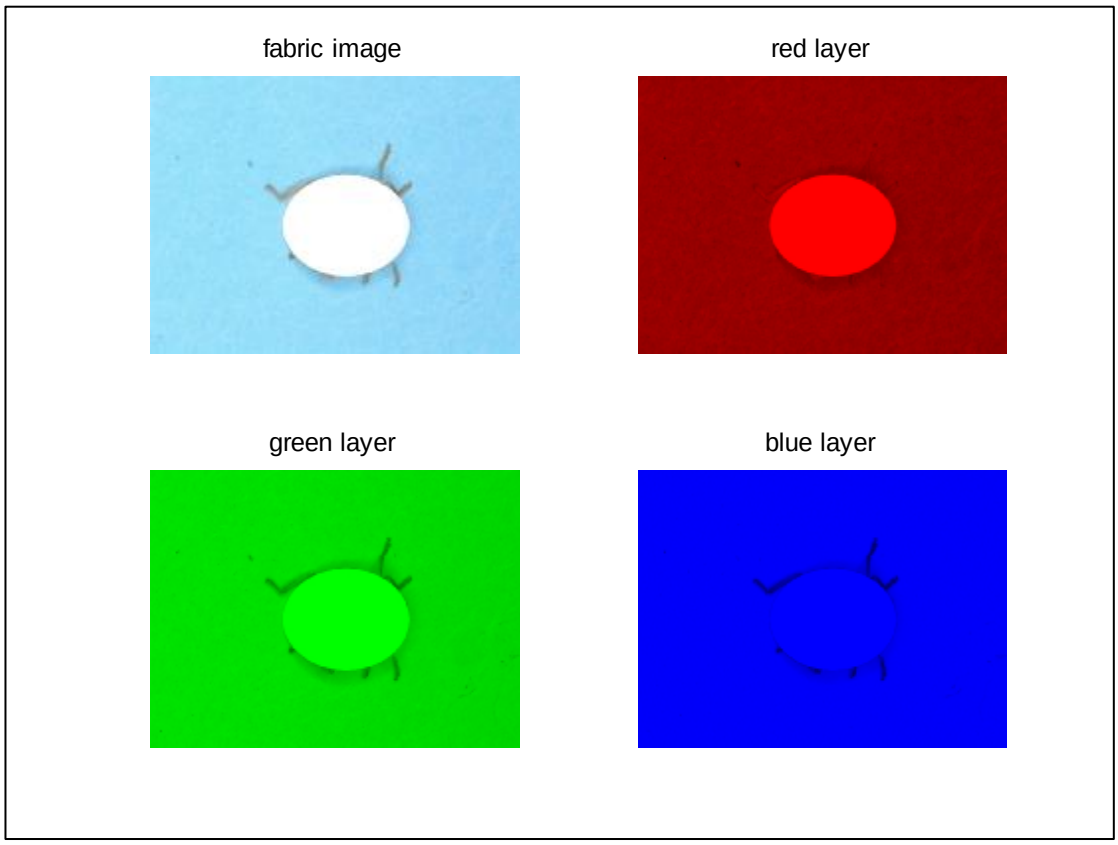


Figure A27: H16 image (elytra cutout) in normalised RGB

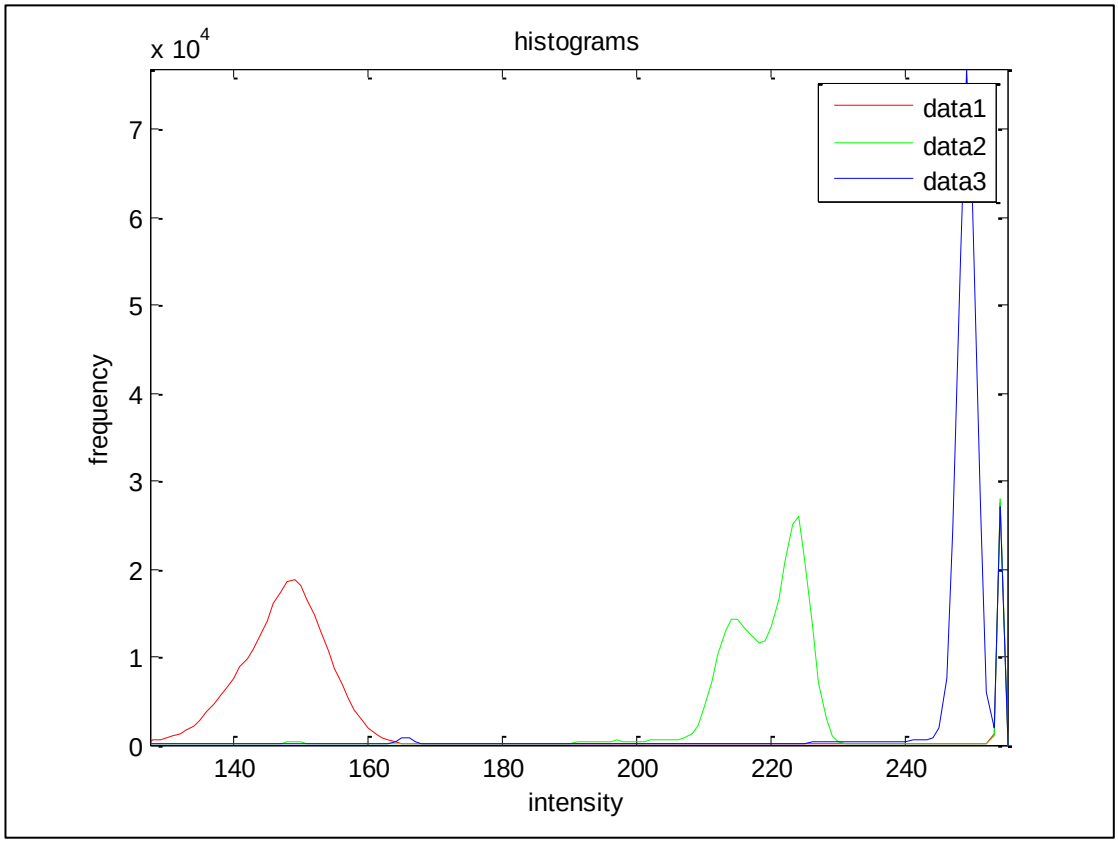


Figure A28: Histogram of H16 image (elytra cutout) in RGB

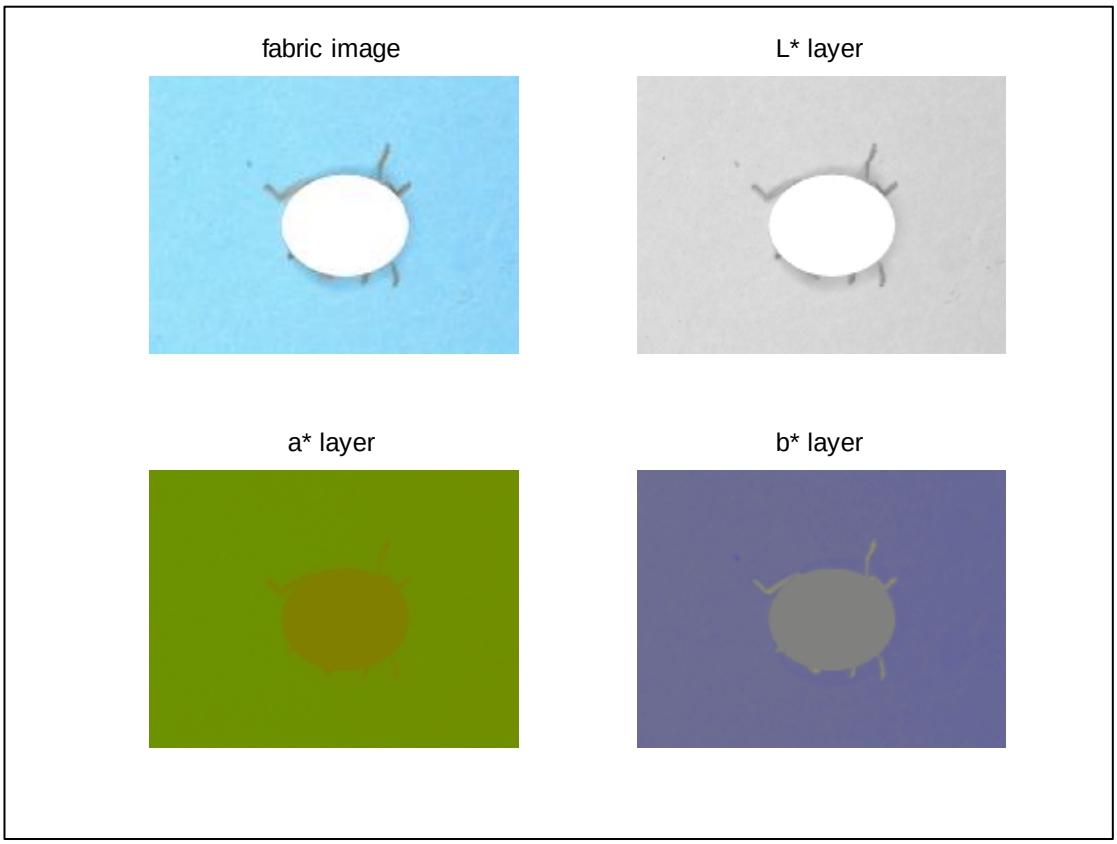


Figure A29: H16 image (elytra cutout) in CIELAB

Overall Observations:

1. Referring to **Figure A4** and **Figure A7**, red colour segmentation works in Mandril image, of which values for thresholding obtained from visual checking using GretagMacbeth colour checker.
2. In **Figure A11**, yellow colours can be segmented using *'b*'* channel.
3. Images with complicated backgrounds are difficult to segment from the object, and vice versa. Perhaps more data on the chrominance density function and the use of colour metrics (eg. Euclidean or Mahalanobis distance) may be able to reduce errors.

APPENDIX III

DISSIMILARITY COEFFICIENTS CALCULATIONS

Intra-OTU dissimilarity calculations

|H1-H2|	|H1-H3|	|H2-H3|							
0.085016	0.267334	0.201164							
0.115587	0.347429	0.236818							
0.071736	0.205272	0.133576							
0.04223	0.161891	0.172923							
0.146103	0.196719	0.133638							
0.107106	0.258153	0.193314							
0.242085	0.33536	0.116256							
0.132619	0.180451	0.077195							
0.30681	0.343052	0.039863							
0.216227	0.265717	0.183066							
0.219725	0.27278	0.053311							
0.169409	0.141174	0.301283							
0.157583	0.294636	0.167918							
0.137004	0.268601	0.135321							
0.077803	0.173511	0.115305							
0.192553	0.314481	0.125437							
0.256877	0.348586	0.093111							
0.138246	0.238974	0.156628							
0.211521	0.327761	0.157898							
0.096021	0.28567	0.190283							
0.058173	0.224436	0.169047							
0.22145	0.361573	0.148572							
0.209817	0.269135	0.061483							
0.259595	0.300903	0.074435							
0.03126	0.164571	0.164936							
0.210761	0.2827	0.073182							
0.24819	0.344297	0.159287							
0.0885	0.140719	0.197583							
0.064274	0.153787	0.214045							
0.092472	0.139495	0.231438							
0.104408	0.232845	0.131855							
0.237433	0.334523	0.167728							
0.29793	0.385673	0.089899							
0.250058	0.238044	0.270529							
0.042077	0.195959	0.158459							
0.197008	0.314708	0.119333							
0.057272	0.214678	0.16026							
0.135485	0.165437	0.034059							
0.03626	0.159535	0.123341							
0.188884	0.325022	0.136885							
0.153839	0.25439	0.146767		**POP. AVERAGE**		0.184999	10-SAMPLES AV(	0.183824	
0.078878	0.072788	0.059749						0.085558	
0.006222	0.005298	0.00357							

Inter-OTU dissimilarity calculations

|C5-E4|	|C5-C14|	|C5-H16|	|C5-H1|	|C5-H2|	|C5-H3|	|C7-E4|	|C7-C14|	|C7-H16|	|E4-C14|	|C14-H16|
0.360954	0.015008	0.005394	0.266582	0.198532	0.160181	0.25977	0.088022	0.097093	0.34733	0.009811
0.300248	0.105495	0.033733	0.351423	0.240983	0.169913	0.306458	0.115796	0.044848	0.308999	0.083517
0.389279	0.009669	0.046787	0.203481	0.131985	0.08019	0.355672	0.047311	0.017759	0.384276	0.038174
0.229965	0.065815	0.062878	0.15538	0.114406	0.116641	0.214781	0.051142	0.049174	0.1644	0.008748
0.249738	0.078918	0.084664	0.215302	0.147686	0.033978	0.22673	0.074537	0.061651	0.179244	0.013811
0.270266	0.045697	0.040738	0.280542	0.208936	0.138208	0.257928	0.0215	0.052545	0.237313	0.052254
0.294306	0.130253	0.009913	0.329845	0.105229	0.076751	0.255615	0.088107	0.038175	0.182454	0.125848
0.381934	0.043741	0.015199	0.173326	0.059897	0.035752	0.381189	0.043084	0.014122	0.338413	0.03385
0.393609	0.003202	0.009009	0.356035	0.060261	0.033463	0.388272	0.008373	0.016108	0.39611	0.00993
0.392864	0.00313	0.001712	0.268169	0.187749	0.105879	0.392831	0.003225	0.0013	0.395886	0.001942
0.326316	0.050093	0.031003	0.260009	0.145566	0.095096	0.303925	0.05411	0.039277	0.293442	0.037788
0.064298	0.044599	0.027767	0.071988	0.062457	0.051518	0.069903	0.037192	0.028332	0.094055	0.039906
0.004134	0.001989	0.000771	0.005182	0.003901	0.002654	0.004886	0.001383	0.000803	0.008846	0.001593
POP. AVE	0.133717									
	0.125931									

APPENDIX IV

TEST DATA SET AND CLASSIFIER RESULTS

Table A2: Test data set for White-spotted group

Spot col. (a*)	Spot col. (b*)	Spot Hue Angle	BG col. (a*)	BG col. (b*)	Base Hue Angle	Area	Perim	Maj Axis	Min Axis	Area Ratio	Asp. Ratio	Species
0.50	0.51	0.04	0.6	0.78	0.08	0.02	0.14	0.0	0.08	0.68	0.6	C14
0.49	0.51	-	0.5	0.58	0.11	0.02	0.11	0.1	0.06	0.69	0.55	C14
0.51	0.54	0.04	0.6	0.63	0.11	0.02	0.1	0.0	0.07	0.76	0.66	C14
0.55	0.62	0.07	0.6	0.59	0.14	0.01	0.07	0.0	0.05	0.76	0.68	C14
0.49	0.51	-	0.5	0.58	0.11	0.02	0.11	0.1	0.06	0.69	0.55	C14
0.52	0.56	0.06	0.6	0.67	0.12	0.02	0.09	0.0	0.06	0.7	0.62	C14
0.55	0.62	0.07	0.6	0.59	0.14	0.01	0.07	0.0	0.05	0.76	0.68	C14
0.54	0.53	0.14	0.5	0.56	0.14	0.01	0.08	0.0	0.06	0.72	0.59	C14
0.49	0.50	-	0.5	0.76	0.05	0.06	0.11	0.0	0.29	0.73	0.56	H16
0.51	0.53	0.05	0.6	0.75	0.09	0.12	0.27	0.1	0.12	0.54	0.46	H16
0.50	0.55	0.02	0.6	0.67	0.1	0.09	0.15	0.1	0.12	0.7	0.55	H16
0.49	0.50	-	0.5	0.76	0.05	0.06	0.11	0.0	0.29	0.73	0.56	H16
0.49	0.50	-	0.5	0.82	0.03	0.05	0.16	0.0	0.07	0.59	0.54	H16
0.54	0.50	0.24	0.6	0.74	0.08	0.15	0.18	0.2	0.17	0.65	0.59	H16
0.51	0.56	0.04	0.6	0.71	0.09	0.1	0.18	0.1	0.15	0.62	0.51	H16
0.49	0.50	-	0.5	0.78	0.05	0.09	0.14	0.1	0.11	0.66	0.48	H16

Table A3: Test data set for Red-spotted group

Spot col. (a*)	Spot col. (b*)	Spot Hue Angle	BG col. (a*)	BG col. (b*)	Base Hue Angle	Area	Perim	Maj Axis	Min Axis	AreaRatio	Asp. Ratio	Species
0.764	0.748	0.13	0.496	0.511	-	0.0	0.1	0.0	0.0	0.7	0.6	E4
0.772	0.627	0.18	0.530	0.435	-	0.0	0.1	0.1	0.0	0.7	0.4	E4
0.691	0.662	0.14	0.500	0.505	0.01	0.0	0.1	0.1	0.0	0.6	0.4	E4
0.704	0.680	0.13	0.498	0.505	-	0.0	0.1	0.1	0.0	0.6	0.5	E4
0.630	0.687	0.1	0.5	0.5	0	0.0	0.1	0.0	0.0	0.7	0.7	E4
0.704	0.680	0.13	0.498	0.505	-	0.0	0.1	0.1	0.0	0.6	0.5	E4
0.691	0.662	0.14	0.500	0.505	0.01	0.0	0.1	0.1	0.0	0.6	0.4	E4
0.704	0.680	0.13	0.498	0.505	-	0.0	0.1	0.1	0.0	0.6	0.5	E4
0.619	0.590	0.15	0.502	0.501	0.18	0.2	0.3	0.3	0.3	0.6	0.8	H1
0.622	0.680	0.09	0.501	0.497	-	0.1	0.3	0.3	0.3	0.6	0.8	H1
0.645	0.739	0.09	0.486	0.528	-	0.4	0.5	0.5	0.4	0.5	0.7	H1
0.580	0.798	0.04	0.510	0.480	-	0.2	0.3	0.3	0.3	0.6	0.9	H1
0.661	0.674	0.12	0.5	0.5	0.25	0.2	0.3	0.3	0.3	0.6	0.8	H1
0.748	0.750	0.12	0.504	0.508	0.07	0.1	0.2	0.2	0.2	0.6	0.8	H1
0.523	0.802	0.01	0.5	0.5	0	0.2	0.3	0.3	0.3	0.6	0.6	H1
0.545	0.794	0.02	0.495	0.511	-	0.4	0.6	0.6	0.5	0.4	0.6	H1
0.601	0.587	0.14	0.491	0.510	-	0.2	0.5	0.5	0.5	0.4	0.7	H2
0.642	0.535	0.21	0.492	0.454	0.03	0.1	0.2	0.2	0.2	0.6	0.7	H2
0.709	0.719	0.12	0.496	0.503	-	0.2	0.2	0.2	0.2	0.6	0.6	H2
0.595	0.618	0.11	0.503	0.488	-	0.2	0.2	0.2	0.2	0.6	0.7	H2
0.610	0.645	0.1	0.493	0.485	0.06	0.4	0.5	0.5	0.5	0.4	0.4	H2
0.636	0.572	0.17	0.498	0.504	-	0.4	0.6	0.6	0.5	0.4	0.6	H2
0.748	0.693	0.14	0.507	0.496	-	0.1	0.3	0.2	0.2	0.6	0.7	H2
0.605	0.56	0.17	0.479	0.475	0.11	0.0	0.1	0.1	0.0	0.7	0.7	H2

Table A4: Test data set for Black-spotted group

Spot col. (a*)	Spot col. (b*)	Spot Hue Angle	BG col. (a*)	BG col. (b*)	BaseHue Ang	Area	Perim	Maj Axis	Min Axis	AreaRatio	Asp. Ratio	Species
0.478	0.450	0.07	0.570	0.595	0.1	0.1	0.4	0.5	0.3	0.6	0.4	A2
0.499	0.503	-	0.627	0.618	0.1	0.2	0.5	0.5	0.4	0.5	0.5	A2
0.503	0.499	-	0.651	0.636	0.1	0.1	0.5	0.3	0.6	0.4	0.7	A2
0.497	0.485	0.03	0.749	0.798	0.1	0.2	0.5	0.6	0.3	0.5	0.5	A2
0.502	0.500	0.22	0.667	0.654	0.1	0.1	0.4	0.4	0.5	0.5	0.6	A2
0.508	0.513	0.09	0.628	0.615	0.1	0.1	0.4	0.3	0.5	0.4	0.6	A2
0.484	0.449	0.05	0.514	0.532	0.0	0.1	0.4	0.6	0.2	0.5	0.4	A2
0.502	0.500	0.22	0.607	0.595	0.1	0.1	0.4	0.3	0.6	0.5	0.8	A2
0.498	0.504	-	0.693	0.685	0.1	0.6	0.8	0.8	0.8	0.6	0.8	C5
0.506	0.495	-	0.676	0.655	0.1	0.4	0.5	0.5	0.5	0.6	0.6	C5
0.498	0.504	-	0.693	0.685	0.1	0.6	0.8	0.8	0.8	0.6	0.8	C5
0.488	0.505	-	0.770	0.717	0.1	0.2	0.3	0.3	0.2	0.6	0.5	C5
0.498	0.503	-	0.642	0.642	0.1	0.2	0.4	0.4	0.4	0.7	0.9	C5
0.506	0.495	-	0.676	0.655	0.1	0.4	0.5	0.5	0.5	0.6	0.6	C5
0.488	0.505	-	0.770	0.717	0.1	0.2	0.3	0.3	0.2	0.6	0.5	C5
0.499	0.507	-	0.667	0.675	0.1	0.4	0.6	0.6	0.7	0.7	0.9	C5
0.501	0.515	0.01	0.720	0.773	0.1	0.4	0.4	0.4	0.4	0.6	0.8	C7
0.52	0.544	0.07	0.713	0.795	0.1	0.2	0.2	0.3	0.2	0.7	0.7	C7
0.498	0.508	-	0.744	0.744	0.1	0.7	0.9	0.9	0.8	0.6	0.7	C7
0.575	0.570	0.13	0.702	0.713	0.1	0.5	0.7	0.7	0.8	0.6	0.8	C7
0.499	0.523	-	0.725	0.792	0.1	0.7	0.8	0.8	0.8	0.6	0.7	C7
0.52	0.544	0.07	0.713	0.795	0.1	0.2	0.2	0.3	0.2	0.7	0.7	C7
0.498	0.516	-	0.729	0.734	0.1	0.2	0.3	0.3	0.3	0.7	0.8	C7
0.514	0.512	0.13	0.779	0.735	0.1	0.3	0.4	0.4	0.4	0.6	0.8	C7
0.582	0.748	0.05	0.500	0.504	0.0	0.2	0.4	0.4	0.4	0.6	0.7	H3
0.494	0.512	-	0.533	0.534	0.1	0.2	0.4	0.4	0.4	0.4	0.6	H3
0.503	0.509	0.05	0.538	0.571	0.0	0.2	0.7	0.8	0.6	0.2	0.7	H3
0.497	0.525	-	0.649	0.684	0.1	0.1	0.2	0.2	0.2	0.6	0.9	H3
0.520	0.490	-	0.620	0.658	0.1	0.0	0.1	0.1	0.0	0.7	0.6	H3
0.501	0.497	-	0.650	0.640	0.1	0.6	0.8	0.8	0.8	0.6	0.8	H3
0.493	0.512	-	0.737	0.736	0.1	0.2	0.3	0.3	0.3	0.6	0.9	H3
0.518	0.457	-	0.574	0.539	0.1	0.5	0.7	0.7	0.7	0.6	0.9	H3

CLASSIFIER RESULTS

6.2.4.2 SVM using SMO algorithm: test results

Table 6.11d: Metrics for SVM using SMO (C14H16 white group, unbalanced class, all features)

Class	TP rate	FP rate	Precision	Recall	AUC
C14	0.9	0	1	0.9	0.95
H16	1	0.1	0.976	1	0.95
Weighted average	0.98	0.08	0.98	0.98	0.95

Table 6.11e: Metrics for SVM using SMO (C14H16 white group, unbalanced class, colour features)

Class	TP rate	FP rate	Precision	Recall	AUC
C14	0.6	0	1	0.6	0.8
H16	1	0.4	0.909	1	0.8
Weighted average	0.92	0.32	0.927	0.92	0.8

Table 6.11f: Metrics for SVM using SMO (C14H16 white group, unbalanced class, geometrical features)

Class	TP rate	FP rate	Precision	Recall	AUC
C14	1	0	1	1	1
H16	1	0	1	1	1
Weighted average	1	0	1	1	1

Balanced class

Table 6.12d: Metrics for SVM using SMO (C14H16 white group, balanced class, all features)

Class	TP rate	FP rate	Precision	Recall	AUC
C14	1	0	1	1	1
H16	1	0	1	1	1
Weighted average	1	0	1	1	1

Table 6.12e: Metrics for SVM using SMO (C14H16 white group, balanced class, colour features)

Class	TP rate	FP rate	Precision	Recall	AUC
C14	0.8	0	1	0.8	0.9
H16	1	0.2	0.833	1	0.9
Weighted average	0.9	0.1	0.917	0.9	0.9

Red-spotted group: Balanced class

Table 6.13d: Metrics for SVM using SMO (E4H1H2 red group, balanced class, all features)

Class	TP rate	FP rate	Precision	Recall	AUC
E4	1	0.05	0.909	1	0.981
H1	0.625	0.138	0.694	0.625	0.744
H2	0.675	0.163	0.675	0.675	0.816
Weighted average	0.767	0.117	0.76	0.767	0.847

Table 6.13e: Metrics for SVM using SMO (E4H1H2 red group, balanced class, colour features)

Class	TP rate	FP rate	Precision	Recall	AUC
E4	0.7	0.15	0.7	0.7	0.838
H1	0.4	0.225	0.471	0.4	0.588
H2	0.8	0.175	0.696	0.8	0.844
Weighted average	0.633	0.183	0.622	0.633	0.757

Table 6.13f: Metrics for SVM using SMO
(E4H1H2 red group, balanced class, geometrical features)

Class	TP rate	FP rate	Precision	Recall	AUC
C4	1	0.113	0.816	1	0.944
H1	0.7	0.313	0.528	0.7	0.694
H2	0.275	0.088	0.611	0.275	0.745
Weighted average	0.658	0.171	0.652	0.658	0.794

Black-spotted group: Balanced class

Table 6.14d: Metrics for SVM using SMO
(A2C5C7H3 black group, balanced class, all features)

Class	TP rate	FP rate	Precision	Recall	AUC
A2	1	0.117	0.741	1	0.95
C5	0.8	0.175	0.604	0.8	0.806
C7	0.9	0.05	0.857	0.9	0.963
H3	0.275	0	1	0.275	0.695
Weighted average	0.744	0.085	0.8	0.744	0.854

Table 6.14e: Metrics for SVM using SMO
(A2C5C7H3 black group, balanced class, colour features)

Class	TP rate	FP rate	Precision	Recall	AUC
A2	0.6	0.092	0.686	0.6	0.748
C5	0.8	0.15	0.64	0.8	0.81
C7	1	0.2	0.625	1	0.769
H3	0.275	0	1	0.275	0.643
Weighted average	0.669	0.11	0.738	0.669	0.775

Table 6.14f: Metrics for SVM using SMO
(A2C5C7H3 black group, balanced class, geometrical features)

Class	TP rate	FP rate	Precision	Recall	AUC
A2	0.975	0.133	0.709	0.975	0.931
C5	0.3	0.208	0.324	0.3	0.585
C7	0.6	0.242	0.453	0.6	0.764
H3	0.15	0.075	0.4	0.15	0.514
Weighted average	0.506	0.165	0.472	0.506	0.699

6.2.5 Tests using Learning Vector Quantisation (LVQ)

White-spotted group

Table 6.15d: Metrics using LVQ (C14H16 white group, unbalanced, all features)

Class	TP rate	FP rate	Precision	Recall	AUC
C14	0.9	0	1	0.9	0.95
H16	1	0.1	0.976	1	0.95
Weighted average	0.98	0.08	0.98	0.98	0.95

Table 6.15e: Metrics using LVQ (C14H16 white group, unbalanced, colour features)

Class	TP rate	FP rate	Precision	Recall	AUC
C14	0.7	0	1	0.7	0.85
H16	1	0.3	0.93	1	0.85
Weighted average	0.94	0.24	0.944	0.94	0.85

White-spotted group : Balanced class

Table 6.16d: Metrics using LVQ (C14H16 white group, balanced, all features)

Class	TP rate	FP rate	Precision	Recall	AUC
C14	1	0	1	1	1
H16	1	0	1	1	1
Weighted average	1	0	1	1	1

Table 6.16e: Metrics using LVQ (C14H16 white group, balanced, colour features)

Class	TP rate	FP rate	Precision	Recall	AUC
C14	0.85	0.05	0.944	0.85	0.9
H16	0.95	0.15	0.864	0.95	0.9
Weighted average	0.9	0.1	0.904	0.9	0.9

Table 6.16f: Metrics using LVQ (C14H16 white group, balanced, geometrical features)

Class	TP rate	FP rate	Precision	Recall	AUC
C14	1	0	1	1	1
H16	1	0	1	1	1
Weighted average	1	0	1	1	1

Note: *Results obtained at learning rate = 0.3, total iterations = 1000, linear decay learning function*

Red-spotted group : Balanced class

Table 6.17d: Metrics using LVQ (E4H1H2 red group, balanced, all features)

Class	TP rate	FP rate	Precision	Recall	AUC
E4	1	0.038	0.93	1	0.981
H1	0.625	0.15	0.676	0.625	0.738
H2	0.7	0.15	0.7	0.7	0.775
Weighted average	0.775	0.113	0.769	0.775	0.831

Table 6.17e: Metrics using LVQ (E4H1H2 red group, balanced, colour features)

Class	TP rate	FP rate	Precision	Recall	AUC
E4	0.975	0.125	0.796	0.975	0.925
H1	0.625	0.125	0.714	0.625	0.75
H2	0.675	0.113	0.75	0.675	0.781
Weighted average	0.758	0.121	0.753	0.758	0.819

Table 6.17f: Metrics using LVQ (E4H1H2 red group, balanced, geometrical features)

Class	TP rate	FP rate	Precision	Recall	AUC
E4	1	0.05	0.909	1	0.975
H1	0.575	0.25	0.535	0.575	0.663
H2	0.475	0.175	0.576	0.475	0.65
Weighted average	0.683	0.158	0.673	0.683	0.763

Note: *Results obtained at learning rate = 0.3, total iterations = 1000, linear decay learning function*

Black-spotted group

Table 6.18d: Metrics using LVQ (A2C5C7H3 black group, balanced, all features)

Class	TP rate	FP rate	Precision	Recall	AUC
A2	1	0.05	0.87	1	0.975
C5	0.8	0.15	0.64	0.8	0.825
C7	0.8	0.083	0.762	0.8	0.858
H3	0.375	0.058	0.682	0.375	0.658
Weighted average	0.744	0.085	0.738	0.744	0.829

Table 6.18e: Metrics using LVQ (A2C5C7H3 black group, balanced, colour features)

Class	TP rate	FP rate	Precision	Recall	AUC
A2	0.575	0.083	0.697	0.575	0.746
C5	0.8	0.142	0.653	0.8	0.829
C7	0.9	0.1	0.75	0.9	0.9
H3	0.525	0.075	0.7	0.525	0.725
Weighted average	0.7	0.1	0.7	0.7	0.8

Table 6.18f: Metrics using LVQ (A2C5C7H3 black group, balanced, geometrical features)

Class	TP rate	FP rate	Precision	Recall	AUC
A2	1	0.05	0.87	1	0.975
C5	0.6	0.15	0.571	0.6	0.725
C7	0.675	0.15	0.6	0.675	0.763
H3	0.35	0.108	0.519	0.35	0.621
Weighted average	0.656	0.115	0.64	0.656	0.771

6.2.6 Tests using Probabilistic Neural Network (PNN)

White-spotted group

Table 6.19d: Metrics using PNN (C14H16 white group, unbalanced, all features)

Class	TP rate	FP rate	Precision	Recall	AUC
C14	0.9	0	1	0.9	1
H16	1	0.1	0.976	1	1
Weighted average	0.98	0.08	0.98	0.98	1

Table 6.19e: Metrics using PNN (C14H16 white group, unbalanced, colour features)

Class	TP rate	FP rate	Precision	Recall	AUC
C14	0.9	0	1	0.9	0.93
H16	1	0.1	0.976	1	0.93
Weighted average	0.98	0.08	0.98	0.98	0.93

Table 6.19f: Metrics using PNN (C14H16 white group, unbalanced, geometrical features)

Class	TP rate	FP rate	Precision	Recall	AUC
C14	0.9	0	1	0.9	0.93
H16	1	0.1	0.976	1	0.93
Weighted average	0.98	0.08	0.98	0.98	0.93

Note: *Results obtained at MinStdDev = 0.1, no. of clusters = 2*

White-spotted group : Balanced

Table 6.20d: Metrics using PNN (C14H16 white group, balanced, all features)

Class	TP rate	FP rate	Precision	Recall	AUC
C14	1	0	1	1	1
H16	1	0	1	1	1
Weighted average	1	0	1	1	1

Table 6.20e: Metrics using PNN (C14H16 white group, balanced, colour features)

Class	TP rate	FP rate	Precision	Recall	AUC
C14	0.925	0	1	0.925	0.986
H16	1	0.075	0.93	1	0.986
Weighted average	0.963	0.038	0.965	0.963	0.986

Table 6.20f: Metrics using PNN (C14H16 white group, balanced, geometrical features)

Class	TP rate	FP rate	Precision	Recall	AUC
C14	1	0	1	1	1
H16	1	0	1	1	1
Weighted average	1	0	1	1	1

Note: *Results obtained at MinStdDev = 0.1, no. of clusters = 2*

Red-spotted group

Table 6.21d: Metrics using PNN (E4H1H2 red group, balanced, all features)

Class	TP rate	FP rate	Precision	Recall	AUC
E4	1	0.013	0.976	1	0.993
H1	0.7	0.213	0.622	0.7	0.825
H2	0.575	0.138	0.676	0.575	0.847
Weighted average	0.758	0.121	0.758	0.758	0.888

Table 6.21e: Metrics using PNN (E4H1H2 red group, balanced, colour features)

Class	TP rate	FP rate	Precision	Recall	AUC
E4	0.85	0.138	0.756	0.85	0.932
H1	0.625	0.188	0.625	0.625	0.796
H2	0.675	0.1	0.771	0.675	0.806
Weighted average	0.717	0.142	0.717	0.717	0.845

Table 6.21f: Metrics using PNN ((E4H1H2 red group, balanced, geometrical features)

Class	TP rate	FP rate	Precision	Recall	AUC
E4	1	0.038	0.93	1	0.988
H1	0.625	0.338	0.481	0.625	0.711
H2	0.3	0.163	0.48	0.3	0.724
Weighted average	0.642	0.179	0.63	0.642	0.808

Note: *Results obtained at MinStdDev = 0.1, no. of clusters = 3*

Black-spotted group

Table 6.22d: Metrics using PNN (A2C5C7H3 black group, balanced, all features)

Class	TP rate	FP rate	Precision	Recall	F-measure	AUC
A2	0.975	0	1	0.975	0.987	1
C5	0.85	0.083	0.773	0.85	0.81	0.954
C7	0.975	0	1	0.975	0.987	1
H3	0.75	0.067	0.789	0.75	0.769	0.935
Weighted average	0.888	0.038	0.891	0.888	0.888	0.972

Table 6.22e: Metrics using PNN (A2C5C7H3 black group, balanced, colour features)

Class	TP rate	FP rate	Precision	Recall	F-measure	AUC
A2	0.75	0.042	0.857	0.75	0.8	0.956
C5	0.95	0.05	0.864	0.95	0.905	0.972
C7	0.95	0.017	0.95	0.95	0.95	0.999
H3	0.725	0.1	0.707	0.725	0.716	0.873
Weighted average	0.844	0.052	0.845	0.844	0.843	0.95

Table 6.22f: Metrics using PNN (A2C5C7H3 black group, balanced, geometrical features)

Class	TP rate	FP rate	Precision	Recall	F-measure	AUC
A2	1	0	1	1	1	1
C5	0.675	0.133	0.628	0.675	0.651	0.877
C7	0.8	0.1	0.727	0.8	0.762	0.95
H3	0.525	0.1	0.636	0.525	0.575	0.837
Weighted average	0.75	0.083	0.748	0.75	0.747	0.916

Note: *Results obtained at MinStdDev = 0.1, no. of clusters = 4*

6.7.2 Tests using J48 decision tree

6.7.2.1 Unbalanced class 1:4

Table 6.25d: Metrics for unbalanced class using MLP (all features)

Class	TP rate	FP rate	Precision	Recall	AUC
E4	0	0	0	0	0.5
H1	1	1	0.8	1	0.5
Weighted average	0.8	0.8	0.64	0.8	0.5

Table 6.25e: Metrics for J48 decision tree (all features)

Class	TP rate	FP rate	Precision	Recall	AUC
E4	0.4	0.05	0.667	0.4	0.825
H1	0.95	0.6	0.864	0.95	0.825
Weighted average	0.84	0.49	0.824	0.84	0.825

Table 6.25f: Metrics from confusion matrix

Class	TP rate	FP rate	Precision	Recall	AUC
E4	0.9	0.075	0.75	0.9	0.913
H1	0.925	0.1	0.974	0.925	0.913
Weighted average	0.92	0.095	0.929	0.92	0.913

Table 6.26d: Metrics for MLP (balanced class)

Class	TP rate	FP rate	Precision	Recall	AUC
E4	1	0.05	0.952	1	0.975
H1	0.95	0	1	0.95	0.975
Weighted average	0.975	0.025	0.976	0.975	0.975

Table 6.26e: Metrics for J48 decision tree (balanced class)

Class	TP rate	FP rate	Precision	Recall	AUC
E4	0.925	0.075	0.925	0.925	0.959
H1	0.925	0.075	0.925	0.925	0.959
Weighted average	0.925	0.075	0.925	0.925	0.959

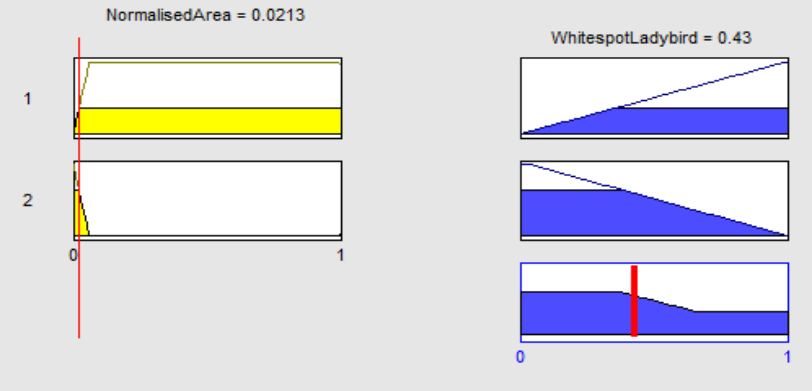
Table 6.26f: Identification metrics for combination of J48 and MLP (balanced class)

Class	TP rate	FP rate	Precision	Recall	AUC
E4	1	0.075	0.93	1	0.963
H1	0.925	0	1	0.925	0.963
Weighted average	0.963	0.038	0.965	0.963	0.963

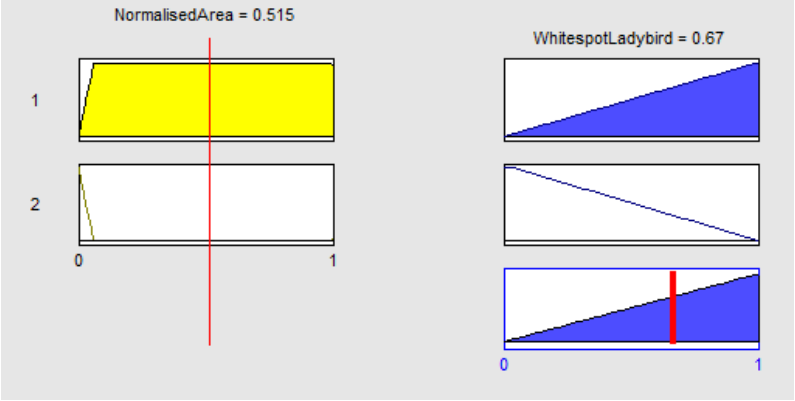
APPENDIX V

SYSTEM INTEGRATION TEST RESULTS

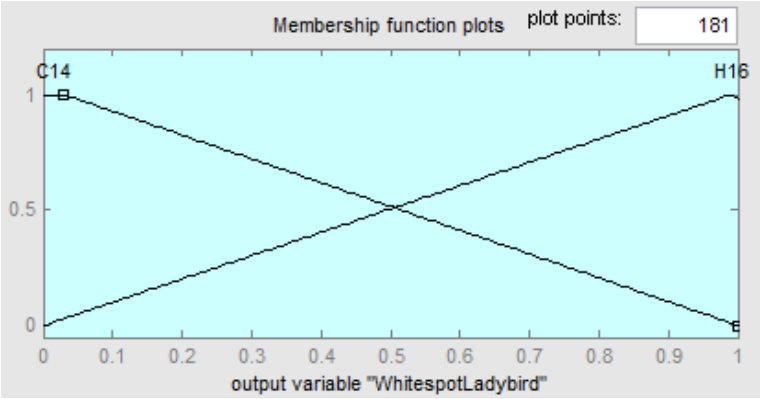
Test results for White-spotted ladybird



Rule viewer for the feature 'NormalisedArea = 0.0213'

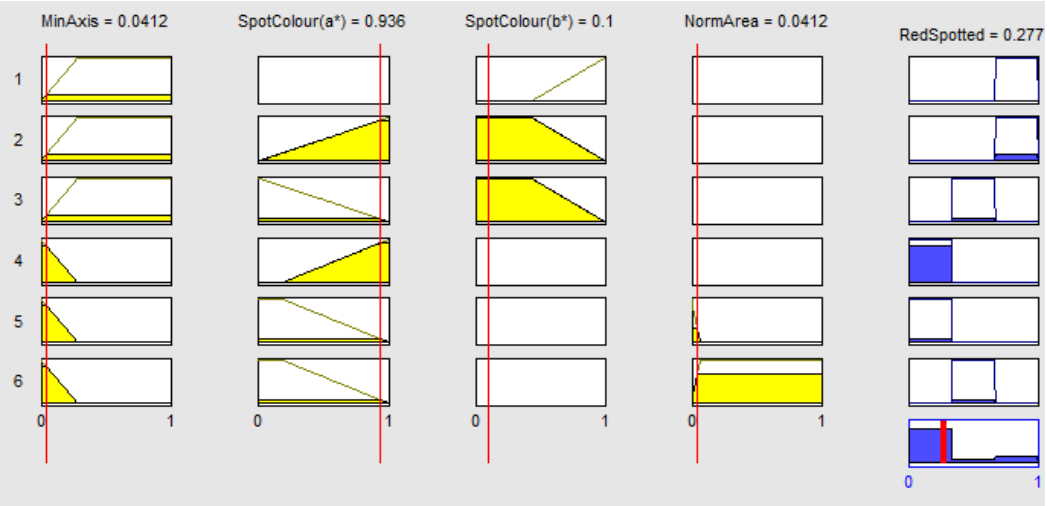


Rule viewer for the feature 'NormalisedArea = 0.515'

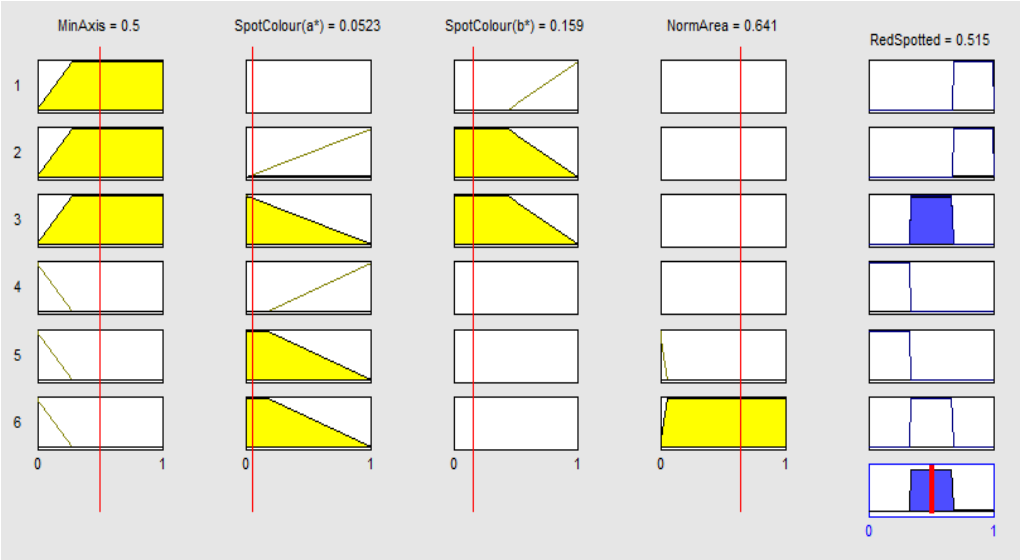


Membership functions for output variables
(White-spotted ladybird group)

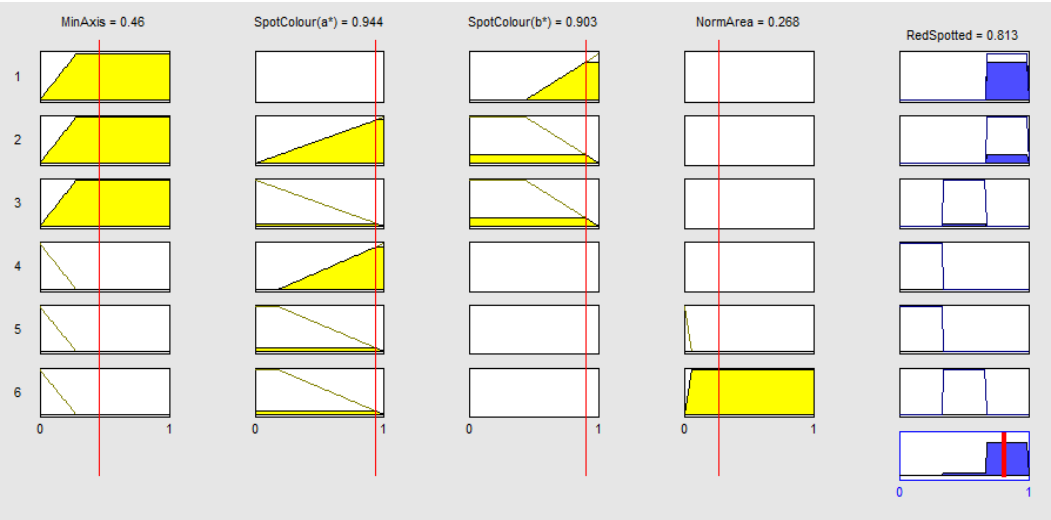
Test results for Red-spotted ladybird



Rule viewer for resultant = *E. 4-pustulatus*

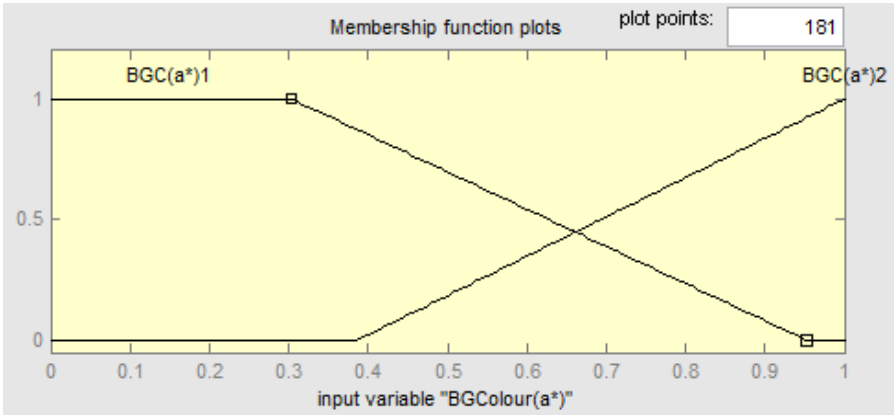


Rule viewer for resultant = *Harmonia axyridis form conspicua*

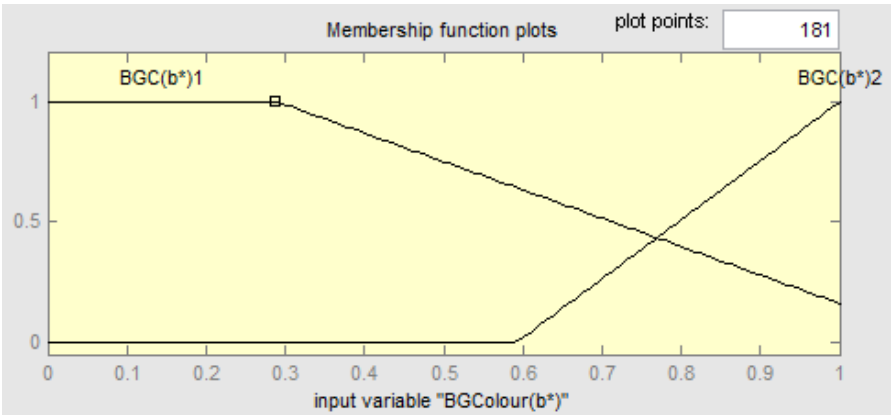


Rule viewer for resultant = *Harmonia axyridis form spectabilis*

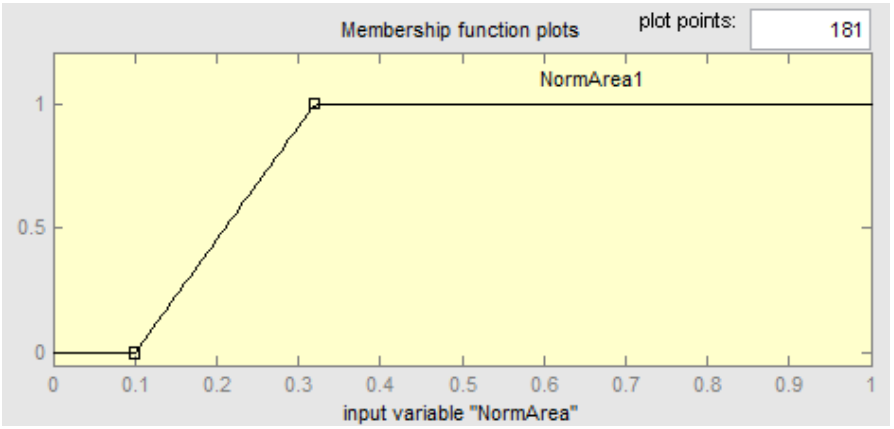
Test results for Black-spotted ladybird



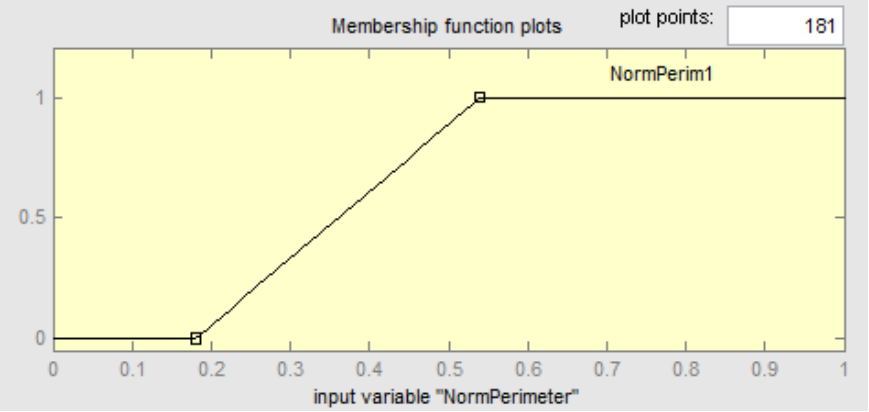
Membership functions for input variables 'BGColour(a*)'



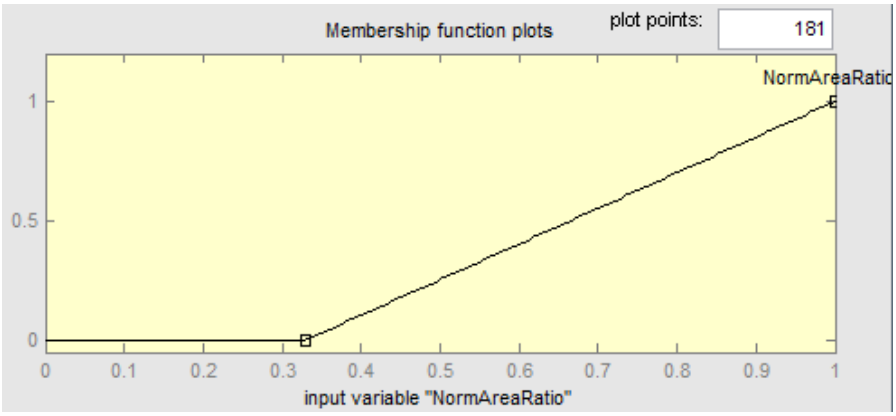
Membership functions for input variables 'BGColour(b*)'



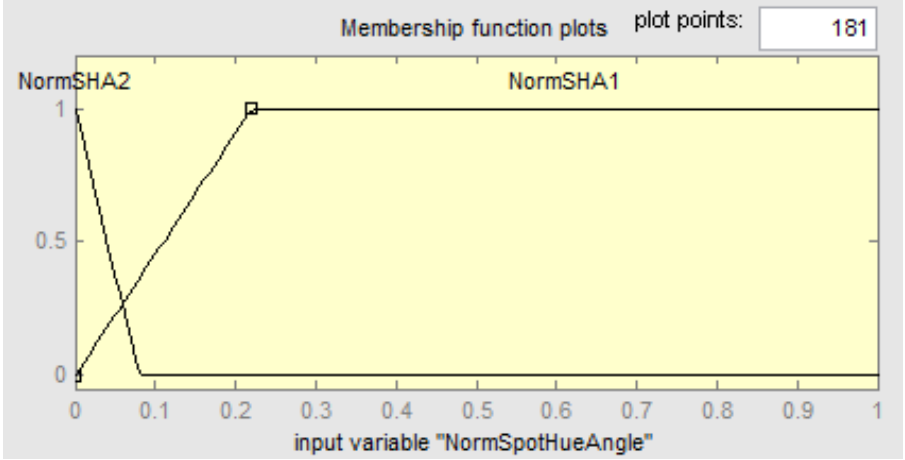
Membership functions for input variable 'NormArea'



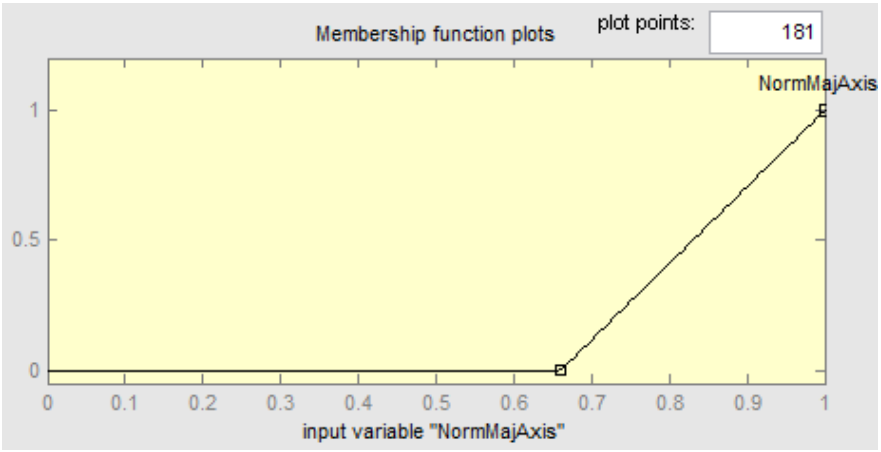
Membership functions for input variable 'NormPerimeter'



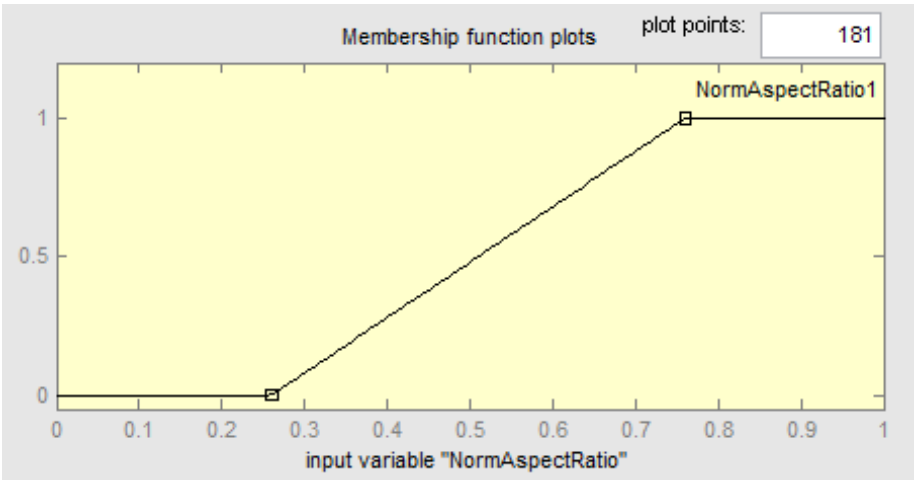
Membership functions for input variable 'NormAreaRatio'



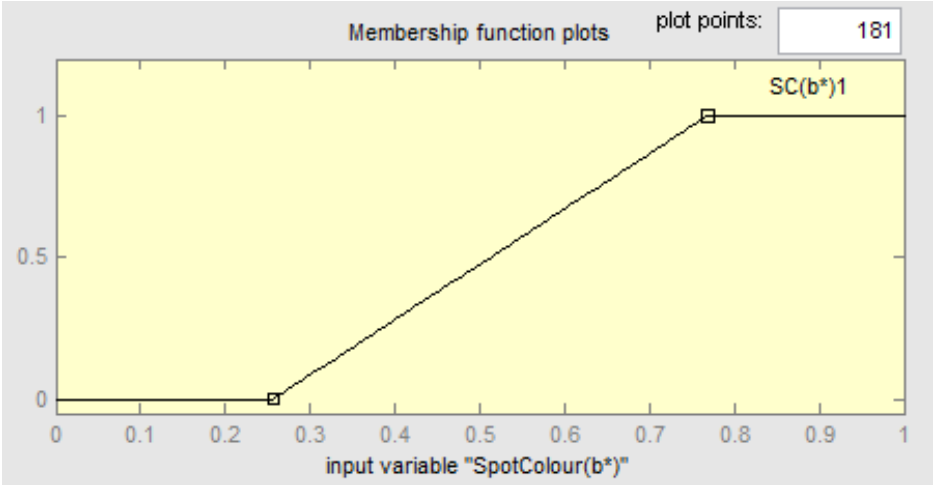
Membership functions for input variable 'NormSpotHueAngle'



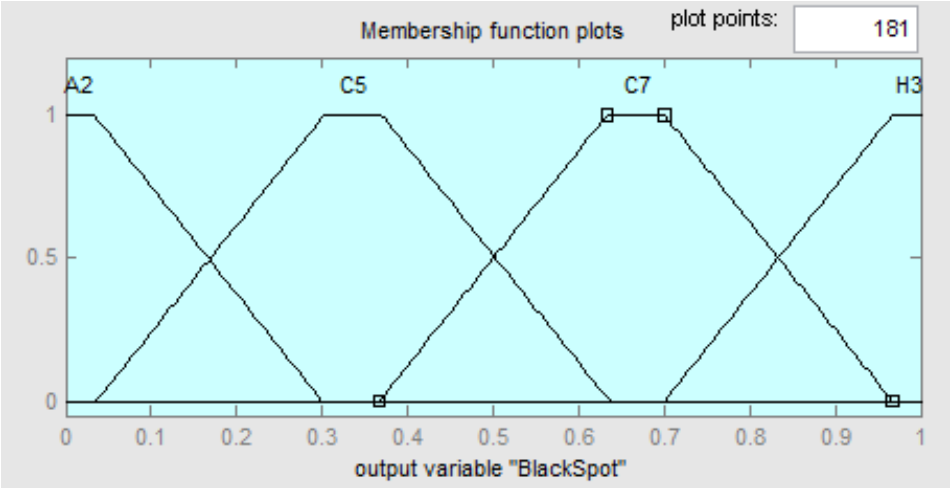
Membership functions for input variable 'NormMajAxis'



Membership functions for input variable 'NormAspectRatio'



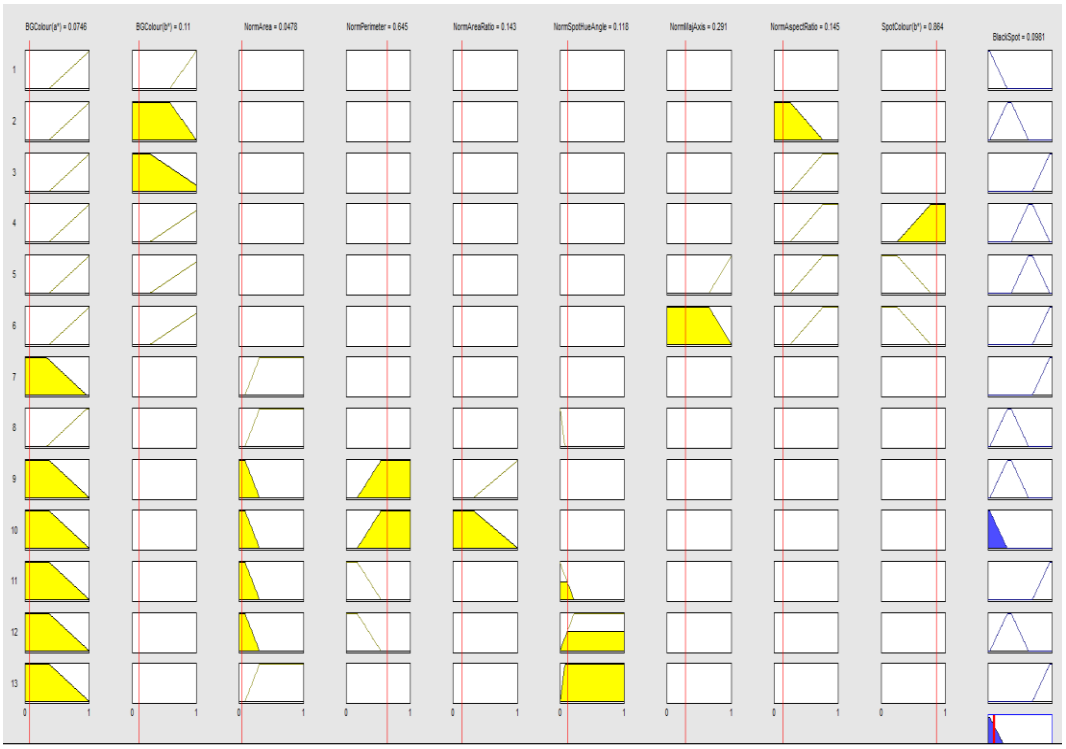
Membership functions for input variable 'SpotColour(b*)'



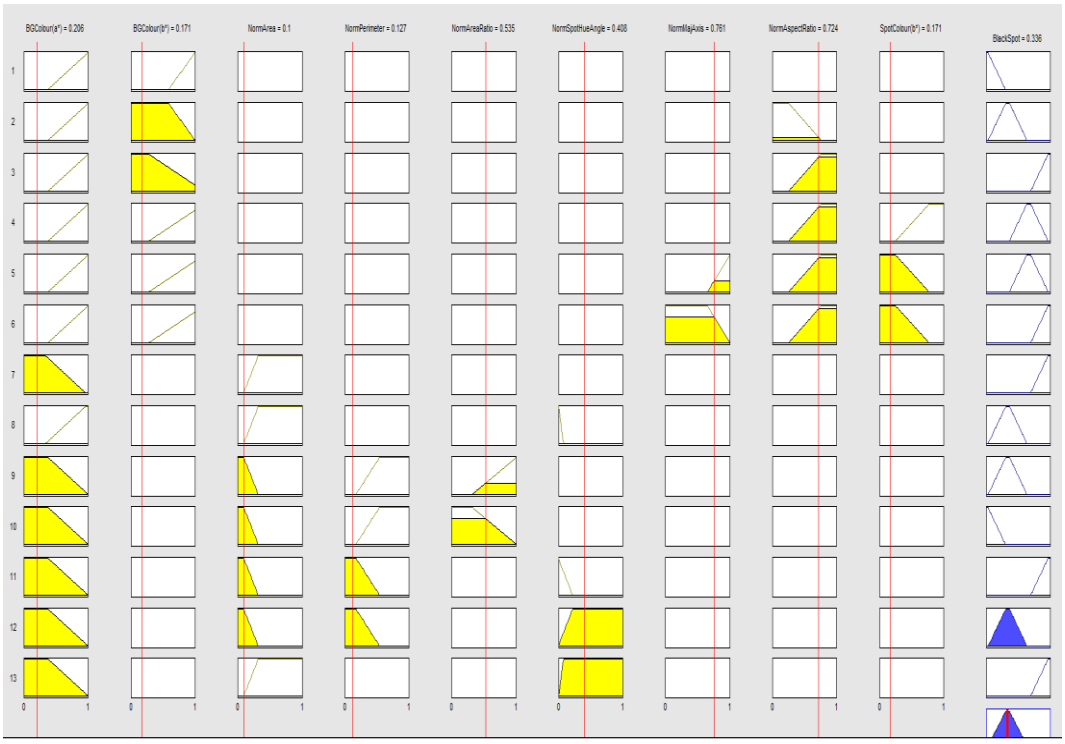
Membership functions for output variables
(Black-spotted ladybird group)

The rules are:

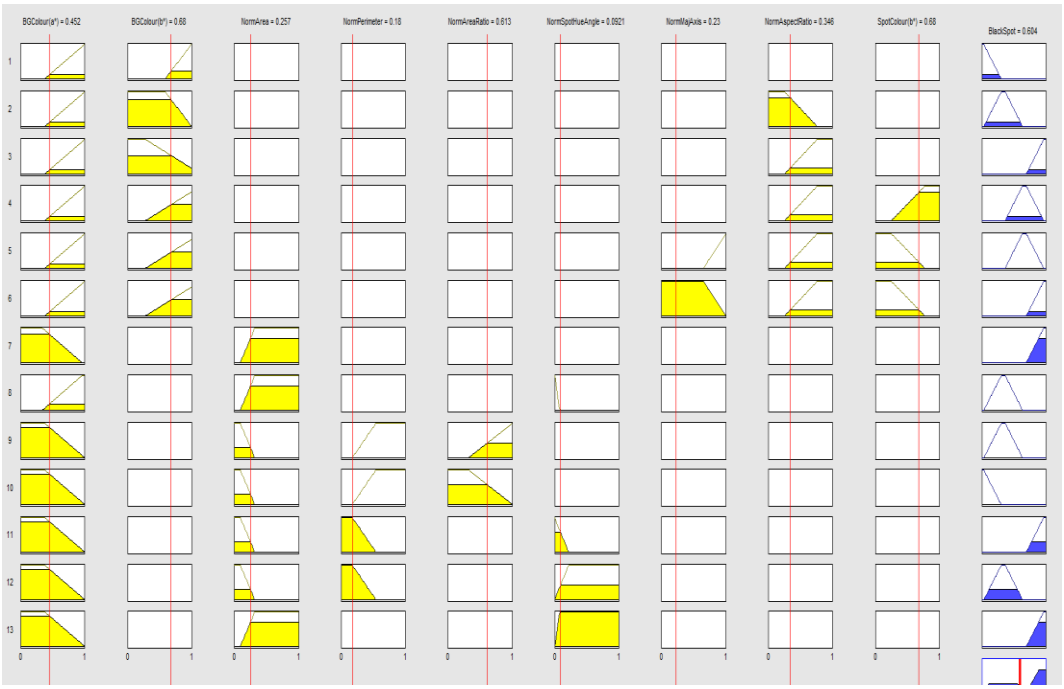
1. IF (BGColour(a*) is BGC(a*)2) and (BGColour(b*) is BGC(b*)2) THEN (BlackSpot is A2)
2. IF (BGColour(a*) is BGC(a*)2) and (BGColour(b*) is not BGC(b*)2) and(NormAspectRatio is not NormAspectRatio1) THEN (BlackSpot is C5)
3. IF (BGColour(a*) is BGC(a*)2) and (BGColour(b*) is BGC(b*)1) and (NormAspectRatio is NormAspectRatio1) THEN (BlackSpot is H3)
4. IF (BGColour(a*) is BGC(a*)2) and (BGColour(b*) is not BGC(b*)1) and (NormAspectRatio is NormAspectRatio1) and (SpotColour(b*) is SC(b*)1 THEN (BlackSpot is C7)
5. IF (BGColour(a*) is BGC(a*)2) and (BGColour(b*) is not BGC(b*)1) and (NormMajAxis is NormMajAxis1) and (NormAspectRatio is NormAspectRatio1) and (SpotColour(b*) is not SC(b*)1 THEN (BlackSpot is C7)
6. IF (BGColour(a*) is BGC(a*)2) and (BGColour(b*) is not BGC(b*)1) and (NormMajAxis is not NormMajAxis1) and (NormAspectRatio is NormAspectRatio1) and (SpotColour(b*) is not SC(b*)1 THEN (BlackSpot is H3)
7. IF (BGColour(a*) is BGC(a*)1) and (NormArea is NormArea1) THEN (BlackSpot is H3)
8. IF (BGColour(a*) is not BGC(a*)1) and (NormArea is NormArea1) and (NormSpotHueAngle is NormSHA2) THEN (BlackSpot is C5)
9. IF (BGColour(a*) is not BGC(a*)2) and (NormArea is not NormArea1) and (NormPerimeter is NormPerim1) and(NormAreaRatio is NormAreaRatio1) THEN (BlackSpot is C5)
10. IF (BGColour(a*) is not BGC(a*)2) and (NormArea is not NormArea1) and (NormPerimeter is NormPerim1) and (NormAreaRatio is not NormAreaRatio1) THEN (BlackSpot is A2)
11. IF (BGColour(a*) is not BGC(a*)2) and (NormArea is not NormArea1) and (NormPerimeter is not NormPerim1) and (NormSpotHueAngle is not NormSHA1) THEN (BlackSpot is H3)
12. IF (BGColour(a*) is not BGC(a*)2) and (NormArea is not NormArea1) and (NormPerimeter is not NormPerim1) and (NormSpotHueAngle is NormSHA1) THEN (BlackSpot is C5)
13. IF (BGColour(a*) is not BGC(a*)2) and (NormArea is NormArea1) and (NormSpotHueAngle is not NormSHA1) THEN (BlackSpot is H3)



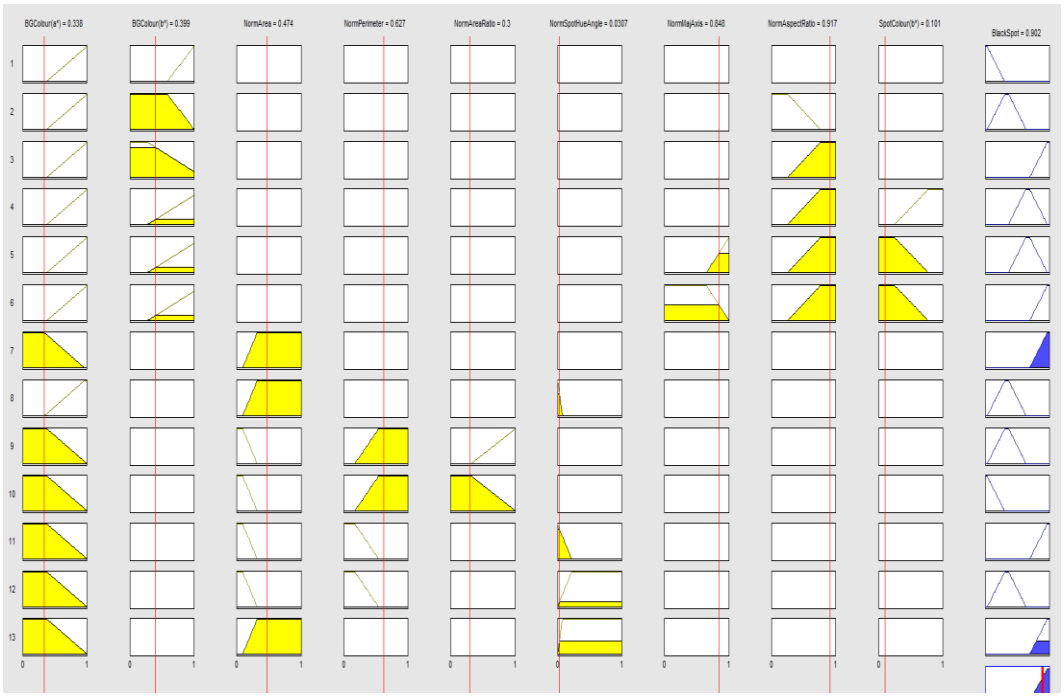
Rule viewer for resultant = A2



Rule viewer for resultant = C5



Rule viewer for resultant = C7



Rule viewer for resultant = H3

APPENDIX VI

PUBLICATION LIST

Seminar Presentation

Ayob, M. Z. and Chesmore, E.D., 2010. Automated Identification of Harlequin Ladybirds using Colour Image Processing. *Entomologist Special Interest Group Seminar*, Rothamsted Research, Harpenden, UK.

Ayob, M. Z. and Chesmore, E.D., 2010. Automated Identification of Harlequin Ladybirds using Colour Image Processing. *RES National Science Meeting 26-28 July*, University of Swansea, UK.

Ayob, M. Z. and Chesmore, E.D., 2011. Intelligent Systems Approach for the Identification of UK Ladybirds. *Departmental Postgraduate Seminar*, Department of Electronics, University of York, UK.

Poster

Ayob, M. Z., 2011. Comparison of Classifier Accuracy for the Identification of Invasive Ladybird Species in UK. *Departmental Postgraduate Poster Competition*, Department of Electronics, University of York, UK.

Conference Paper

Ayob, M. Z. and Chesmore, E.D., 2012. Hybrid feature extractor for Harlequin ladybird identification using color images. *2012 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, San Diego, pp. 214-221.

LIST OF REFERENCES

REFERENCES

An, S. and Hu, Q., 2012. Fuzzy rough decision trees. In: J. T. Yao et al., eds. The 8th International Conference on Rough Sets and Current Trends in Computing RSCTC 2012, *Lecture Notes in Artificial Intelligence LNAI 7413,* Springer-Verlag Berlin Heidelberg, pp. 397-404.

Angel, P.N., 1999. Multiscale image analysis for the automated localisation of taxonomic landmark points and the identification of speceis of parasitic wasp. *Ph.D Thesis,* University of Glamorgan.

Atkinson, W.D. and Gammerman, A., 1987. An application of expert systems technology to biological identification. *Taxon,* 36(4), pp. 705-714.

Ayob, M.Z. and Chesmore, E.D., 2012. Hybrid feature extractor for Harlequin ladybird identification using colour images. *2012 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB),* San Diego, pp. 214-221.

Blackburn, N., Hagstrom, A. Wikner, J. Cuadros-Hansson, R. and Bjornsen, P.K., 1998. Rapid determination of bacteria abundance, biovolume, morphology and growth by neural network-based image analysis. *Applied and Environmental Microbiology,* 64, pp. 3246-3255.

Boddy, L., Morris, C.W., Wilkins M.F., Tarran, G.A. and Burkill, P.H., 1994. Neural network analysis of flow cytometric data for 40 marine phytoplankton species. *Cytometry,* 15, pp. 283-293.

Boddy, L., Morris, C.W. and Morgan, A., 1998. Development of artificial neural networks for identification. In: P. Bridge, P. Jeffries, D.R. Morse and P.R. Scott, eds. 1998, *Information Technology, Plant Pathology and Biodiversity*. Wallingford, UK: CAB International, p. 21.

Boddy, L., Morris C.W., Wilkins M.F., Al-Haddad L., Tarran, G.A., Jonker, R.R. and Burkill, P.H., 2000. Identification of 72 phytoplankton species by radial basis function neural network analysis of flow cytometric data. *Marine Ecology. Progress Series*, 195, pp. 47-59.

Bovik, A., 2000. Introduction to Digital Image and Video Processing. In: *Handbook of Image and Video Processing* (ed. Al Bovik), San Diego: Academic Press, Ch. 1, pp.3-17.

Boyle, R.D. and Thomas, R.C., 1988. *Computer Vision: A First Course*. Oxford: Blackwell Scientific Publications.

Bradley, A., 1997. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30, pp. 1145-1159.

Brown, P.M.J., Frost, R., Doberski, J., Sparks, T., Harrington, R. and Roy, H.E., 2011. Decline in native ladybirds in response to the arrival of *Harmonia axyridis*: early evidence from England. *Ecological Entomology*, 36, pp. 231–240.

Canny, J.F., 1986. A computational approach to edge detection. *IEEE Trans. Pattern Anal.*, 8(6), pp. 679-698.

Chang, C.-Y., Hong, Y.-C., Chung, P.-C. and Tseng, C.-H., 2011. A neural network for thyroid segmentation and volume estimation in CT images. *IEEE Computational Intelligence Magazine*, 6(4), pp. 43-55.

Chesmore E.D. and Monkman G., 1994. Automated analysis of variation in *Lepidoptera*. *The Entomologist*, 113, pp.171-182.

Chesmore E.D., 1998. Methodologies for automating the identification of species. In: *Proceedings of Inaugural Meeting of The BioNET-INTERNATIONAL Group For Computer-Aided Taxonomy (BIGCAT) 1997*, 1, pp. 3-12.

Chesmore, E.D. 2000. Automated species identification the next step in species monitoring. *Proceedings of the Second BioNET-INTERNATIONAL Global Workshop (BIGW2)* (eds. Prof. Tecwyn Jones and Simon Gallagher), BioNET-International, pp. 268-274.

Chesmore, D., Bernard, T., Inman, A.J. and Bowyer, R.J., 2003. Image analysis for identification of the quarantine pest *Tilletia indica. Bulletin OEPP/EPPO Bulletin*, 33, pp. 495–*49*9.

Chesmore, E.D., 2007. The Automated Identification of Taxa: Concepts and Applications. In: Norman MacLeod, ed. *Automated Taxon Identification in Systematics: Theory, Approaches and Applications*. London: CRC Press, 6, pp.83-100.

Clark, J.Y., 2004. Identification of botanical specimens using artificial neural networks. In: *Proceedings of the 2004 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology*, CIBCB'04, pp. 87-94.

Clark, J.Y., 2007. Plant Identification from Characters and Measurements Using Artificial Neural Networks. In: Norman MacLeod, ed. *Automated Taxon Identification in Systematics: Theory, Approaches and Applications*. London: CRC Press, 6, pp. 207-224.

Clark, J.Y., Corney, D.P.A. and Tang, H.L., 2012. Automated Plant Identification Using Artificial Neural Networks. In: *2012 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*. San Diego, California, USA: IEEE, pp. 343–348.

Cortes, C. and Vapnik, V., 1995. Support-vector network. *Machine Learning*, 20, pp. 273-297.

Cotterill, F.P.D., 1995. Systematics, biological knowledge and environmental conservation. *Biodiversity and Conservation*, 4, pp. 183-205.

Dai, J. and Chesmore, D., 2005. Identification of Diptera species by image analysis of wing venation. *Royal Entomological Society National Meeting: Entomology 2005,* 12 -14 September 2005, University of Sussex Royal Entomological Society.

Daly, H.V., Hoelmer, K., Norman, P. and Allen, T., 1982. Computer-assisted measurement and identification of honeybees *(Hymenoptera Apidae). Annals of the Entomological Society of America*, 75, pp. 591-594.

Dallwitz, M.J., 1980. A general system for coding taxonomic descriptions. *Taxon*, 29, pp. 41–46.

Dallwitz, M. J., Paine, T. A., and Zurcher, E. J. 1998. Interactive keys. In: Eds. P. Bridge, P. Jeffries, D. R. Morse, and P. R. Scott. *Information technology, plant pathology and biodiversity.* Wallingford: CAB International, pp. 201–212.

Dawson, W.L., 1913. Identification by camera. *The Condor*. University of California Press, 15, pp. 204-205.

Dietrich, C.D. and Pooley, C.D., 1994. Automated Identification of leafhoppers (Homoptera: Cicadellidae: *Draeculacephala* Ball). *Annals of the Entomological Society of America*, 87, pp. 412-423.

Do, M.T., Harp J.M. and Norris K.C., 1999. A test of a pattern recognition system for identification of spiders. *Bulletin of Entomological Research*, 89, pp. 217-224.

Dorge, T., Carstensen, J.M. and Frisvad J.C., 2000. Direct identification of pure *Penicillium* species using image analysis. *Journal of Microbiological Methods,* 41, pp. 121-133.

Du, C.-J and Sun, D.-W., 2004. Recent developments in the applications of image processing techniques for food quality evaluation. *Trends in Food Science & Technology,* 15, pp. 230-249.

Dubois, D. and Prade, H. 1990. Rough Fuzzy Sets and Fuzzy Rough Sets. *General Systems*, 17, pp. 191–209.

Dunn, C.P, 2003. Keeping taxonomy based in morphology. *Trends in Ecology and Evolution*, 18, pp. 270-271.

Egmont-Petersen M., de Ridder, D. and Handels, H., 2002. Image processing with neural networks- a review. *Pattern Recognition Society,* Elsevier Science, 35, pp.2279-2301.

Evangelista, P.F., Embrechts, M.J. and Szymanski, B.K., 2006. Taming the Curse of Dimensionality in Kernels and Novelty Detection. *Applied Soft Computing Technologies: The Challenge of Complexity eds. Ajith Abraham, Bernard de Baets, Mario K¨oppen, and Bertam Nickolay,* Springer Verlag, pp.431–444.

Fawcett, T., 2006. An Introduction to ROC Analysis. *Pattern Recognition Letters*, 27(8), pp. 861-874.

Foody, G.M., 2001. Thematic mapping from remotely sensed data with neural networks: MLP, RBF and PNN based approaches, In: *Journal of Geographical Systems*, 3, pp.217–232.

Foreroa, M.G., Sroubek, F. and Cristobal, G., 2004. Identification of tuberculosis bacteria based on shape and color. *Real Time Imaging,* 10, pp. 251-262.

Gaston, K. J. and O'Neill, M. A., 2004. Automated species identification: why not? *Philosophical Transactions of the Royal Society London B,* 359, pp. 655–667.

Gauld, I.D., O'Neill, M.A. and Gaston, K.J., 2000. Driving Miss Daisy: the performance of an automated insect identification system. In: A.D. Austin & M. Dowton M. eds. *Hymenoptera: Evolution, Biodiversity and Biological Control,* Collingwood, Victoria: CSIRO, pp. 303-312.

Geman, S., Bienenstock, E. and Doursat, R., 1992. Neural networks and the bias/variance dilemma. *Neural Computation*, 4, pp. 1-58.

Gilkinson, A.K., Pearson, H.C., Weltz, F. and Davis, R.W., 2007. Photo-identification of sea otters using nose scars. *Journal of Wildlife Management*, 71(6), pp. 2045–2051.

Glickstein, N. 1987. Introducing Dichotomous Keys and Taxonomy. *The American Biology Teacher*, University of California Press, 49(8), pp. 438-439.

Gonzalez, R.C. and Woods, R.E., 1992. *Digital Image Processing*. Massachusetts: Addison-Wesley Publishing Company.

Goodall, D.W., 1968. Identification by Computer. *BioScience*, 18(6), pp. 485-488.

Graham, A., 2010. *Understand Statistics*, 6th ed. London: Hodder Education.

Hagan, M.T., Demuth, H.B. and Beale, M., 2002. *Neural Network Design.* Beijing: PWS Publishing Company, ISBN 7-111-10841-8.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. and Witten, I.H., 2009. The WEKA Data Mining Software: An Update; *SIGKDD Explorations*, 11(1).

Ham, F.M. and Kostanic, I., 2001. *Principles of Neurocomputing for Science & Engineering*, McGraw-Hill.

Haralick, R.M., Sternberg, S.R. and Zhuang, X., 1987. Image Analysis Using Mathematical Morphology. *IEEE Transactions on Pattern Recognition and Machine Intelligence*, 9(4), pp. 532–550.

Hart, N. H., and Huang, L. 2011. An image based approach to monitor New Zealand native bees. In: *2011 IEEE Conference on Robotics, Automation and Mechatronics (RAM)*, pp. 353-357.

Hopkins, G.W. and Freckleton, R.P., 2002. Decline in the numbers of amateur and professional taxonomists: implications for conservation. *Animal Conservation,* 5, pp. 245-249.

Iperti, G. & Bertand, E., 2001. Hibernation of *Harmonia axyridis* (*Coleoptera: Coccinellidae*) in South-Eastern France. *Acta Societatis Zoologicae Bohemicae*, 65, pp.207–210.

Jähne, B., 1995. *Digital Image Processing*. Springer-Verlag.

Jing Dai, 2006. Automated Identification of Insect Taxa using Structural Image Processing. *Ph.D Thesis*, Department of Electronics, University of York.

Katsinis, C., Poularikas, A.D. and Jeffries, H.P., 1984. Image processing and pattern recognition with applications to marine biological images. *SPIE 28th Annual International Technical Symposium on Optics and Electro-optics,* San Diego, Society of Photo-optical Instrumentation Engineers, pp. 150-155.

Katsoyannos, P., Kontodimas, D.C., Stathas, G.J. & Tsartsalis, C.T., 1997. Establishment of Harmonia axyridis on citrus and some data on its phenology in Greece. *Phytoparasitica*, 25, pp. 183–191.

Kecman, V., 2001. Learning and soft computing: support vector machines, neural networks, and fuzzy logic models. MIT press.

Kelly, M.J., 2001. Computer-aided photograph matching in studies using individual identification: An example from Serengeti cheetahs. *Journal of Mammalogy*, 82(2), pp. 440-449.

Kipling, M.L. and Chesmore, D., 2005. Automated recognition of moth species using image processing and artificial neural networks. In: *Royal Entomological Society National Meeting, Entomology 2005*, 12-14 September, University of Sussex, Royal Entomological Society.

Kohavi, R., 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, San Francisco, CA, Morgan Kaufmann, pp. 1137-1143.

Kohonen, T., 1990. The Self-Organizing Map. *Proceedings of the IEEE,* 78(9), pp. 1464-1480.

Kohonen, T., 2001. Self-Organizing Map, 3rd ed., Springer-Verlag, New York, pp. 106-127.

Lang, R.I.W., 2007. Neural Networks in Brief. In: Norman MacLeod, ed. *Automated Taxon Identification in Systematics: Theory, Approaches and Applications*. London: CRC Press, Ch.4.

Lang, R.I.W. and Warwick, K., 2002. The plastic self-organising map. In: *Proceedings of International Joint Conference on Neural Networks (IJCNN), part of IEEE World Congress on Computational Intelligence,* Hawaii, pp. 727-732.

Lebbe, J. and Vignes, R., 1998. State of the art in computer-aided identification in biology. *Oceanis*. 24, pp. 305–317.

Liu, J.D., 1996. The expert system for identification of Tortricinae (Lepidoptera) using image analysis of venation. *Entomologica Sinica*, 3(1), pp. 1-8.

Looney, C.G., 1997. *Pattern recognition using neural networks: theory and algorithms for engineers and scientists.* Oxford University Press, Oxford.

Low, A. 1991. *Introductory computer vision and image processing*. London: McGraw-Hill.

Macleod, N., 2007. The Automated Identification of Taxa: Concepts and Applications. In: Norman MacLeod, ed. *Automated Taxon Identification in Systematics: Theory, Approaches and Applications*. London: CRC Press, 1, Ch.1.

Majerus, M.E.N., Strawson, V. and Roy, H., 2006. The potential impacts of the arrival of the harlequin ladybird, *Harmonia axyridis* (Pallas) (Coleoptera:Coccinellidae), in Britain. *Ecological Entomology*, 31, pp. 207-215.

Mayo, M. and Watson, A., 2006. Automatic species identification of live moths. In: Ellis et. al., ed. *Proceedings of the 26th SGAI International Conference on Innovative Techniques and Applications of Artificial Intelligence*, pp. 195-202.

McConkey, S.D., 1999. Photographic identification of the New Zealand sea lion: A new technique. *New Zealand Journal of Marine and Freshwater Research*, 33(1), pp. 63-66.

Meier, R., Shiyang, K., Vaidya, G. and Ng, K.L., 2006. DNA Barcoding and Taxonomy in Diptera: A Tale of High Intraspecific Variability and Low Identification Success. *Systematic Biology*, 55(5), pp. 715-728.

Mohamed, M., 2000. Building Taxonomic Capacity in Developing Countries. In: Prof. Tecwyn Jones and Simon Gallagher, eds. *Proc. of the Second BioNET-INTERNATIONAL Global Workshop (BIGW2)* BioNET-International, pp. 30-34.

Negnevitsky, M., 2005. *Artificial Intelligence: A Guide to Intelligent Systems*, 2nd ed., London: Addison Wesley, pp. 165-217.

Nikolaou, N., Sampaziotis, P., Aplikioti, M., Drakos, A., Kirmitzoglou, I., Argyrou, M., Papamarkos, N. and Promponas V.J., 2010. VeSTIS: A Versatile Semi-Automatic Taxon Identification System from Digital Images. In: Nimis P. L., Vignes Lebbe R., eds. *Tools for Identifying Biodiversity: Progress and Problems.* pp. 231-236.

Nixon, M.S. and Aguado, A.S., 2008. *Feature Extraction and Image Processing*, 2nd ed., London: British Library Cataloguing in Publication Data, pp. 103-178.

Omid, M., 2011. Design of an expert system for sorting pistachio nuts through decision tree and fuzzy logic classifier. *Expert Systems with Applications*, 38, pp. 4339-4347.

Page, L.M., Bart H.L., Jr., Beaman R., Bohs L., Deck L.T., Funk V.A., Lipscomb D., Mares, M A, Prather L.A., Stevenson J, Wheeler Q.D., Woolley J.B., and Stevenson D.W., 2005. LINNE: Legacy infrastructure network for natural environments. Champaign, Illinois, Natural History Survey.

Pajak, M., 2000. Identification of British Bombus and Megabombus using DAISY. *B.A. 3rd Year Honours Project*, Oxford.

Pankhurst, R.J. 1998. A Historical Review of Identification by Computer. In: P. Bridge, P. Jeffries, D.R. Morse and P.R. Scott, eds. *Information Technology, Plant Pathology and Biodiversity*, CAB International, pp. 289-303.

Pawlak, Z., 1982. Rough sets. *International Journal of Computer and Information Sciences*. 11, pp. 341–356.

Payne, A.J., 2001. Processing and Analysis of Hawaiian Happy-Face Spider Images. *MEng Thesis*, Department of Computer Science, University of York.

Pell, J.K., Baverstock, J., Roy, H.E., Ware, R.L. and Majerus, M.E.N., 2008. Intraguild predation involving Harmonia axyridis: a review of current knowledge and future perspectives. *BioControl*, 53, pp. 147-168.

Prechelt, L., 1995. Some Notes on Neural Learning Algorithm Benchmarking. *Neurocomputing*, 9(3), pp. 343-347.

Prechelt, L., 1998. Automatic Early Stopping Using Cross Validation: Quantifying the Criteria, *Neural Networks,* 11(4), pp. 761-767.

Quinlan, J.R., 1996. Improved use of continuous attributes in C4.5. *Journal of Artificial Research*, 4, pp. 77-90.

Richard, M.D. and Lippmann, R.P., 1991. Neural Network Classifiers Estimate Bayesian *a posteriori* Probabilities. *Neural Computation*, 3, pp. 461-483.

Roth, V., Pogoda, A., Steinhage, V. and Schroder, S., 1999. *Integrating Feature-Based and Pixel-Based Classification for the Automated Identification of Solitary Bees.* Informatik aktuell, pp. 120-129.

Russell, K.N., Do, M.N., Huff, J.C. and Platnick, N.I., 2007. Introducing SPIDA-Web: Wavelets, Neural Networks and Internet Accessibility in an Image-Based Automated Identification System. In: Norman MacLeod, ed. *Automated Taxon Identification in Systematics: Theory, Approaches and Applications*. London: CRC Press, pp. 131-152.

Saravanan, N., Cholairajan, S., and Ramachandran, K. I., 2009. Vibration-based fault diagnosis of spur bevel gear box using fuzzy technique. *Expert Systems with Applications*, 36(2), pp. 3119-3135.

Schalkoff, R., 1992. *Pattern Recognition, Statistical, Structural and Neural Approaches*, New York: Wiley.

Schroder S., Drescher W., Steinhage V. and Kastenholz B., 1995. An automated method for the identification of bee species (Hymenoptera, Apoidea). In: *Proceedings of the International Symposium on Conserving Europe's Bees, International Bee Research Association and Linnean Society, London,* International Bee Research Association, pp. 6-7.

Shouche, S. P., Rastogi, R., Bhagwat, S. G., and Sainis, J. K., 2001. Shape analysis of grains of Indian wheat varieties. *Computers and Electronics in Agriculture*, 33, pp. 55–76.

Shri, T. K., and Sriraam, N., 2012. Performance Evaluation of Classifiers for Detection of Alcoholics Using Electroencephalograms (EEG). *Journal of Medical Imaging and Health Informatics*, 2(3), pp. 289-295.

Simpson, M.G., 2010. Chapter 1 Plant Systematics: an Overview. *Plant Systematics* 2nd ed., Academic Press.

Steinhage, V., Kastenholz, B., Schroder, S. and Drescher, W., 1997. A Hierarchical Approach to Classify Solitary Bees Based on Image Analysis. *Informatik aktuell Springer*, pp. 419-426.

Steinhage, V., Schroder, S., Lampe, K. and Cremers, A.B., 2007. Automated Extraction and Analysis of Morphological Features for Species Identification. In: Norman MacLeod, ed. *Automated Taxon Identification in Systematics: Theory, Approaches and Applications*. London: CRC Press, Ch.8.

Sugumaran, V., Muralidharan, V., and Ramachandran, K. I., 2007. Feature selection using decision tree and classification through proximal support vector machine for fault diagnostics of roller bearing. *Mechanical Systems and Signal Processing*, 21(2), pp. 930-942.

Torres, L., Reutter, J.Y. and Lorente, L., 1999. The importance of the color information in face recognition. In: *International Conference on Image Procesing (ICIP 99)*, Oct. 1999, 3, pp. 627-631.

Vízhányó, T. and Felföldí, J., 2000. Enhancing colour differences in images of diseased mushrooms. *Computers and Electronics in Agriculture*, 26, pp. 187–198.

Walker, R. and Kumagai, M., 2000. Image analysis as a tool for quantitative phycology, a computational approach to cyanobacterial taxa identification. *Limnology,* 1, pp. 107-115.

Ware, R.L. and Majerus, M.E., 2008. Intraguild predation of immature stages of British and Japanese coccinellids by the invasive ladybird *Harmonia axyridis*. *BioControl*, 53, pp. 169-188.

Watson, A.T., O'Neill M.A. and Kitching, I.J., 2003. A qualitative study investigating automated identification of living macro lepidoptera using the digital Automated Identification System (DAISY). *Systematics & Biodiversity*, 1, pp. 287-300.

Weeks, P.J.D., O'Neill, M.A., Gaston, K.J. and Gauld, I.D., 1997a. Automating the identification of insects, a new solution to an old problem. *Bulletin of Entomological Research*, 87, pp. 203-211.

Weeks, P.J.D., O'Neill, M.A., Gaston, K.J. and Gauld, I.D., 1997b. Development of an automated insect identification system. *Journal of Applied Entomology*, 123, pp. 1-8.

Weeks, P.J.D., O'Neill, M.A., Gaston, K.J. and Gauld, I.D., 1999. Automating insect identification: exploring the limitations of a prototype system. *Journal of Applied Entomology*, 123, pp. 1-8.

White, I.M. and Sandlant, G., 1998. Computerized insect identification: a comparison of differing approaches and problems. In: Bridge, P., Morse, D.R., Jeffries, P. and Scott, P.R., eds. *Information Technology, Plant Pathology and Biodiversity*. CAB International, Wallingford, pp. 261-272.

White, R. and L. Winokur, 2003. Quantitative Description and Discrimination of Butterfly Wing Patterns Using Moment Invariant Analysis. *Bulletin of Entomological Research*, 93(4), pp. 361-374.

Will, K. W., Mishler, B. D. and Wheeler, Q. D., 2005. The perils of DNA barcoding and the need for integrative taxonomy. *Systematic Biology*, 54, pp. 844–851.

Will, K. W. and Rubinoff, D., 2004. Myth of the molecule: DNA barcodes for species cannot replace morphology for identification and classification. *Cladistics*, 20, pp. 47–55.

Witten, I. H. and Frank, E., 2005. *Data Mining: Practical Machine Learning Tools and Techniques.* 2nd ed., San Francisco: Morgan Kaufmann.

Wolpert, D. H., 1992. Stacked Generalization. *Neural Networks*, 5, pp. 241-259.

Woolley, J.B. and Stone, N.D., 1987. Application of Artificial Intelligence to Systematics: SYSTEX - A Prototype Expert System for Species Identification. *Systematic Zoology*, 36(3), pp. 248-267.

Wu, S. G., Bao, F.S., Xu, E.Y., Wang, Y.X., Chang, Y.F. and Xiang, Q.L., 2007. A Leaf Recognition Algorithm for Plant Classification Using Probabilistic Neural Network. *2007 IEEE International Symposium on Signal Processing and Information Technology*. 15-18 December 2007, Giza, pp. 11-16.

Yip, A. and Sinha, P., 2001. Role of color in face recognition. In: *MIT tech report (ai.mit.com) AI Memo 2001-035,* Massachusetts Institute of Technology, Cambridge, USA.

Yu, D.S., Kokko, E.G., Barron, J.R., Schaalje, G.B. and Gowen, B.E., 1992. Identification of ichneumonid wasps using image analysis of wings. *Systematic Entomology*, 17, pp. 389-395.

Zadeh, L.A., 1983. The role of fuzzy logic in the management of uncertainty in expert systems. *Fuzzy Sets and Systems*, 11, pp. 199-127.

Zakri, A.H., 2000. The Convention on Biological Diversity and the Global Taxonomy Initiative. In: Prof. Tecwyn Jones and Simon Gallagher, eds. *Proc. of the Second BioNET-INTERNATIONAL Global Workshop (BIGW2)* BioNET-International, pp. 35-46.

Internet sources:

Activation function, n.d. [image online]. Available at: <http://tx.liberal.ntu.edu.tw/~purplewoo/Literature/DataAnalysis/threeactivation functions.files/NN1.gif> [Accessed July 2012].

Artificial Neural Networks-A neural network tutorial, n.d. [online]. Available at: <http://www.learnartificialneuralnetworks.com> [Accessed December 2009].

AskOxford Online Dictionary, n.d. [online]. Available at: <http://www.askoxford.com/dictionaries> [Accessed November 2009].

Bullinaria, J. A. 2012 [pdf]. Learning Vector Quantisation. Available at: www.cs.bham.ac.uk/~jxb/INC/l18.pdf [Accessed July 2012].

Cambridge Dictionaries Online, n.d. [online]. Available at: <http://dictionary.cambridge.org/> [Accessed November 2009].

Chen, Y. W., & Lin, C. J., 2005 [pdf]. Combining SVMs with various feature selection strategies. Available at: < http://www.csie.ntu.edu.tw/~cjlin/papers/features.pdf > [Accessed Oct 2012].

CIELAB colour models – Technical Guides, n.d. [online]. Available at: < http://www.dba.med.sc.edu/price/irf/Adobe_tg/models/cielab.html> [Accessed July 2009].

Colour models, n.d. [online]. Available at: < http://www.booksmartstudio.com/color_tutorial/colortheory4.html> [Accessed July 2009].

Community Heritage Initiative, n.d. P. Mabbott, ed. *A colour key for identifying ladybirds in Leicester & Rutland*. [pdf] Available at: <http://www.leics.gov.uk/celebrating_wildlife> [Accessed June 2011].

Eddins, S., n.d. [online]. MathWorks: Steve on Image Processing. Available at: <blogs.mathworks.com/steve/> [Accessed December 2009].

Hsu, C.W., Chang, C.C. and Lin, C.J., 2003. [pdf]. A Practical Guide to Support Vector Classification. Available at: https://www.cs.sfu.ca/people/Faculty/teaching/726/spring11/svmguide.pdf [Accessed Oct 2012].

Just Green, n.d. [online]. Available at: <http://www.just-green.com/Article/170/Harlequin-Ladybirds.html> [Accessed Oct 2009].

MathWorks documentation, n.d. [online]. http://www.mathworks.com/help/nnet/ug/probabilistic-neural-networks.html [Accessed July 2012].

MathWorks documentation, n.d. [online]. http://www.mathworks.com/help/nnet/ug/collect-and-prepare-the-data.html#bss33zw-9 [Accessed Dec 2012].

MORPHIDAS: Morphometric Herbarium Image Data Analysis, 2012. [online]. Available at: http://www.computing.surrey.ac.uk/morphidas/ [Accessed August 2012].

NERC / Field Studies Council-NBN Gateway, 2010. *H. axyridis* distributions in the UK [image online]. Available at: http://data.nbn.org.uk/gridMap/gridMap.jsp?allDs=1&srchSpKey=NHMSYS0000712592 [Accessed 2 August 2012].

Neuro AI-Intelligent Systems and Neural Network, 2007. [online] Supervised learning. Available at: http://www.learnartificialneuralnetworks.com [Accessed August 2012].

O'Neill, M.A., 2007. [pdf] DAISY: A Practical Tool For Semi-Automated Species Identification. Available at: www.fao.org/ag/AGP/AGPS/C-CAB/Castudies/pdf/3-001.pdf [Accessed Oct 2009].

Royal Horticultural Society, 2012. Lily beetle survey. Available at: < http://www.rhs.org.uk/Science/Plant-pests/Lily-beetle> [Accessed August 2012].

Schneider, J., Fri Feb 7 1997 [online]. Cross Validation. http://www.cs.cmu.edu/~schneide/tut5/node42.html [Accessed December 2012].

Southampton Natural History Society, 2005. [pdf]. Ladybirds of Southampton.

sotonnhs.org/docs/LadybirdAll.pdf [Accessed August 2005].

UK Harlequin Survey, n.d. [online]. Available at: http://www.harlequin-survey.org/recognition_and_distinction.htm [Accessed 2009].

UK Ladybird Survey, n.d. [online]. Available at: <http://www.coleoptera.org.uk> [Accessed Oct 2012].

Weiss, D., 2011 [video online]. Machine Learning & Model Selection. Available at: www.youtube.com [Accessed December 2012].

X. Hong, 2009 [pdf]. Probabilistic neural network (PNN). Available at: www.personal.reading.ac.uk/~sis01xh/teaching/CY2D2/Pattern3.pdf [Accessed August 2012].