HYPERSPECTRAL IMAGING AND MACHINE LEARNING OF TEXTURE FOODS FOR CLASSIFICATION

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF INFORMATICS
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

MUSA ATAŞ

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
IN
THE DEPARTMENT OF INFORMATION SYSTEMS

OCTOBER 2011

Approval of the Graduate School of Informatics

_____

Prof. Dr. Nazife BAYKAL

Director

I certify that this thesis satisfies all the requirements as a thesis for the degree of Doctor of Philosophy.

_____

Prof. Dr. Yasemin YARDIMCI

Head of Department

This is to certify that we have read this thesis and that in our opinion it is fully adequate, in scope and quality, as a thesis for the degree of Doctor of Philosophy.

_____	_____
Asst. Prof. Dr. Alptekin TEMİZEL	Prof. Dr. Yasemin YARDIMCI
Co-Supervisor	Supervisor

Examining Committee Members

Assoc. Prof. Dr. Ferda Nur ALPASLAN	(METU, CENG) _____
Prof. Dr. Yasemin YARDIMCI ÇETİN	(METU, II) _____
Asst. Prof. Dr. Erhan EREN	(METU, II) _____
Asst. Prof. Dr. Habil KALKAN	(SDU, CENG) _____
Asst. Prof. Dr. Tuğba TAŞKAYA TEMİZEL	(METU, II) _____

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last name : Musa Ataş

Signature : _____

ABSTRACT

HYPERSPECTRAL IMAGING AND MACHINE LEARNING OF TEXTURE FOODS FOR CLASSIFICATION

Ataş, Musa

Ph.D., Department of Information Systems

Supervisor: Prof. Dr. Yasemin Yardımcı Çetin

October 2011, 137 pages

In this thesis the main objective is to design a machine vision system that classifies aflatoxin contaminated chili peppers from uncontaminated ones in a rapid and non-destructive manner via hyperspectral imaging and machine learning techniques. Hyperspectral image series of chili pepper samples collected from different regions of Turkey have been acquired under halogen and UV illuminations. A novel feature set based on quantized absolute difference of consecutive spectral band features is proposed. Spectral band energies along with absolute difference energies of the consecutive spectral bands are utilized as features and compared with other feature extraction methods such as Teager energy operator and 2D wavelet Linear Discriminant Bases (2D-LDB). For feature selection, Fisher discrimination power, information theoretic Minimum Redundancy Maximum Relevance (mRMR) method and proposed Multi Layer Perceptron (MLP) based feature selection schemes are utilized.

Finally, Linear Discriminant Classifier (LDC), Support Vector Machines (SVM) and MLP are used as classifiers. It is observed that MLP outperforms other learning models in terms of predictor performance. We verified the performance and robustness of our proposed methods on different real world datasets. It is suggested that to achieve high classification accuracy and predictor robustness, a machine vision system with halogen excitation and quantized absolute difference of consecutive spectral band features should be utilized.

Keywords: Machine Vision, Feature Extraction and Selection, Artificial Neural Network, Hyperspectral Imaging, Food Safety.

ÖZ

DOKULU GIDALARIN SINIFLANDIRILMASINDA HIPERSPEKTRAL GÖRÜNTÜLEME VE MAKİNE ÖĞRENMESİ

Ataş, Musa

Doktora, Enformatik Enstitüsü

Tez Yöneticisi: Prof. Dr. Yasemin Yardımcı Çetin

Ekim 2011, 137 sayfa

Bu doktora çalışmasının temel amacı hiperspektral görüntüleme ve makine öğrenmesi ile aflatoksinli pul biberleri temiz biberlerden, hızlı ve tahribatsız bir şekilde ayırabilecek bir bilgisayarla görü sistemi geliştirmektir. Türkiye'nin değişik bölgelerinden toplanmış değişik biberlerin halojen ve UV altındaki hiperspektral görüntüleri elde edilmiştir. Nicemlenmiş ardışık bantların piksel farklarının mutlak değeri temelli yeni bir öznitelik kümesi önerilmiştir. Spektral bant enerjisi ve ardışık bantların piksel farklarının mutlak değeri kullanılarak elde edilen öznitelikler ile Teager enerji işlemi ve iki boyutlu dalgacık dönüşümü Yerel Ayırtaç Tabanları (YAT) temelli öznitelikler karşılaştırılmıştır. Öznitelik seçimi için Fisher ayrımsallık gücü, bilgi teorisi yaklaşımı en küçük fazlalık en büyük ilişki (KFBİ) ve önerilen çok katmanlı algılayıcı (ÇKA) tabanlı teknikler kullanılmıştır.

Son olarak, Doğrusal Ayrımsallık Sınıflandırıcısı (DAS), Destek Vektör Makineleri (DVM) ve Yapay Sinir Ağları (YSA) modelleri sınıflandırıcı olarak kullanılmıştır. Ortalama doğruluk başarım ölçütüne göre ÇKA'ların daha iyi sonuçlar verdiği gözlemlenmiştir. Önerdiğimiz yöntemlerin başarımı ve gürbüzlüğü, değişik veri kümeleri üzerinde gösterilmiştir. Yüksek sınıflandırma başarımı ve gürbüz sınıflandırıcı elde etmek için, halojen ışıklandırma ile birlikte ardışık spektral bantların mutlak değer fark özniteliklerinden meydana gelen bilgisayarla görü sisteminin kullanılması tavsiye edilmektedir.

Anahtar Kelimeler: Makine Görüsü, Öznitelik Çıkarımı ve Seçimi, Yapay Sinir Ağları, Hiperspektral Görüntüleme, Gıda Güvenliği.

To My Parents and My Family

ACKNOWLEDGEMENTS

First and foremost I would like to thank my supervisor Prof. Dr. Yasemin YARDIMCI ÇETİN who has accepted me as a PhD student and provided me with her invaluable academic guidance, stimulating ideas, endless support and friendly attitude during my PhD study. I must extend my special thanks to Assist. Prof. Dr. Alptekin TEMİZEL for his invaluable advices and strong encouragements throughout the research. I would also like to acknowledge the academic support of my committee members, Assoc. Prof. Dr. Ferda Nur ALPASLAN, Assist. Prof. Dr. Erhan EREN, Assist. Prof. Dr. Habil KALKAN and Assist. Prof. Dr. Tuğba TAŞKAYA TEMİZEL. Dr. Habil KALKAN who has not only given me precious insight into the machine vision and aflatoxin detection problem but also provided me guidance and full support, Assist. Prof. Dr. Erhan EREN for his invaluable guidance and ideas about variable selection on neural network, and Assist. Prof. Dr. Tuğba TAŞKAYA TEMİZEL for the fruitful insights about feature selection based on information theoretic approach. Also I wish express my gratitude to Assoc. Prof. Dr. Ferda Nur ALPASLAN for her excellent teaching of machine learning and artificial intelligence courses. I want to express my appreciation to all the faculty and friends in Informatics Institute at Middle East Technical University for their support. Finally, I would like to express my deepest gratitude to my wife Leyla, my son Muhammed Said, my daughter Zeynep Nur and my parents for their patience, courage, understanding and support. I would like to thank to my brother Instructor İsa Ataş for sharing his academic views and support. Special thanks to Tahsin DURAN and Siraceddin MUSABOĞLU for supplying different species of chili peppers.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ABSDIF_CSBE	: Absolute Difference of Consecutive Spectral Band Energy
ADALINE	: Adaptive Linear Element
ANN	: Artificial Neural Network
ANOVA	: Analysis of Variance
BGYF	: Bright Greenish Yellow Fluorescence
CCD	: Charge Coupled Device
CFS	: Correlation-based Feature Selection
CSS_DG	: Complementary Subset Search by Dependency Graph
ELISA	: Enzyme-Linked Immunosorbent Assay
FDB	: Fisher Distance Based
FWHM	: Full width at half maximum
GLCM	: Gray Level Co-Occurrence Matrix
HCC	: Hepatocellular Carcinoma
HBBE	: Hierarchical Bottleneck Backward Elimination
HPLC	: High Performance Liquid Chromatography
IARC	: International Agency for Research on Cancer
IFS	: Incremental Feature Selection
ISBE	: Individual Spectral Band Energy
IR	: Infrared
LDA	: Linear Discriminant Analysis
LDB	: Local Discriminant Bases
LDC	: Linear Discriminant Classifier

LMS	: Least Mean Square
LOO-CV	: Leave One Out Cross Validation
MI	: Mutual Information
MLP	: Multi Layer Perceptron
mRMR	: Minimum Redundancy Maximum Relevance
NID	: Neural Interpretation Diagram
QHM	: Quantized Histogram Matrix
PCA	: Principle Component Analysis
ppb	: Parts per billion
RBF	: Radial Basis Function
SA	: Simulated Annealing
SVM	: Support Vector Machines
TEO	: Teager Energy Operator
TLC	: Thin Layer Chromatography
UV	: Ultraviolet
VLCETF	: Varispec Liquid Crystal Electronically Tunable Filter

CHAPTER 1

INTRODUCTION

Machine vision employs computer vision to examine natural objects and materials, human artifacts and manufacturing processes in order to improve the quality and to provide control in a wide variety of industrial applications (Mark and G., 2003). It typically concerns with inspection of materials like electronic devices, automobiles, food and pharmaceuticals. Developing a machine vision system requires knowledge of diverse disciplines such as mechanical, optical, electronic and software engineering (Mark and G., 2003). Off all the components, software systems play crucial role in the overall performance of the machine vision system. Image processing and machine learning are two significant parts of the software system. Image processing techniques are widely used during the preprocessing and feature extraction stages. Machine learning techniques are employed during the classification and evaluation phases. The classification problems for which there are no predefined rules but a series of cases or observations fall into the area of interest of machine learning. Such problems are found in many domains ranging from medicine, defense, remote sensing, business and robotics. Some widespread applications are speech, character, target, face and object recognition. Given a number of training samples associated with class labels, machine learning aspires

to determine the relationship between the input patterns and the outcomes using only the training examples (Guyon et al., 2006).

Hyperspectral imaging is gaining popularity in many application areas. A machine vision system which is made up of polychromatic image acquisition sensors has more identification and discrimination potential than a monochromatic spectral machine vision system. A typical human eye can sense the wavelengths between 390 nm and 750 nm, known as visible range or visible window (Starr et al., 2005). There are two type of receptors in the human eye: cones and rods. Cones are responsible for color sensing whereas rods are sensitive to low levels of illumination giving a general and overall picture of the scene (Gonzales and Woods, 2002). Humans and some primates are trichromats in general. Trichromacy is the ability of organisms to differentiate colors by the three different types of cone receptors exist in the retina (Svaetichin, 1956). These three cones are called as S, M and L types. The peak wavelength intensities are 420–440 nm for S, 534–545 nm for M and 564–580 nm for L cones respectively (Günther and W.S., 1982). **Figure 1.1** depicts normalized responsivity spectra of human cone cells of S, M and L types. It can be concluded that, human visual perception consists of 3 bins in the visible window.

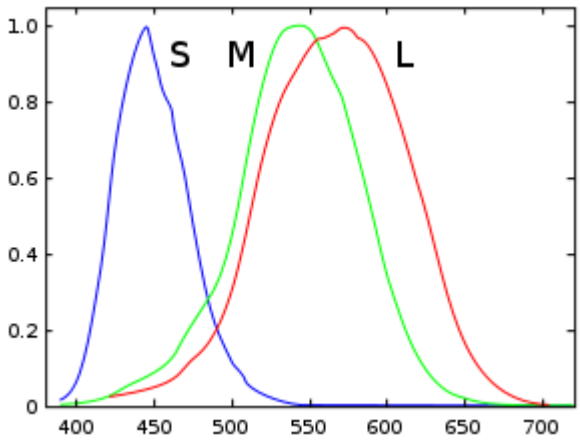


Figure 1.1 Normalized responsivity spectra of human cone cells.

As **Figure 1.1** indicates, visible spectrum is split into three bins. These bins are not separated well especially for the M and L receptors. Therefore, human visual perception system cannot analyze the detailed inspection and recognition of the spectral signature perfectly. However, with machine vision and hyperspectral imaging we can examine various bins with narrow bandwidth along with wider spectrum ranging from ultraviolet to infrared bands.

Hyperspectral imaging incorporates spectroscopy and digital imaging techniques to provide spectral and spatial information together. A hyperspectral image comprises a series of images, containing the intensity distribution at a certain spectral band (ElMasrya et al., 2009). Nevertheless, hyperspectral images contain highly correlated spectral band signature and noises which affect classifier performance adversely. Novel algorithms to mitigate these adverse factors will be developed in this thesis.

1.1 Motivation

Aflatoxins are the toxic metabolites of Aspergillus molds, especially by Aspergillus flavus and Aspergillus parasiticus. They have been studied extensively because they are associated with various chronic and acute diseases especially immunosuppression and cancer. Aflatoxin occurrence is influenced by certain environmental conditions such as drought seasons and agronomic practices. Chili pepper may also be contaminated by aflatoxins during harvesting, production and storage. Turkey is the second largest chili pepper producer in the world (FAO, 2009), but it has less than 3% international market share (Taydaş and Aşkın, 1995, Beriat, 2009). This is partly due to the aflatoxin contamination. Aflatoxin detection based on chemical methods is fairly accurate (Egmond and Paulsch, 1986). However, they are labor intensive, time consuming, expensive and destructive. In this thesis, we used hyperspectral imaging as an alternative for detection of such contaminants in

a rapid and nondestructive manner. In order to classify aflatoxin contaminated chili peppers from uncontaminated ones, we proposed a compact machine vision system based on hyperspectral imaging and machine learning. In the previous work of our research group only UV excitation was employed. We utilized both UV and Halogen excitations to gather more information. Different features with higher discrimination power were extracted from hyperspectral image series. However, the feature set with minimum length is preferred for simple machine vision system. Therefore, we examined various dimensionality reduction techniques and proposed a novel dimension reduction method. Several benchmark tests were performed in order to verify the robustness and reliability of our machine vision system.

1.2 Thesis Objective

In this thesis, we intend to establish a machine vision system based on hyperspectral imaging and machine learning techniques enabling aflatoxin detection in chili pepper with high classification accuracy in a rapid and non destructive manner. Proposed system utilizes both halogen and UV excitations. **Figure 1.2** depicts typical demonstration of the proposed machine vision system. Similarly, **Figure 1.3** illustrates the flowchart and interrelated components of the proposed system.

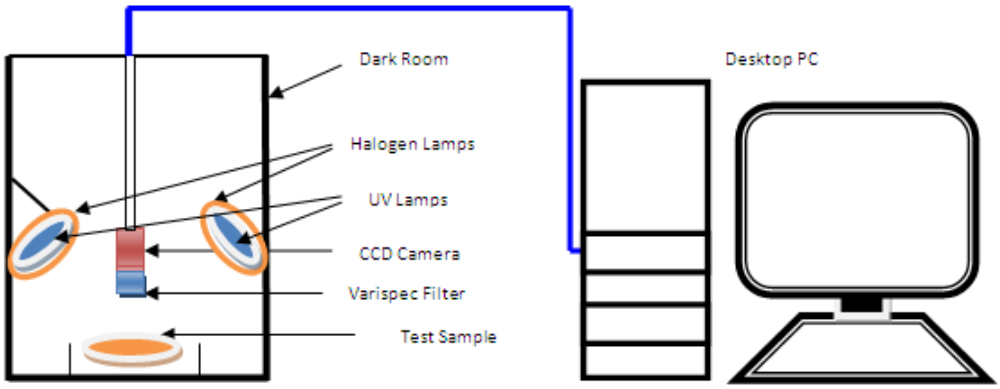


Figure 1.2 Typical demonstration of the proposed machine vision system.

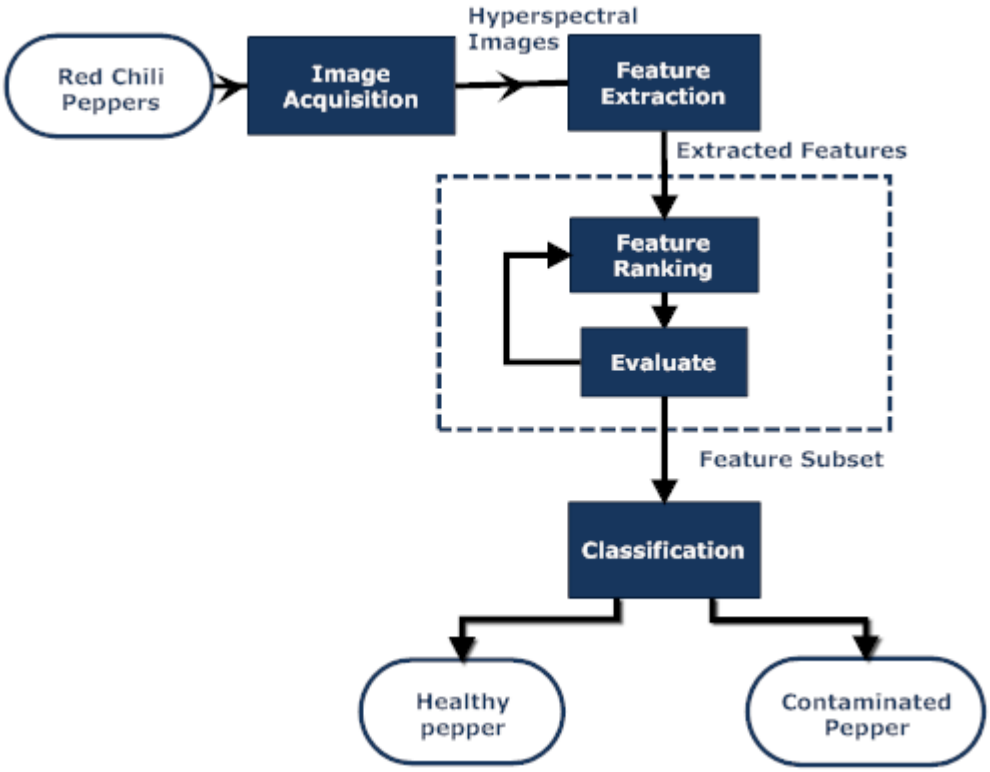


Figure 1.3 Flowchart and interrelated components of the proposed system.

1.3 Contributions of the Thesis

Main original contributions of this dissertation are as follows.

- **Illumination**

  Both halogen and UV illuminations were utilized in the proposed machine vision system for aflatoxin detection of chili pepper. Previous studies acquired spectral images only under the UV excitation (Kalkan, 2008, Beriat, 2009). In this thesis, merit of the halogen illumination was emphasized. In addition to this, it was ascertained that electronic tunable filter can also be used as an alternative to the optical filters in such a machine vision system providing faster acquisition of hyperspectral image of the samples.

- **Feature Extraction**

  Novel feature extraction methods based on absolute difference of consecutive spectral band energies and quantized histogram matrix features (QHM) for hyperspectral data were proposed. Robustness of our proposed feature extraction methods were verified against well known feature extraction methods such as spectral energy features, Teager energy features and wavelet based features. The validity of our proposed feature extraction methods were tested on different chili pepper spectral dataset.

- **Feature Selection**

  A new feature selection approach which we named as hierarchical bottleneck backward elimination (HBBE) method for dimensionality reduction was proposed. HBBE method was verified on different chili

pepper and 10 publicly available UCI machine learning repository datasets. Experimental results revealed that our proposed method was superior to the other dimensionality reduction techniques such as PCA, Fisher discrimination and mRMR from the perspective of providing higher or equal accuracy.

- **Weight Initialization of MLP**

  We observed that the generalization performance of MLP is dependent on the choice of weight initialization. A novel weight initialization method for MLP was developed. Proposed V-Shape weight initialization scheme was compared with different weight initialization techniques. Our approach outperformed other initialization schemes in both feature selection and classification stages. Furthermore, this approach shortens the validation stage because the procedure is always initialized with the same weight vectors. This makes different types of MLP based classifiers more comparable.

Apart from these, we developed a multivariate feature extraction and selection tool that can be used either for hyperspectral images or any other datasets i.e. UCI machine learning repository datasets.

1.4 Thesis Overview

Chapter 2 covers the Aflatoxin in Foodstuff. It gives detailed information about aflatoxin formation. Section 2 emphasizes the widely known chemical methods, their advantages and disadvantages. As an alternative to chemical methods, hyperspectral imaging for aflatoxin detection is described and brief literature review is given in section 3.

Chapter 3 reviews the literature and gives background information about the principles of dimensionality reduction, feature extraction, feature selection and machine learning algorithms.

Chapter 4 gives detailed information about chili pepper sample preparation, acquisition of the hyperspectral images and preprocessing of the hyperspectral data before the feature extraction stage.

Chapter 5 presents novel feature extraction approaches. These are energy features of individual spectral bands and absolute difference of consecutive spectral bands, quantized histogram matrix of preceded features, Teager energy operator features. Wavelet LDB features of Kalkan's are utilized as well for comparison purposes.

Chapter 6 demonstrates detailed information about our proposed hierarchical bottleneck backward elimination method for feature subset selection. In addition to this, other ranking based feature selection methods are covered as well.

Chapter 7 illustrates multivariate feature extraction and selection analysis tool that is developed in this thesis.

Chapter 8 presents the findings and results of the experiments which are performed throughout the PhD study with detailed discussions.

Chapter 9 presents conclusions and suggestions for future works.

CHAPTER 2

AFLATOXIN CONTAMINATION IN FOODSTUFF

2.1 Aflatoxins in Foods

A wide variety of foods (hazelnut, pistachio nut, almond, dried fig, wheat, corn, chili pepper, etc...) are generally prone to Aflatoxin contamination that degrades food quality and also threatens human health. Aflatoxins are toxic compounds produced by many species of Aspergillus molds, especially by Aspergillus flavus and Aspergillus parasiticus (Zeringue and Shih, 1998). The term “aflatoxin” comes from **A**spergillus **fla**vus**toxin**. As International Agency for Research on Cancer (IARC) pointed out, aflatoxin causes human liver cancer (IARC, 2002). Moreover it is reported that, in China and sub-Saharan Africa regions, at least 250,000 deaths from human hepatocellular carcinoma (HCC) occur annually suspected to be due to the aflatoxin contaminated food consumption (Environment, 2011). Therefore, in order to limit exposure to aflatoxin, several countries have taken strict regulations to control aflatoxin contamination level (Environment, 2011). Generally accepted aflatoxin level in food is, 20 ppb in both USA and Turkey. On the other hand maximum level of aflatoxin B1 and total aflatoxin was determined as 5 ppb and 10 ppb in European countries, respectively (European Commission Regulation, 2006). **Table 2.1** shows USA legislation for food and feed aflatoxin limits.

Table 2.1 USA legislation of aflatoxin level in food and feed.

Commodity	**Aflatoxin Level (ppb)**
All products, except milk, designated for humans	20
Milk	0.5
Corn for immature animals and dairy cattle	20
Corn for breeding beef cattle, swine and mature poultry	100
Corn for finishing swine	200
Corn for finishing beef cattle	300
Cottonseed meal (as a feed ingredient)	300
All feedstuff other than corn	20

There are four major aflatoxin types based on their fluorescence property under UV 365 nm: Aflatoxin B1, B2, G1 and G2. Aflatoxin B_1 (AFB_1) is reported as a highly toxic and carcinogenic metabolite produced by certain *Aspergillus* species on agricultural commodities (Vergopoulou et al., 2001). **Figure 2.1** illustrates chemical structure of B_1. Although B_1 is the most significant parameter among the others, 'total aflatoxin' is also used and is obtained as the sum, B1 + B2 + G1 + G2. We labeled a sample with aflatoxin level exceeding 10 ppb as contaminated based on regulations of European Countries (European Commission Regulation, 2006) . Otherwise, we labeled it as uncontaminated.

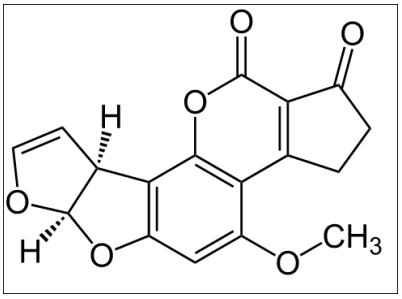


Figure 2.1 Chemical structure of Aflatoxin B_1

According to EU Draft Legislation, aflatoxin limits on some foods for European countries and USA is given in **Table 2.2** (Kithu, 2001).

Table 2.2 EU Draft Legislation of aflatoxin limits on some foods for EU and USA

Country	Permitted Levels	Products	Comments
Austria	B1<1ppb	All Food stuffs	Except mechanically prepared cereals
Belgium	<5 ppb for Peanuts EU legislation is expected		
Germany	B1+B2+G1+G2<4ppb	All foodstuffs	
Denmark	B1<2ppb		
Netherlands	B1<5ppb	All foodstuffs	No controls on B2
Switzerland	B1<1ppb B2+G1+G2<5ppb	All foodstuffs	Except maize
United Kingdom	<50ppb for chili B1+B2+G1+G2<4ppb for Dried Fig and Fig products,		
Spain	B1<5ppb B1+B2+G1+G2<10ppb	All Foodstuffs	
Sweden	B1+B2+G1+G2<5ppb	All Foodstuffs	
Finland	B1+B2+G1+G2<5ppb	All Foodstuffs	
Italy, France	< 10 ppb for B1		
USA, Turkey	<20 ppb	All Foodstuffs	Guideline FDA

Aflatoxin contamination can occur during pre-harvesting and post-harvesting periods. High temperature, prolonged drought conditions and high insect activities are the significant factors for aflatoxin contamination during pre-harvesting. For the latter case, warm temperature and high humidity factors become active ingredients

that increase the mold invasion and toxin production (Wagacha and Muthomi, 2008).

2.2 Aflatoxin Detection Methods

Various methods are suggested for detection of aflatoxin. Mass Spectroscopy (MS), Thin Layer Chromatography (TLC), High Performance Liquid Chromatography (HPLC) and Enzyme-Linked Immunosorbent Assay (ELISA) are the most commonly known chemical methods. Off all the methods HPLC gives most accurate and sensitive quantitative results with different types of aflatoxin (Egmond and Paulsch, 1986).

Although chemical methods give quantitative and accurate results, they are slow, expensive and destructive in nature. Recently, machine vision and pattern classification techniques are developed for aflatoxin detection (Marsh et al., 1969, Tyson and Clark, 1974, Bothast and Hesseltine, 1975, Hirano et al., 1998, Zeringue et al., 1999, Pearson et al., 2001, Yao et al., 2006, Kalkan, 2008, Beriat, 2009, Ataş et al., 2010b, Ataş et al., 2010a, Ataş et al., 2011a, Ataş et al., 2011b). Under 365 nm UV light (also known as Black Light), aflatoxin positive food exhibits Bright Green Yellowish Fluorescent (BGYF). However, BGYF produced by Aspergillus is actually due to Kojic Acid, a secondary by-product of Aflatoxin, not the aflatoxin itself (Peter, 2003).

In fact, BGYF is caused by two constituents; the kojic acid and peroxidase enzyme from the host. Kojic acid is formed by the fungus which may be Aspergillus Flavus, Aspergillus Parasiticus or some other variants. According to Marsh et al., (1969) even some bacteria can produce kojic acid and results in BGYF in fiber. **Figure 2.2** summarizes the formation of BGYF.

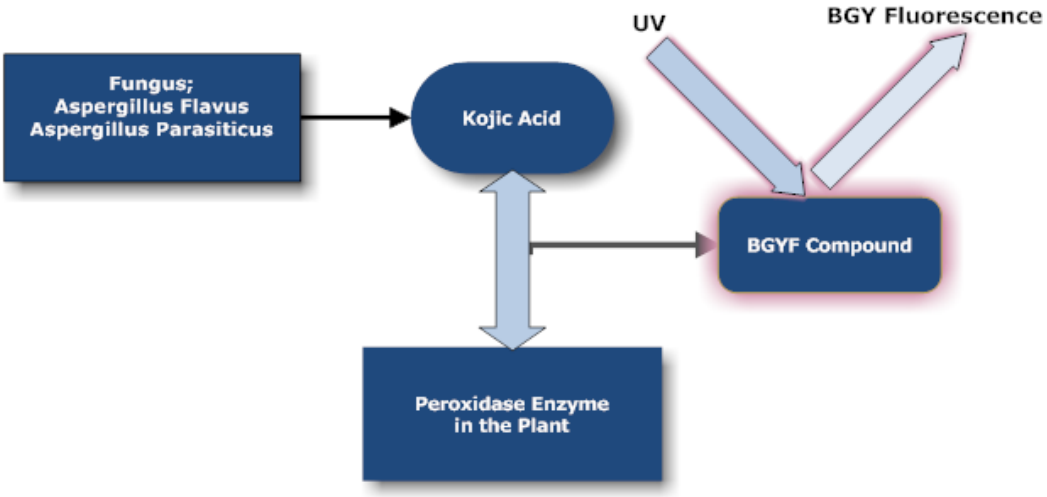


Figure 2.2 Flowchart of BGYF phenomena.

Certain fungi produce kojic acid which may result in BGYF when there is enough peroxidase enzyme in the plant. It is known that not all fungi that produce kojic acid also produce aflatoxins. Similarly, the lack of peroxidase enzyme may conceal the presence of aflatoxins because BGYF will be absent. Thus, BGYF itself does not directly indicate the actual presence of aflatoxin and it may result in false positives and negatives during the evaluation stage. Furthermore, in the previous studies on corn and pistachio the authors (Pearson et al., 2001, Yao et al., 2006) stated that BGYF phenomenon under UV illumination is observed when average aflatoxin level exceeds 100 ppb. Therefore, BGYF based aflatoxin detection is not always recommended (Doster et al., 1996, Fersaie et al., 1978, Herrman, 2002). **Table 2.1** (Bothast and Hesseltine, 1975) indicates that BGY fluorescence does not always guarantee the presence or lack of aflatoxin in some agricultural commodities. Note that, although rice sample in their study originally had 64 µg/g aflatoxin contamination level, BGY fluorescence was not observed at all. On the other hand, strong BGYF is observed for the sorghum, but total aflatoxin level is moderate. NRRL stands for Northern Regional Research Laboratory.

Table 2.1 Aflatoxin levels and observed BGY fluorescence on some commodities.

Commodity	BGY Fluorescence[a]		Aflatoxin $B_1 + B_2$ (μg/g)	
	NRRL 2999[b]	NRRL 482[c]	NRRL 2999	NRRL 482
Wheat	++	+	25	ND[f]
Oats	++	–[d]	38	ND
Barley	+	+	32	ND
Rice	–	–	64	ND
Coconut	(++)[e]		78	
Yellow corn	++	+++	21	ND
White corn	+	++	61	ND
Peanuts	–	–	42	ND
Sorghum	+++	++++	28	ND
Soybeans	–	–	12	ND

[a] Number of plus or minus signs equals relative amounts of BGY fluorescence observed.
[b] NRRL 2999 is a high aflatoxin-producing strain of *Aspergillus parasiticus*.
[c] NRRL 482 is a nonaflatoxin-producing strain of the *Aspergillus flavus* group.
[d] Poor mold growth.
[e] Considerable interfering fluorescence.
[f] ND, Not detected.

2.3 Hyperspectral Imaging for Aflatoxin Detection

Hyperspectral imaging acquires images of simultaneously in several narrow and consecutive spectral bands. They consist of both spatial and spectral information. **Figure 2.3** (Gruna et al., 2010) depicts hyperspectral image cube. Hyperspectral imaging was first developed for remote sensing for military aims approximately 20 years ago and has recently utilized for the inspection of quality and safety of food and agricultural commodities (Lu, 2007, Zude, 2008, Gruna et al., 2010).

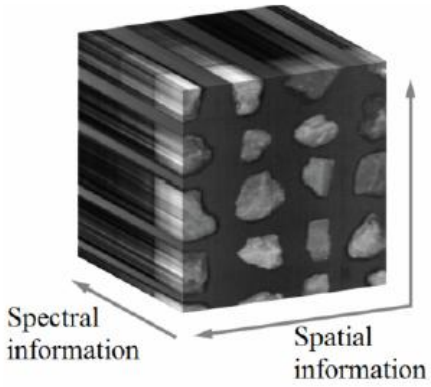


Figure 2.3 Hyperspectral data cube.

Some researchers used BGYF in their studies. Utilizing the reflectance ratios of 440/490 nm and 450/490 nm, Tyson and Clark, (1974) achieved 90% classification rate by examining aflatoxin-infected pecans under UV fluorescence. By analyzing corn kernels, again Yao et.al, (2006) achieved 87% and 88% classification rates for 20 ppb and 100 ppb aflatoxin level thresholds. Kalkan et al., (2011) studied hazelnuts and red chili peppers and achieved 92.3% and 79.2% classification accuracies respectively. Similarly, Beriat, (2009) analyzed 40 chili pepper samples and achieved 80% generalization performance for aflatoxin detection by using linear discriminant classifier. Another possible excitation mode is halogen illumination. Hirano et al., (1998) used transmittance ratio of T700 nm/ T1100 nm bands for peanuts classification under halogen illumination and achieved 95% classification accuracy. Again, Pearson et al., (2001) achieved 96.6% classification accuracy rate for corn samples under the 100W quartz-tungsten-halogen lamp by utilizing the spectral reflectance ratio of R735 nm / R1005 nm bands. They used discriminant analysis technique for detecting highly contaminated corn kernels (>100 ppb) from low contaminated (<10 ppb) or uncontaminated ones. Hence, in order to combine the advantageous of both excitation modes, we decided to utilize UV and halogen illuminations in our machine vision system.

CHAPTER 3

LITERATURE REVIEW AND BACKGROUND

3.1 Feature Extraction and Dimensionality Reduction

Transforming the original data into a set of descriptive features is known as feature extraction. In pattern recognition and image processing domain, feature extraction process is closely related to the dimensionality reduction concept. Ideally the feature vector should be the simplest and the most compact representation of the phenomenon that is being tested. The effectiveness of the classifier is directly related to the choice of the feature vector. If the feature vector contains redundant features and excessive noise or outliers, the classifier may not converge to the correct solution. In addition to this, larger feature size leads to so called 'Curse of Dimensionality' problem and eventually may over-fit the training data. We should eliminate irrelevant and redundant features from the feature set by applying feature extraction and selection techniques. Feature extraction, projects the original data or high dimensional feature vectors onto another set of basis vectors. Feature selection, on the other hand, is essentially ranking the feature vector components according to their predictive significance and selecting the most discriminative ones. The main advantage of using feature extraction over selection is that, extraction process results in a smaller and richer set of features (Oracle, 2008). Alternatively, feature selection has the benefit of requiring a simpler data acquisition device.

So for both techniques, the main objective is to reduce feature vector size to an acceptable level with minimal loss of information. This process is also known as dimensionality reduction. Fewer features not only improve the classifier performance but also provide faster computation and better understanding the underlying mechanism of the problem (Guyon and Elisseeff, 2003).

For dimensionality reduction, techniques such as PCA (Fukunaga and Koontz, 1970, Jollifie, 2002) and Auto Associative ANN can be used. The structure of the Auto Associative ANN is demonstrated in **Figure 3.1**. PCA is a well known dimensionality reduction technique which aims to identify the most meaningful basis to re-express original features and dataset. There are several studies where PCA is successively utilized. As researchers indicate, multivariate analysis, fault tolerance, data validation problems were addressed by PCA effectively (Wise and L., 1989). Similarly, studies such as data visualization (Stephanopoulos and Guterman, 1989), quality control (MacGregor, 1989) and correlation and prediction (Joback, 1984) also employed PCA successively. Likewise, the Autoassociator can give similar results as PCA, if it is used with single hidden layer as a bottleneck and linear activation function in the network model (Bourlard and Kamp, 1988). When the number of hidden units is less than the number of input or output nodes, Multi Layer Perceptron (MLP) performs dimensionality reduction. **Figure 3.1** demonstrates that, upper region acts as an encoder while, lower part acts as a decoder. As it is seen, outputs of the bottleneck hidden neurons are compact representations of the input data. Autoassociator networks may outperform the PCA when an appropriate non-linear transformation is employed (Bourlard and Kamp, 1988).

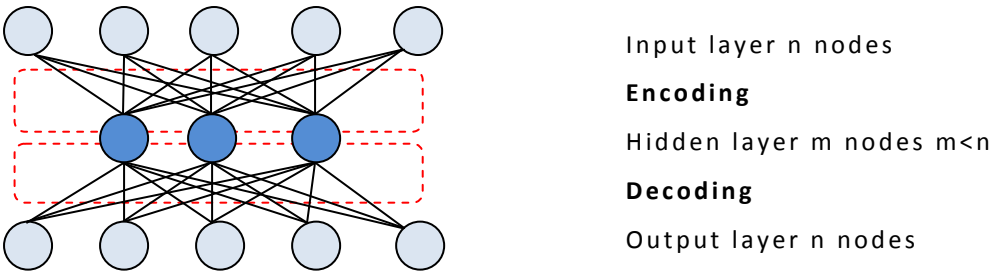


Figure 3.1 Structure of the auto associative ANN. Hidden layer behaves as a bottleneck where encoding and decoding processes take place.

One drawback of using PCA and Autoassociator networks methods is that, they are unsupervised learning procedures and therefore do not consider the relation between the features and the class labels. Moreover, PCA and/or Autoassociator methods, force us to exploit all the variables which actually may not be desirable for hyperspectral data due to the redundancies between spectral bands. In addition to these, PCA may represent the noise instead of the data in the presence of excessive noise (Vijayakumar, 2011).

Despite the simplicity and popularity of PCA, the features extracted from PCA may not always produce better generalization than the original dataset (Cheriyadat and L.M.Bruce, 2003). Discriminative features can also be extracted by first and second order statistics (Lee and Landgrebe, 1993), Wavelet LDB algorithm (Kalkan, 2008, Kumar et al., 2001) or simply average intensity value .

3.2 Feature Selection and Ranking

In machine learning, feature selection, also known as variable selection or feature reduction, is the technique for selecting the most relevant features from the original feature set. It differs from the feature extraction in that, the selection process does not need to alter the data samples whereas the extraction process

performs linear or non-linear transformations which eventually modify the original data. By feature selection, accelerated learning process and reduced processing power and storage requirements is expected. Besides, it provides improvements in the generalization performance and reduces the learning model complexity which enables us to better interpret the underlying mechanisms (Guyon and Elisseeff, 2003). A simple approach for feature selection is, ranking the features based on their predictive significance and discrimination power (Kittler, 1978). According to Kohavi and John (1997), feature ranking is a filtering method since it does not depend on any predictor. It is a preprocessing stage of classification but it may be optimal under certain conditions (Kohavi and John, 1997). Ranking features, based on Fisher discrimination power metric yields favorable results if Fisher's linear discriminant is employed as the classifier under jointly Gaussian priors (Duda et al., 2001). (Guyon and Elisseeff, 2003) pointed out that, even if feature ranking is not optimal it is still preferable to other feature subset selection techniques because of its computational and statistical efficiency. It requires only *N* computations for *N* number of features. On the other hand, feature subset selection needs to compute almost all possible combinational groups. Combinatorial problems are regarded as NP hard.

Given $P(F_i)$ as the discrimination power of the *i*'th feature. Ranking procedures starts with an empty set $F_0=\emptyset$ and then expands by adding *f'*. **Equations 3.1** and **3.2** summarize the ranking procedure as (Michalak and Snicka, 2006):

$$F_n = F_{n-1} \cup \{ f' \} \qquad \text{(Equation 3.1)}$$

Where,

$$f' = arg_{F_{n-1} \cap \{f_i\}=\emptyset} \max P(f_i) \qquad \text{(Equation 3.2)}$$

Note that, ranking procedure does not take into account the dependency amongst the features and examine features one by one. Determining the optimum number of top *N* ranked features is still an open problem Some researchers propose utilizing the first *N* features from the ranked feature set (Ataş et al., 2011b, Kalkan et al., 2011). In fact, applying feature ranking would be more efficient than the feature subset selection in terms of computational cost. Ranking procedures can be based on correlation coefficients, information theoretic approaches like information gain (Hunt et al., 1966), entropy (Mantaras, 1989) and ANOVA technique and their variants mRMR(Peng et al., 2005b) and RELIEF (Kira and Rendell, 1992). Another set of procedures use some saliency metrics derived from connection weights of ANN or SVM (Hall, 1998, Zhao et al., 2002, Robnik-Sikonja and Kononenko, 2003, Guyon et al., 2006, Ataş et al., 2011a) .

3.3 Feature Subset Selection

Size of the extracted feature vector is crucial, especially for a small sample size data set where the number of samples in the training data is lower than the number of features. Increasing the feature vector dimension requires an exponential increase in the data size. Since generation of training data is costly or in some cases not feasible, the size of the feature vector should be reduced to an acceptable level. There is no consensus on the optimal feature vector size for a given training data size. As a good practice, it is recommended that, training data should be about 10 times the feature vector size for each class to produce robust and reliable classification (Bishop, 1995). Feature subset selection is the operation of removing as many redundant features as possible. In this way, better classification rate can be achieved. This also leads to faster and more reliable learning of the classifier. Moreover, it also results in a simpler classifier design, better understanding and interpretation of the data (Marona Noelia et al., 2005). As **Figure 3.2** indicates, feature subset selection process is basically composed of four stages (Dash and Liu,

2011). These are, feature subset generation, evaluation, stopping criterion and validation stages.

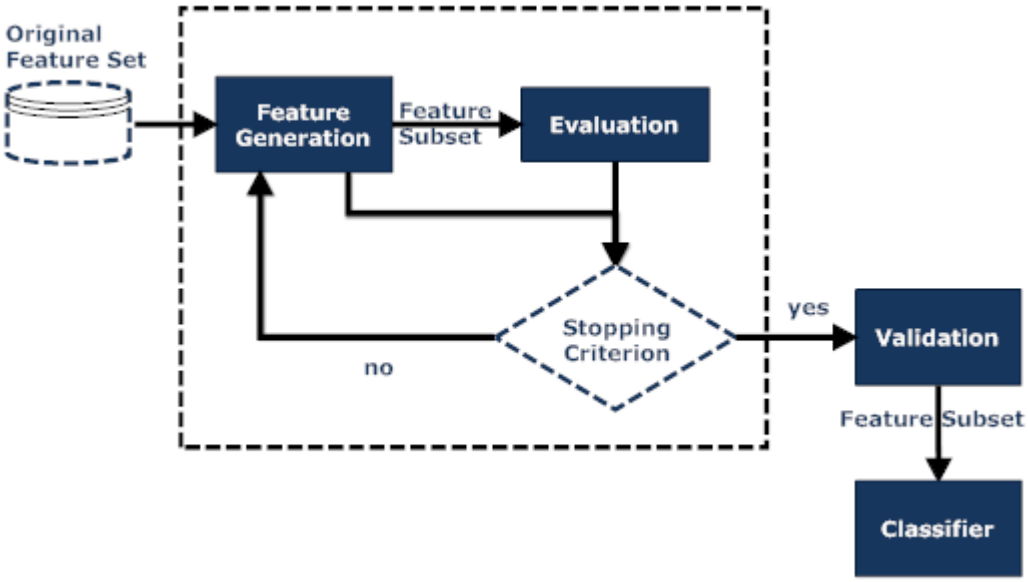


Figure 3.2 Schematic diagram of the feature subset selection.

3.3.1 Feature Subset Generation

Feature subset generation can be performed by three search strategy groups. These are complete, heuristic and stochastic search strategies.

3.3.1.1 Complete Search Method

It is also called the exhaustive search method because all possible subsets are examined individually. Theoretically there exist 2^N possible subsets for a given N dimensional feature space. Although exhaustive search strategy achieves global optimum of the given dataset, it is not advisable due to its high computational cost and was reported as a NP-hard problem (Amaldi and Kann, 1998).

3.3.1.2 Heuristic Search Method

Heuristic search is steered by certain rules contrary to the exhaustive search where search proceeds blindly. Heuristic search cannot guarantee the convergence to the global optimum, but lower computational cost makes it more preferable. It is a sub-optimal solution for multivariate problems. If we are dealing with relatively small feature size, we can still prefer exhaustive search, otherwise heuristic methods are suggested.

Forward selection and backward elimination approaches are extensively used for greedy search and reported as robust to over fitting (Guyon and Elisseeff, 2003). On the other hand, it is pointed out that, it may produce sub optimal results (Pudil et al., 1994, Cover and van Campenhout, 1977). In forward selection, features are continuously added to build larger subsets if they contribute to classifier performance. Hence, forward selection starts with an empty subset and search progresses for a candidate feature which is complementary to the given subset. At the end, it reaches a potentially sub-optimal solution where the final subset is composed of complementary features. Conversely, in backward elimination method, search starts with the highest number of features and the feature subset adaptively shrinks until removing another candidate feature will not give any improvement in terms of generalization performance. Forward selection method is generally faster than backward elimination method. Forward and Backward selection methods can be used in association. It is called stepwise bidirectional selection.

3.3.1.3 Stochastic Search Method

For meta-heuristic search, various methods have been proposed such as genetic algorithm, tabu-search and simulated annealing (Guyon and Elisseeff, 2003, Korycinski et al., 2003).

Optimal feature subset highly depends on the number of trials in the search which in turn relies on available resources (Dash and Liu, 2011). The well known traveling salesman problem can be efficiently solved by using simulated annealing approach (Kirkpatrick et al., 1983). Although simulated annealing does not guarantee convergence to global optimum in finite number of trials, still it can give us preliminary insight. In order to increase the convergence probability we start a series of simulated annealing searches at the same time and take the one with the lowest error into consideration.

3.3.1.3.1 Simulated Annealing

Simulated annealing (SA) resembles the annealing process in metallurgical engineering. In material science, samples are heated at a temperature greater than recrystallization temperature, in order to supply enough energy for atoms to move to the most stable state. Then the temperature is gradually decreased restricting the atomic movement and jumping. Eventually the atoms are stabilized at the minimum energy states. This gives a stronger and defect free product, SA was successfully applied to various optimization problems including the traveling salesman problem, placement and wiring of micro-chips, (Kirkpatrick et al., 1983) and topological optimization of microwave filters (Curtis et al., 2002). The simulated annealing algorithm uses a temperature variable which decays with each iteration. The pseudocode of the simulated annealing algorithm is presented below. **Figure 3.3** illustrates the simulated annealing search algorithm.

```
s ← s0; e ← E(s)                    //initial state, energy.
sbest ← s; ebest ← e                //initial "best" solution
t ← predefined constant             //initial temperature constant
while t > t_min and e < emax        //while time left & not good enough:
  snew ← neighbor(s)                //randomly pick some neighbor.
```

```
 cost ← E(snew)                          //compute its cost
 if exp(-e/t)> random() then             //should we move to it?
  s ← snew; e ← cost                     //yes, change state.
 if cost < ebest then                    //is this a new best?
  sbest ← snew; ebest ← cost             //save 'new neighbor' to 'best found'.
 t ← t*0.9999                            //temperature decays
return sbest                             //return the best solution found.
```

The main difference between traditional descent algorithms like steepest descent or hill climbing and simulated annealing is that, traditional search always moves in a single direction of improvement with the risk of getting trapped at the local minima, whereas SA allows non-improving moves to escape from the local minima. In simulated annealing the move is selected randomly and only if the move is better than current position then SA accepts it, otherwise, it will be accepted only under the some probability constraint. It should be noted that, this acceptance probability also decreases gradually.

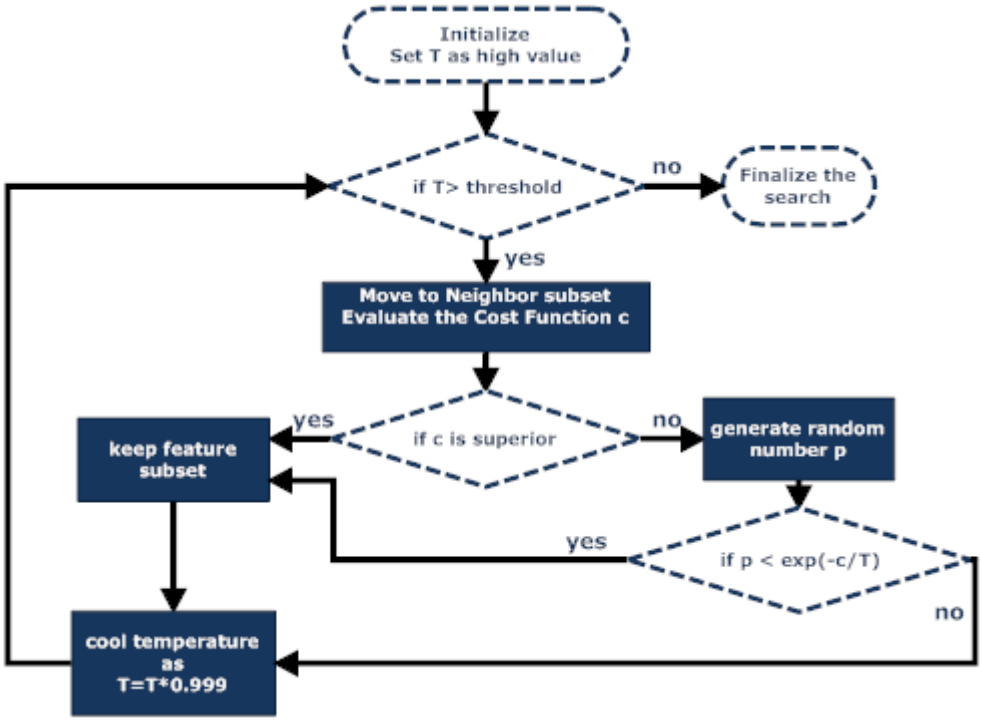


Figure 3.3 Workflow of the simulated annealing search algorithm.

3.3.2 Evaluation

Selected features can be evaluated either by the filter or wrapper model (Kohavi and John, 1997). **Figure 3.4** demonstrates the structure of each model. The choice of the filter model is independent from the choice of the classifier. Feature ranking actually is a kind of filtering process (Kohavi and John, 1997). On the other hand, wrapper method acts as a black-box and directly related to classifier capability. Statistically, the filtering method is reported to have less variance and is robust to over-fitting (Hastie et al., 2009). Although wrapper model yields better classification rates than the filter model, it is slower and its computational cost is higher (Talavera, 2005).

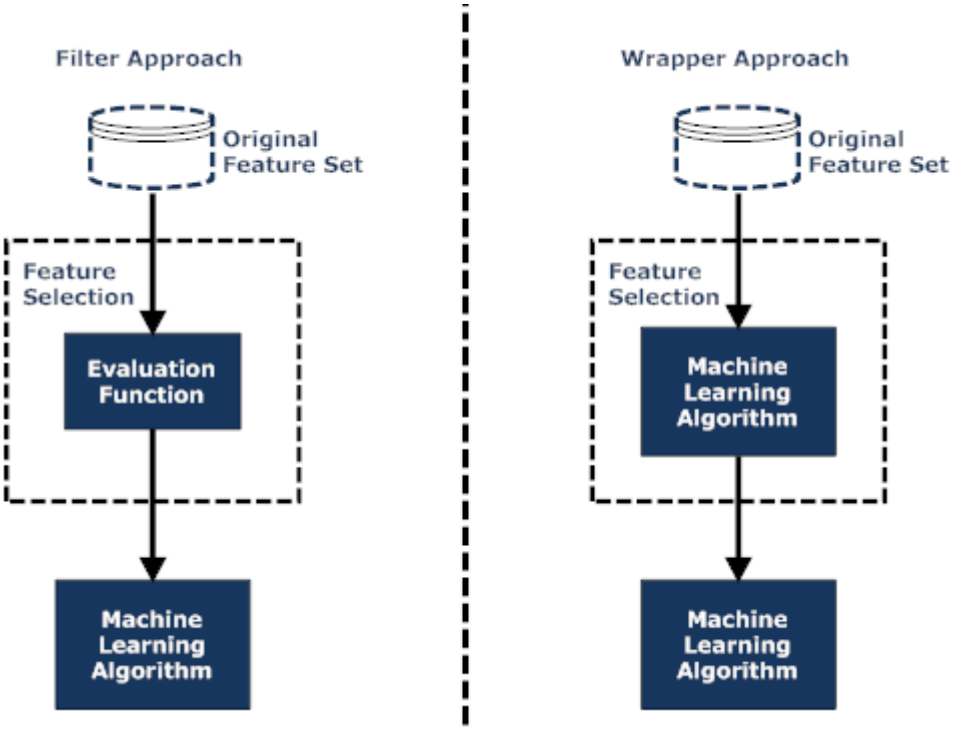


Figure 3.4 Filter and wrapper approaches for feature selection.

There are several studies that combine filter and wrapper approaches for feature subset selection as a hybrid model (Jashki et al., 2009, Yang et al., 2009, Ataş et al., 2011b, Dash and Liu, 2011). In addition to this, there exist numerous evaluation functions playing crucial roles in the feature subset selection process. Some of them are as follows;

1. **Distance Measure**

- **Euclidean Distance**

$$d_{pq} = \sqrt{\sum_{k=1}^{n}(x_{pk} - x_{qk})^2} \qquad \text{(Equation 3.3)}$$

where, *p* and *q* denotes two data points and *k* refers to *n* dimensional space. d_{pq} is the Euclidean distance between *p* and *q* data points in *n* dimensional space.

- **Fisher Distance**

Fisher discriminant was first proposed by Fisher, R.A. (1936) and is utilized in various studies. Fisher discrimination projects data from *n*-dimensional space to a one-dimensional space where between-class scatter is maximum and within-class scatter is minimum. It can be computed as;

$$F_{dp} = \frac{|\mu_1 - \mu_2|^2}{\sigma_1^2 + \sigma_2^2} \qquad \text{(Equation 3.4)}$$

where μ and σ are the mean and standard deviation of the population of each class.

2. **Dependency Measure**

- **Pearson Correlation Coefficient**

It measures the degree of correlation between two features. High correlation (dependency) amongst features implies redundancy whereas it indicates high relevance if class label is used instead of the other feature. Suppose we have *n* observations of variables *X* and *Y* denoted as x_i and y_i where *i=1,2,…. ,n.* The correlation coefficient P_{cor} between these two variables can be computed as

$$P_{cor}(x,y) = \frac{cov(X,Y)}{\sigma_x \sigma_y} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2 \sum_{i=1}^{n}(y_i-\bar{y})^2}} \qquad \text{(Equation 3.5)}$$

3. **Information Theoretic Approach**

- **Minimum Redundancy Maximum Relevance**

Various studies based on information theoretic criteria have been proposed (Hall, 1998, Zhao et al., 2002, Guyon and Elisseeff, 2003, Robnik-Sikonja and Kononenko, 2003, Forman, 2003, Bekkerman et al., 2003, Peng et al., 2005b). Correlation coefficient or Fisher based evaluation functions examine the discrimination power of features individually without considering complementarities of the features to each other. Especially for hyperspectral data, there might be excessive redundancies amongst features. mRMR method tries to minimize the mutual information through features while preserving high correlation to the class labels (Peng et al., 2005b, Peng et al., 2005a). To be more specific, mRMR tries to maximize the difference between maximum relevance and the minimum redundancy value as

$$max_{i \in S}[I(i,h) - \frac{1}{|S|}\sum_{j \in S} I(i,j)] \qquad \text{(Equation 3.6)}$$

Where *h* is the target class, *S* is the set of features, *I(i,h)* denotes mutual information of *i*'th feature and the class label. Similarly, *I(i,j)* is the mutual information between features *i* and *j*. For discrete variables, Mutual Information (MI) between features *X* and *Y* is defined as

$$I(X;Y) = \sum_{y\in Y}\sum_{x\in X} p(x,y) log\left(\frac{p(x,y)}{p_1(x)p_2(y)}\right) \quad \text{(Equation 3.7)}$$

Maximum relevance can be expressed as

$$max\, V_I,\ V_I = \frac{1}{|S|}\sum_{i\in S} I(h,i) \quad \text{(Equation 3.8)}$$

Minimum redundancy is defined as

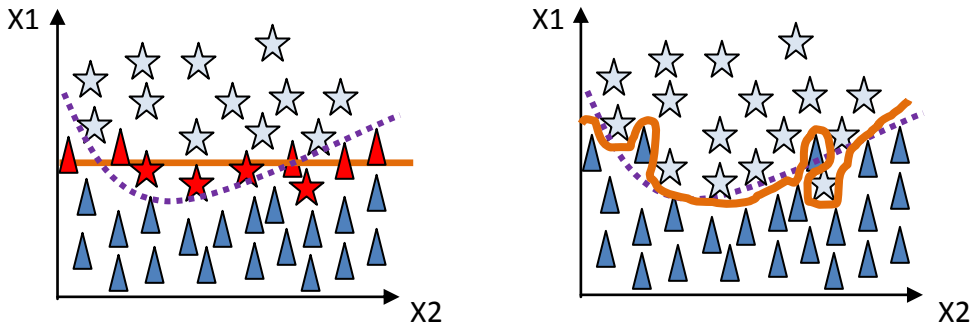
$$max\, W_I,\ W_I = \frac{1}{|S|^2}\sum_{i,j\in S} I(i,j) \quad \text{(Equation 3.9)}$$

3.3.3 Stopping Criterion

The feature subset selection process is finalized when some predefined condition is satisfied. This can be reaching a predetermined number of trials or achieving minimum classification error.

3.3.4 Bias Variance Tradeoff

A high dimensional model is more likely to over-fit the training data which in turn degrades the generalization performance. This increases the variance of prediction. Likewise, decreasing the number of parameters increases predictor's bias. This is known as Bias-Variance dilemma. **Figure 3.5** demonstrates over-fitting and under-fitting phenomena. The default optimal decision function is depicted as the dotted line. As it seen in **Figure 3.5 a**, using simple linear classifiers results in high number misclassified data points for each side which indicates that model cannot learn from the training samples or simply it under-fits the data. In **Figure 3.5 b** a complex nonlinear classifier with high number of parameters classifies whole data points correctly. Yet it is poor in generalization performance with respect to unseen test dataset (over-fitting).



a b

Figure 3.5 Under-fitting and Over-fitting effects are occurred with a) simple linear classifier, b) complex nonlinear classifier

As the number of parameters increases, variance term also rises but the bias value decreases. **Figure 3.6** summarizes this compromise between bias and variance. Desirable models are the ones which have both small bias and variance values. Derivation of the bias-variance tradeoff can be accessed at (Meyer, 2008).

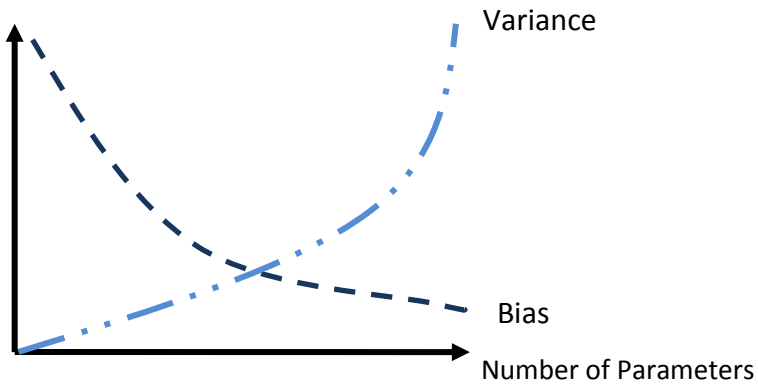


Figure 3.6 Bias – Variance tradeoff

3.3.5 Validation

Generalization performance is the capability of the learning models to accomplish desired prediction on unseen test data (Hastie et al., 2009, Meyer, 2008). In order

to assess the generalization performance of the learning models and the feature subsets, validation is performed. Cross validation is commonly accepted and used technique in data mining and machine learning communities for performance evaluation and model selection (Refaeilzadeh et al., 2009). Cross validation is a statistical method to evaluate and compare learning algorithms by dividing the data set into non-overlapping subsets called train and validation/test set. Typically, train and validation/test subsets must cross over so that they can be validated against each other at successive rounds (Refaeilzadeh et al., 2009). Broadly, validation techniques are divided into three groups.

1. **Resubstitution Method**: Learning algorithm trains on the whole dataset and then validation process runs on the same data set. This validation method suffers from over-fitting in the case of limited data. It is admissible if we had unlimited number of samples in the dataset. Note that, training error eventually converges to generalization error in the case of unlimited number of training samples.

2. **Hold-out**: This method is more suitable for large dataset again. Dataset is divided into a training and validation set with equal size. Training process proceeds only on the training set. Validation set is held out and not used during the training phase. A risk of using hold-out method is that, it does not utilize the entire dataset so the results are biased to the samples in the train and validation sets for limited data. Some instances in the test set may not be generalized from the training samples.

3. **K-Fold Cross Validation**: It is the most popular and commonly used method for performance assessment in data mining. As **Figure 3.7** depicts, first entire dataset is decomposed into K partitioned segments. Then, *K* iterations of training and validation are realized such that at each iteration, a different

segment is held for validation. Next, the training process proceeds on the remaining K-1 groups. Average of the *K* trials is accepted as the generalization performance of the current predictor. The data normally is stratified before splitting into *K* folds. Stratification is the process of reorganizing the training and validation dataset so that they have the same class distribution as the original dataset. For example, if an original dataset comprises a dispersion of 70% class-1 and 30% class-2, then train and validation sets should also have the same ratio. A special case of the K-Fold cross validation is leave one out cross validation (LOO-CV) where *K* is equal to the number of samples in the dataset. LOO-CV is reported as an unbiased estimator but has high variance (Efron, 1983). Another important drawback of LOO-CV is that, it may be computationally expensive particularly for the non-linear classifiers which makes search the intractable for the relatively large datasets (Hastie et al., 2009).

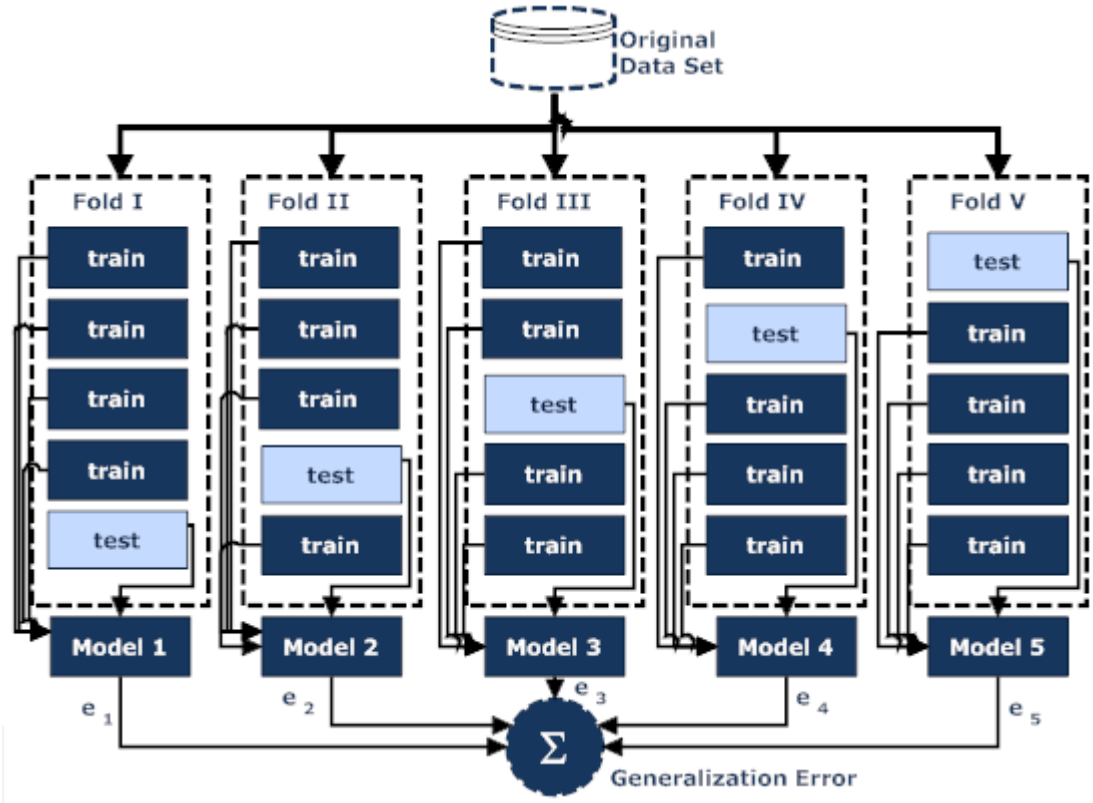


Figure 3.7 Typical illustration of 5-fold cross validation procedure.

3.4 Machine Learning Algorithms

We briefly cover here the main aspects of the three popular learning models namely, the linear discriminant classifier (LDC), support vector machines (SVM) and finally artificial neural network (ANN). Machine learning algorithms can be compared among themselves firstly according to their prediction performances. Convergence speeds, computational complexities, and interpretability of the results are other significant criteria (Harper, 2005). Additional information can be accessed from these sources, (Michie et al., 1994, Bishop, 1995, Haykin, 1999, Alpaydın, 2004, Witten and Frank, 2005).

3.4.1 Linear Discriminant Classifier

LDC is used only for categorical classification problem and it is not suitable for regression problem. LDC is very efficient in terms of computational cost for high dimensional and large datasets provided that the feature dimension does not exceed the sample size. A weakness of LDC is that, it yields poor results when the means of two classes are close to each other.

Given a linearly separable two-class problem, LDC tries to find a linear boundary (best hyper plane)

$$g(x) = w_0 + \sum_{i=1}^{d} w_i x_i \qquad \text{(Equation 3.10)}$$

that, divides the feature space into two parts . In **Equation 3.10**, w_i denotes weight of the corresponding variable (feature) in *d* dimensional space. During the training phase the weight vector is continuously updated so that the categories are well separated from each other. **Figure 3.8** illustrates the two-class problem in two-dimensional space. As it is seen, projection of data onto either X_1 or X_2 axes does not provide good class separability. To achieve best separation of data on projection-axes, the default coordinate system is rotated until at least one axis provides good separation between classes as it is shown in **Figure 3.9**. The line which is orthogonal to X'_1 axis is the linear discriminant function of the specific two-class and two-dimensional problem. The goal is to maximize the inter class mean differences while minimizing the variances of the intra classes. More precisely, this approach is known as Fisher discrimination. It can be computed as;

$$F_{dp} = \frac{|\mu_1 - \mu_2|^2}{\sigma_1^2 + \sigma_2^2} \qquad \text{(Equation 3.11)}$$

where, μ and σ are the mean and standard deviation of the population of each class.

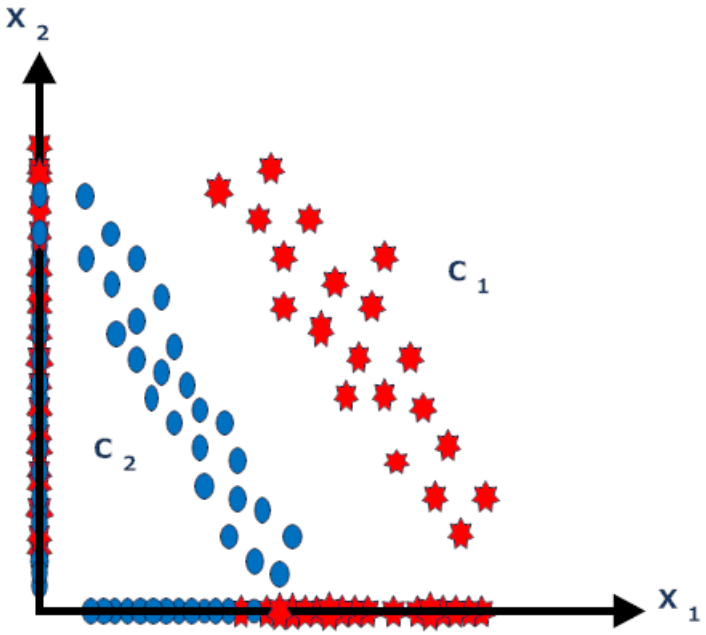


Figure 3.8 Projection of two-class data onto X_1 and X_2 axes yields overlapped separation.

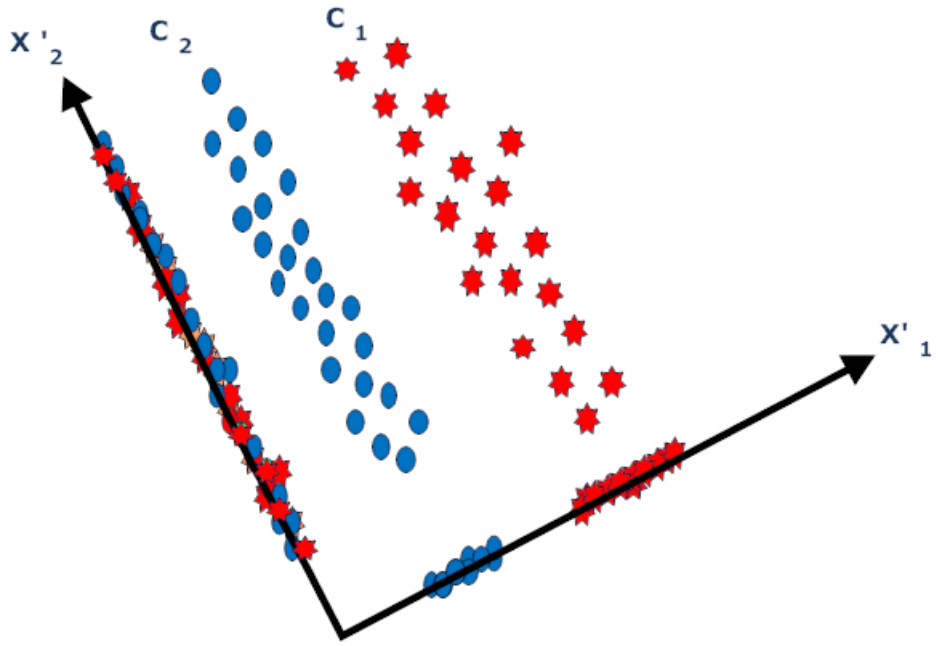


Figure 3.9 Projection of two-class data onto rotated X'_1 and X'_2 axes yields best separation on X'_1 axis and worst separation on X'_2 axis.

3.4.2 Support Vector Machines

SVM is a type of universal feed forward network (Haykin, 1999) first pioneered by Vapnik (Boser et al., 1992, Vapnik, 1995). It can be used for classification and nonlinear regression. The rationale behind the SVM is to construct the optimal hyper-plane as the decision boundary in such a way that the margin of the separation between opposite classes is maximized (Haykin, 1999). This hyper-plane satisfies the maximum margin constraint. **Figure 3.10** and **Figure 3.11** illustrate the phenomenon. The pattern vectors closest to the decision boundary in the transformed space are called as the support vectors. Basically, the decision boundary in the transformed feature space is a hyper-plane. Yet, it corresponds to a nonlinear decision surface in the original space (Lin, 1998). SVM theoretically

utilizes statistical learning theory (Vapnik, 1995) and algorithmically uses optimization techniques (Nocedal and Wright, 1999).

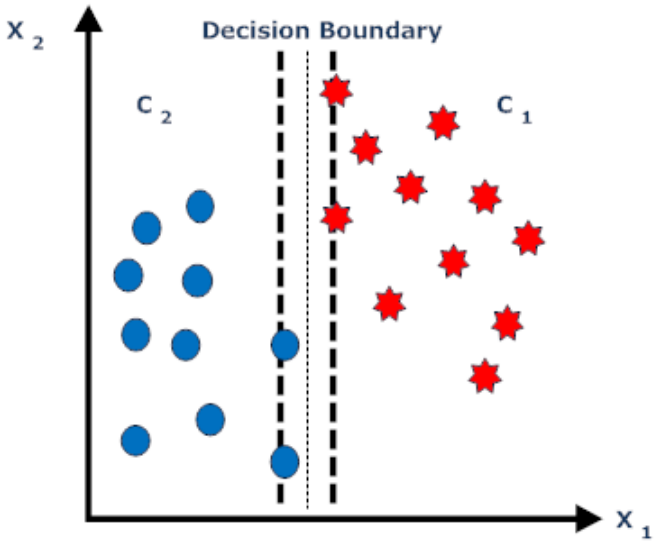


Figure 3.10 A suboptimal decision boundary for a given two-class problem.

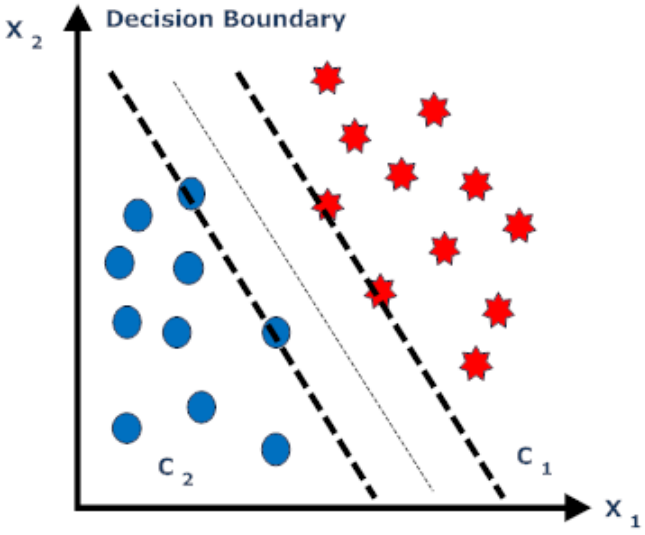


Figure 3.11 An optimal decision boundary for a given two-class problem.

As it is seen in **Figure 3.10** and **Figure 3.11**, there can be many decision boundaries which split opposite classes but we pursue the hyper-plane which has the largest margin.

For a given n number of samples, optimal decision boundary should satisfy

$$\frac{y_i F(x_i)}{\|w\|} \geq r \qquad i = 1,2,\ldots\ldots,n \qquad \text{(Equation 3.12)}$$

where, y_i denotes class label of the i'th sample, *r* designates margin of the decision boundary and $F(x)$ can be computed as

$$F(x) = w^T x + b \qquad \text{(Equation 3.13)}$$

Note that, linear SVM is not suitable for complex and non-linear problems. In order to solve nonlinear and complicated problems we should employ polynomial, RBF or Gaussian kernels in the SVM. For linearly non-separable case, the transformed feature space can be defined by a non-linear computation of base function of $\varphi_i(x)$ which is defined in input space. So, optimal hyper-plane function becomes

$$F(x) = \sum_{i=1}^{n} \alpha_i y_i K(x_i, x) + b \qquad \text{(Equation 3.14)}$$

Where, $\alpha_i$is the Lagrange multiplier, y_i denotes class label, $K(x_i, x)$ equals to inner product kernel of base functions and x_i is the input data (Antkowiak, 2006).

Commonly used kernel functions are:

- Linear: $K(x_i, x_j) = {x_i}^T x_j$ (Equation 3.15)
- Polynomial: $K(x_i, x_j) = ({x_i}^T x_j)^d$ (Equation 3.16)
- Gaussian: $K(x_i, x_j) = exp(-\frac{\|x_i - x_j\|^2}{2\sigma^2})$ (Equation 3.17)
- Sigmoid: $K(x_i, x_j) = tanh(\beta_0 {x_i}^T x_j + \beta_1)$ (Equation 3.18)

3.4.3 Artificial Neural Networks

ANN comprises many simple processing elements whose function is determined by network topology, synaptic weights and nodes (Pearson and Lippmann, 1988). More specifically, those interconnected units are the abstraction of the biological neuronal system in such a way that, the neural network mimics the operation of the brain in two ways: first, information is acquired by the network from its environment through the input signal at the learning phase, and the second, acquired knowledge is stored in the interneuron connection strengths which are also known as connection weights (Haykin, 1999). ANN is regarded as a powerful statistical tool for capturing the complex input output relations without using the domain knowledge. ANN studies began with the cutting edge work of McCulloth and Pits (1943). Provided a sufficient number of simple units (neurons) and proper set of synaptic connections, a single layer network in principle can compute any computable function (McCulloch and Pitts, 1943, Haykin, 1999). However it should be noted that information and processing capability of the single layer network is very limited (Antkowiak, 2006). The second major improvement was the so called *perceptrons convergence theorem* by Rosenblatt (1958) and is regarded as a novel method of supervised learning. In 1960, Widrow and Hoff proposed the least mean square (LMS) algorithm and used it in adaptive linear element (ADALINE). ANN is extensively studied for a wide range of areas including, pattern and object recognition, regression, function approximations, medical diagnosis, data mining applications. ANN with a single layer and neuron is called as a *perceptron*. Perceptron is similar to the linear classifiers. On the other hands, an ANN with multiple layers and a series of neurons is known as a multi layer perceptron (MLP).

As it is shown in **Figure 3.13**, a simple neuronal model starts its operation by accepting inputs signal x_i from the input layer, which is then multiplied by

connection weights w_i and the sum of the weighted inputs is transformed through either a linear or non-linear activation function into the output value of *y*.

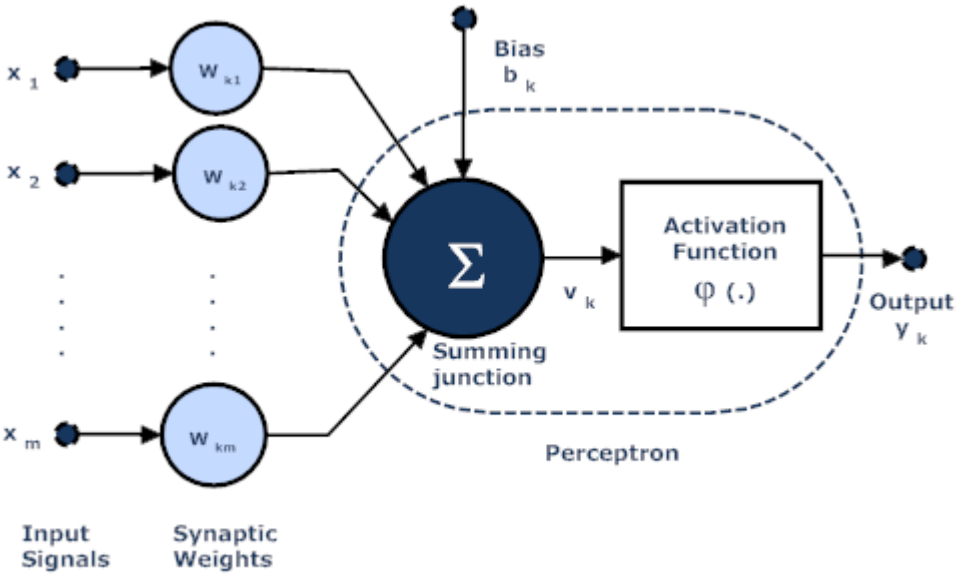


Figure 3.13 A learning model of the perceptron.

As **Figure 3.13** indicates, we can describe the output of the neuron *k* by the following equations

$$v_k = \sum_{j=1}^{m} w_{kj} x_j \qquad \text{(Equation 3.19)}$$

and

$$y_k = \varphi(v_k + b_k) \qquad \text{(Equation 3.20)}$$

where, $x_1, x_2, \ldots\ldots, x_m$ are the input signals, w_{kj} are the connection weights of neuron k, v_k is the weighted sum of the product of input signals and connection weights (also called as induced local fields (Haykin, 1999)), b_k is the bias value and $\varphi(.)$ is the activation function. Commonly used activation functions are as follows;

1. **Threshold function**: $\varphi(v) = \begin{cases} 1, & v \geq 0 \\ 0, & v < 0 \end{cases}$ (Equation 3.21)

2. **Piecewise-Linear**: $$\varphi(v) = \begin{cases} 1, & v \geq +\frac{1}{2} \\ v, & +\frac{1}{2} \geq v \geq -\frac{1}{2} \\ 0, & v \leq -\frac{1}{2} \end{cases}$$ (Equation 3.22)

3. **Sigmoid function**: $$\varphi(v) = \frac{1}{1+e^{-\beta v}}$$ (Equation 3.23)

Where, *β* is a given parameter which modifies the inclination of the sigmoid function.

4. **Hyperbolic tangent function**: $\varphi(v) = a\ tanh(bv)$ (Equation 3.24)

Where, *a* and *b* are constants. Suggested values for a and b are *1.7159* and *2/3,* respectively (LeCun, 1993, LeCun, 1989).

Information processing capacity of the single layer perceptrons is limited. Therefore, in order to address nonlinear and complicated problems multi layer perceptrons was developed. MLP consists of one input layer, one or more hidden layers and single output layer. Each layer may contain one or more neurons. Typical illustration of the MLP is demonstrated in **Figure 3.14**.

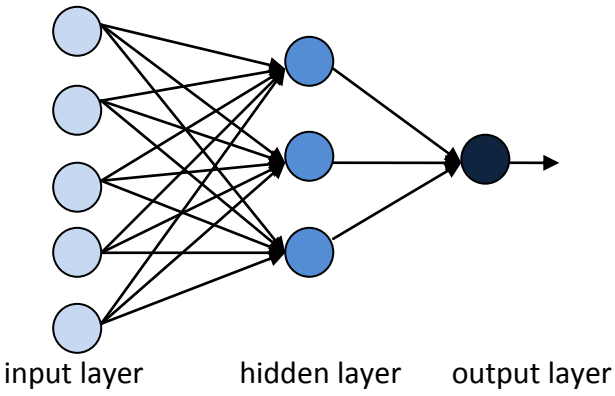


Figure 3.14 One hidden layered MLP

3.4.3.1 Training of the MLP by Back Propagation Algorithm

Basically the process of training of the MLP consists of two passes which are forward pass and backward pass. In the forward pass, the training pattern is fed into the input layer and then nodes of each layer apply the activation function to produce the output as an actual response of the network and this process is continued through the network, layer by layer (Haykin, 1999). During the forward pass, all synaptic connection weights are frozen. In the backward pass, on the other hand, network connection weights are adjusted according to the back-propagation algorithm. To be more specific, the final output is subtracted from the desired output value to form an error signal which is then propagated through the network in the opposite direction of the synaptic connections (Haykin, 1999). **Figure 3.15** illustrates these two passes in MLP (Haykin, 1999). Normal arrows indicate the forward pass, whereas dashed lines depict propagation of the error signal through the MLP network.

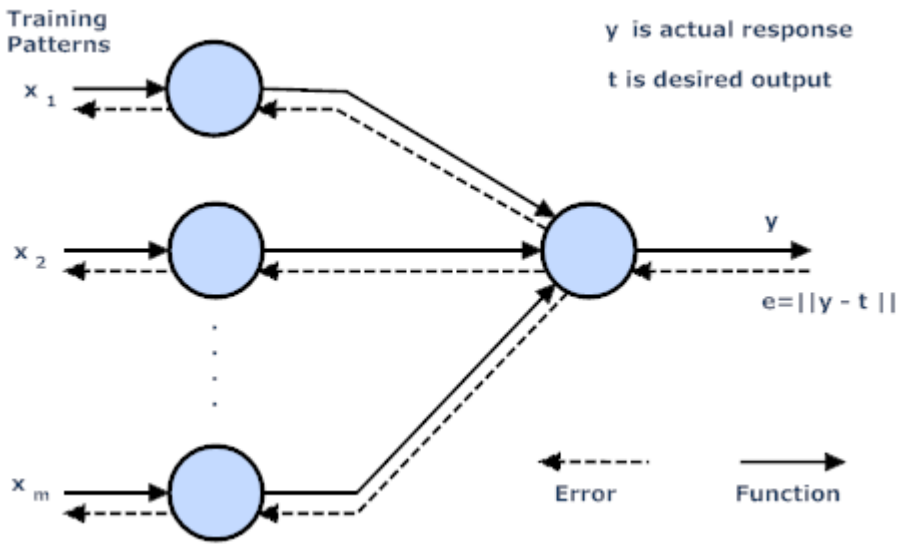


Figure 3.15 Simplified illustrations of forward flow of function signals and backward propagation of error signals in the MLP network.

Thus, each neuron in the network is directed to perform two computations: **Equation** *3.19* is used for the forward pass to compute the output, whereas **Equation** 3.24 is utilized for the gradient vector calculation.

The error signal at the output of the *j*'th output neuron at *n*'th iteration can be expressed as

$$e_j(n) = d_j(n) - y_j(n) \qquad \text{(Equation 3.25)}$$

where, $d_j(n)$ denotes the desired target output and $y_j(n)$ is the current output of the network. One can define error energy as $\frac{1}{2}e^2(n)$. So the total instantaneous error energy at the output layer becomes,

$$\xi(n) = \frac{1}{2}\sum_{j\in C} e_j^2\,(n) \qquad \text{(Equation 3.26)}$$

Where, *C* represents the set of all neurons in the output layer. Let the number of instances in the training set be denoted by *N*. Then average squared error energy can be defined as,

$$\xi_{avg}(n) = \frac{1}{N}\sum_{n=1}^{N} \xi(n) \qquad \text{(Equation 3.27)}$$

Here, ξ_{avg} is a measure of the learning performance. In order to minimize ξ_{avg}, the free parameters of the network is adjusted. For this minimization process, Least Mean Square (LMS) algorithm is employed. Derivation and the details of the LMS algorithm can be found in Haykin (1999). To be more specific, synaptic weights are updated on a pattern-by-pattern basis (for each sample in the training set) until the presentation of all patterns is completed which is also known as one epoch (Haykin, 1999).

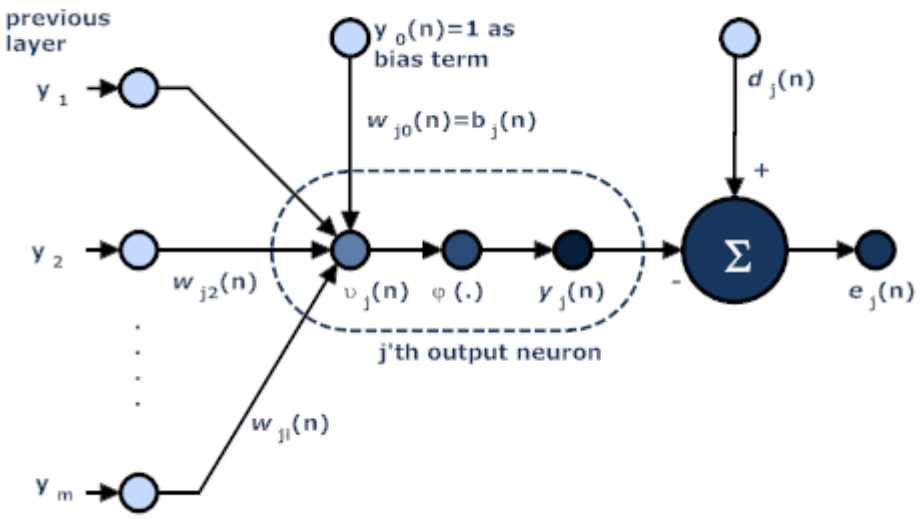


Figure 3.16 Process at the output neuron.

According to **Figure 3.16**, induced local field $v_j(n)$ produced for the activation function $y_j(n)$ at the n'th iteration with neuron *j* can be defined as:

$$v_j(n) = \sum_{i=0}^{m} w_{ji}(n) y_i(n) \qquad \text{(Equation 3.28)}$$

where, m is the total number of nodes in the preceding layer. Applying activation function $\varphi_j(.)$ to the induced local field we get

$$y_j(n) = \varphi_j(v_j(n)) \qquad \text{(Equation 3.29)}$$

As in the LMS algorithm, the backpropagation algorithm also minimizes the error $\xi(n)$ with respect to the connection weights as

$$\partial \xi(n) / \partial w_{ji}(n) \; . \qquad \text{(Equation 3.30)}$$

By applying the chain rule, error gradient can be expressed as

$$\frac{\partial \xi(n)}{\partial w_{ji}(n)} = \frac{\partial \xi(n)}{\partial e_j(n)} \frac{\partial e_j(n)}{\partial y_j(n)} \frac{\partial y_j(n)}{\partial v_j(n)} \frac{\partial v_j(n)}{\partial w_{ji}(n)} \quad \text{(Equation 3.31)}$$

By differentiating both sides of **Equation 3.26** with respect to $e_j(n)$ yields

$$\frac{\partial \xi(n)}{\partial e_j(n)} = e_j(n) \quad \text{(Equation 3.32)}$$

Similarly, if we differentiate both sides of **Equation 3.25** with $y_j(n)$ results in

$$\frac{\partial e_j(n)}{\partial y_j(n)} = -1 \quad \text{(Equation 3.33)}$$

For the third term of **Equation 3.31**, we differentiate output $y_j(n)$ with respect to induced local fields $v_j(n)$ and we get

$$\frac{\partial y_j(n)}{\partial v_j(n)} = \frac{\partial \varphi_j(v_j(n))}{\partial v_j(n)} = \varphi_j'(v_j(n)) \quad \text{(Equation 3.34)}$$

Finally, differentiation of the induced local field $v_j(n)$ is by $w_{ji}(n)$ outputs

$$\frac{\partial v_j(n)}{\partial w_{ji}(n)} = y_j(n) \quad \text{(Equation 3.35)}$$

Substitute all terms into **Equation 3.31** we get

$$\frac{\partial \xi(n)}{\partial w_{ji}(n)} = -e_j(n)\varphi_j'(v_j(n))y_j(n) \quad \text{(Equation 3.36)}$$

From the delta rule also known as gradient descent rule, error correction value of corresponding $w_{ji}(n)$ should be in the negative descent direction in the weight space so that

$$\Delta w_{ji}(n) = -\eta \frac{\partial \xi(n)}{\partial w_{ji}(n)} \quad \text{(Equation 3.37)}$$

where, η is the learning rate. From **Equation 3.37** we can generalize the error correction value of connection weight as

$$\Delta w_{ji}(n) = \eta e_j(n) \varphi_j'(v_j(n)) y_j(n) \quad \text{(Equation 3.38)}$$

In this way, second pass of the training of the MLP is computed and propagated to opposite direction in order to update the associated connection weights.

In this thesis, we preferred to use a feed forward back propagated one hidden layered artificial neural network also known as the MLP. MLP is extensively used in various studies and is regarded as the universal approximator of any continuous function (Hornik, 1989). As seen in (Hornik, 1989, Bochereau, 1992, Jayas, 2000, ElMasrya et al., 2009) ANN is very efficient for identification and classification of agricultural product containing non-linearity. In particular, MLP is superior to linear classifiers in terms of prediction accuracy for the classification of kiwi fruit berries (Kim, 2000). Again, Park and Chen (1996) utilized ANN with a spectral imaging technique and successively achieved 93.3% generalization performance for classifying wholesome chicken carcasses from unwholesome ones (Park, 1996). Thus, we selected MLP as the central classifier in this dissertation.

CHAPTER 4

GENERATION OF DATASETS

4.1 Preparation of the Chili Pepper Samples

In total, 53 ground red chili pepper flake samples were gathered from different regions in Turkey. Most of them were sold as unpackaged. **Figure 4.1** shows their aflatoxin levels. As 10 ppb is the upper level for aflatoxin for spices and herbs in the EU (Commission Regulation [EC], 2006) we used 10 ppb as the threshold to classify the pepper samples into aflatoxin positive and negative groups. As **Table 4.1** indicates, the mean aflatoxin level was measured as 16.78 ppb. Average aflatoxin levels for Afl- and Afl+ groups are 3.2 ppb and 33.3 ppb, respectively. All the pepper samples were sent to TUBITAK Ankara Testing and Analyses Laboratory (ATAL) for HPLC analysis. Chili pepper samples that exceed 10 ppb threshold were labeled as aflatoxin positive otherwise they were labeled as aflatoxin negative for inductive learning. **Table 4.2** demonstrates origins of the obtained chili pepper samples and associated aflatoxin levels.

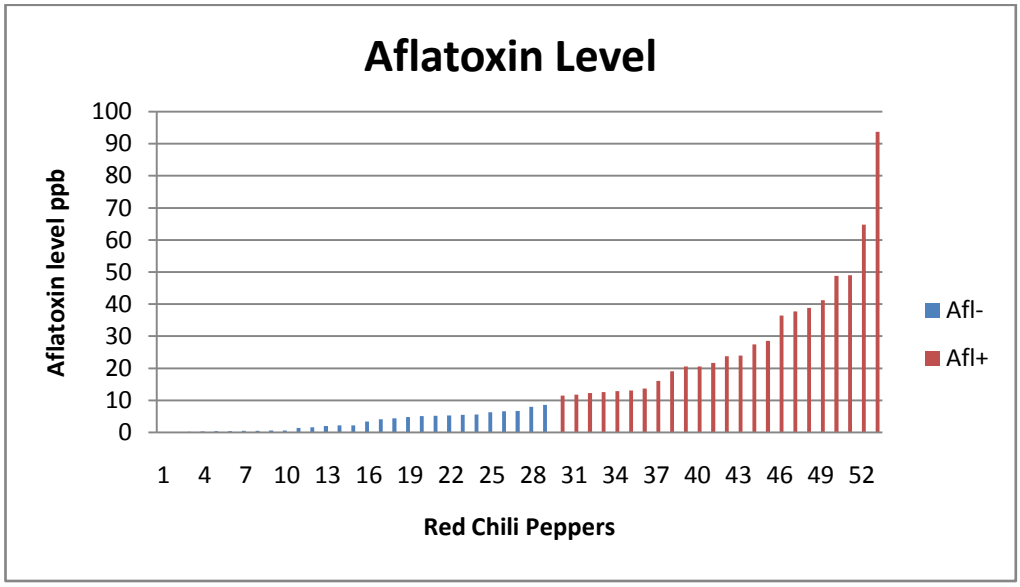


Figure 4.1 Aflatoxin levels of chili pepper samples

Table 4.1: Average aflatoxin levels for Afl+ and Afl- groups.

	# samples	Mean Aflatoxin Level (ppb)
Afl+	24	33.30
Afl-	29	3.20
Total	53	16.75

Table 4.2: Names and corresponding aflatoxin values of the chili pepper samples employed in this thesis.

Pepper Samples	Illumination	Re sampling	aflatoxin value (ppb)
Arifoglu	uv+halogen	3 locations	1.63
KM_7	uv+halogen	3 locations	0.10
Bagdad_1	uv+halogen	3 locations	0.39
Ankara_1	uv+halogen	3 locations	0.61
Ankara_2	uv+halogen	3 locations	27.47
Ankara_3	uv+halogen	3 locations	193.72
Ankara_4	uv+halogen	3 locations	0.64
Ankara_5	uv+halogen	3 locations	0.53
Ankara_6	uv+halogen	3 locations	0.37
Ankara_7	uv+halogen	3 locations	2.19
Antep_1	uv+halogen	3 locations	20.52
Antep_2	uv+halogen	3 locations	0.22

Table 4.2 (cont.)

Antep_3	uv+halogen	3 locations	5.18
Antep_4	uv+halogen	3 locations	4.08
Antep_5	uv+halogen	3 locations	12.59
Diyarbakir_7	uv+halogen	3 locations	13.09
Diyarbakir_8	uv+halogen	3 locations	64.80
Diyarbakir_9	uv+halogen	3 locations	3.44
Diyarbakir_10	uv+halogen	3 locations	6.61
Diyarbakir_11	uv+halogen	3 locations	23.94
Diyarbakir_12	uv+halogen	3 locations	10.78
Diyarbakir_13	uv+halogen	3 locations	8.57
Diyarbakir_14	uv+halogen	3 locations	19.08
Erzurum_1	uv+halogen	3 locations	0.10
Erzurum_2	uv+halogen	3 locations	2.02
Erzurum_3	uv+halogen	3 locations	13.68
Erzurum_4	uv+halogen	3 locations	5.27
Erzurum_5	uv+halogen	3 locations	1.35
KM_1	uv+halogen	3 locations	5.48
KM_2	uv+halogen	3 locations	28.56
KM_3	uv+halogen	3 locations	4.83
KM_4	uv+halogen	3 locations	37.73
KM_5	uv+halogen	3 locations	49.05
KM_6	uv+halogen	3 locations	6.24
Mardin_1	uv+halogen	3 locations	48.82
Mardin_2	uv+halogen	3 locations	5.60
Mardin_3	uv+halogen	3 locations	5.05
Mardin_4	uv+halogen	3 locations	21.69
Mardin_5	uv+halogen	3 locations	0.51
Mersin_1	uv+halogen	3 locations	12.87
Mersin_2	uv+halogen	3 locations	0.32
Mersin_3	uv+halogen	3 locations	6.68
Mersin_4	uv+halogen	3 locations	4.38
Mersin_5	uv+halogen	3 locations	41.22
Siirt_1	uv+halogen	3 locations	11.24
Siirt_2	uv+halogen	3 locations	16.04
Van_1	uv+halogen	3 locations	10.48
Van_2	uv+halogen	3 locations	20.51
Van_3	uv+halogen	3 locations	2.18
Van_4	uv+halogen	3 locations	8.03
Van_5	uv+halogen	3 locations	38.81
Van_6	uv+halogen	3 locations	36.45
Van_7	uv+halogen	3 locations	23.73

4.2 Acquisition of the Hyperspectral Data

In the previous studies (Kalkan et al. 2011, Yao et al., 2006, Pearson et al., 2001), a single illumination source was used. More specifically, some studies were performed only under halogen illumination whereas others were done under UV. UV illumination is utilized for the fluorescence and halogen excitation is for reflectance phenomena. In order to investigate the contribution of those illuminations on the classifier performance, we utilized both excitations in this work. As previous studies stated that BGYF is observed when aflatoxin level in the sample is high, we expect halogen illumination to contribute more for detection of aflatoxin for our chili pepper problem. **Figure 1.2** in Chapter 1 depicts a general overview of the hyperspectral imaging system.

The hardware of the image acquisition system is composed of a Sony FireWire CCD camera with Varispec liquid crystal electronically tunable filter (VLCETF) assembly. Hyperspectral image series ranging from 400 nm to 720 nm (10 nm widths) of 53 different chili pepper samples have been acquired under 100W quartz-tungsten-halogen and UV 365 nm illumination sources. Spectra of both illumination sources are illustrated at **Figure 4.2** and **Figure 4.3** respectively.

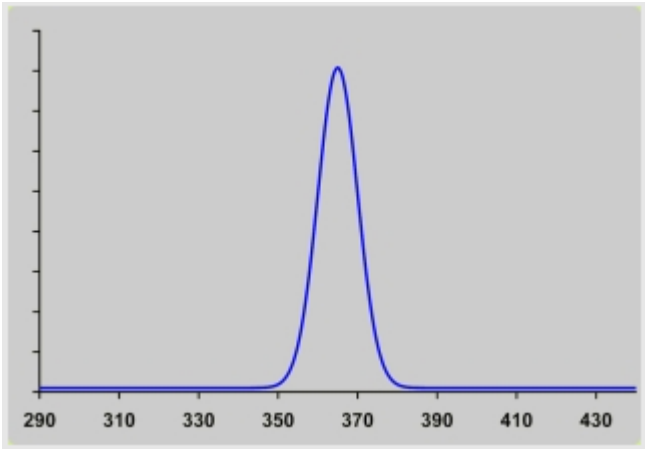


Figure 4.2 Spectrum of the 365 nm UV lamp

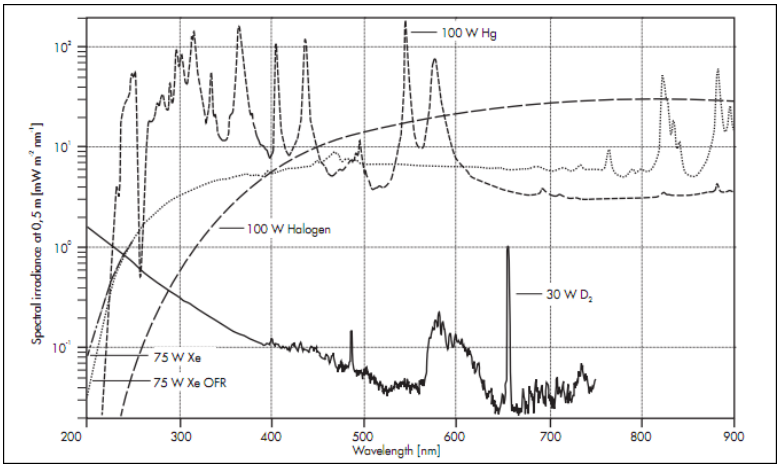


Figure 4.3 Spectrum of the 100 W Quartz-Tungsten-Halogen lamp

Resolution of each image is 1280 X 960 pixels. During the acquisition process the set up was stationary, so there was no need to register the hyperspectral images. Images of three different locations of the same chili pepper sample were obtained in order to increase the training data. This also enables us to analyze a wider surface of the chili pepper sample. **Figure 4.4** depicts representative sample images from the hyperspectral image series of uncontaminated and contaminated peppers for halogen and UV illuminations. As it is seen, the reflectance of chili pepper samples varies with spectral frequency and there is no clear evidence of aflatoxin presence when viewed by the naked eye.

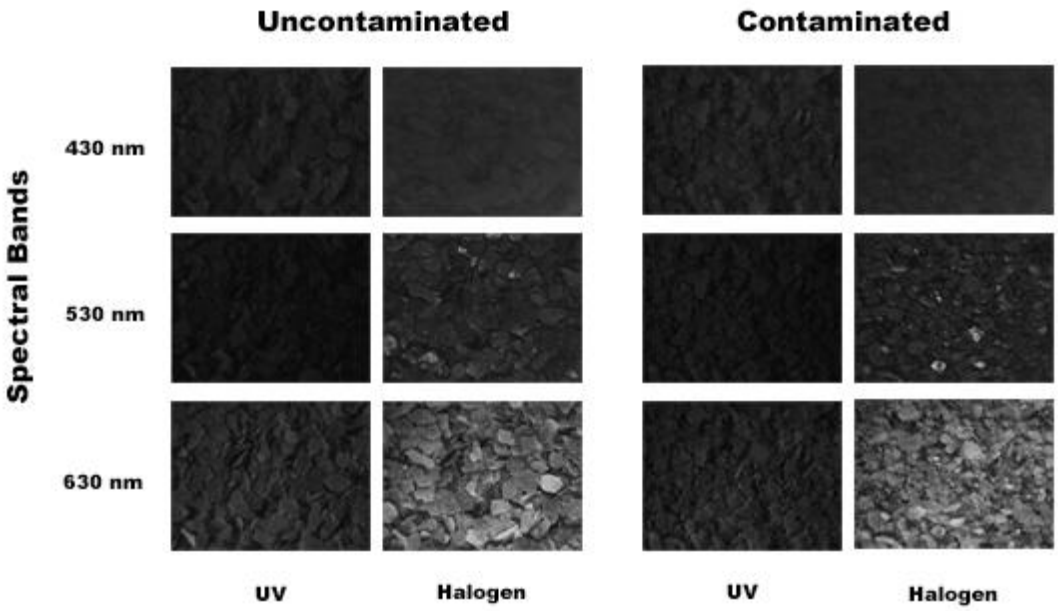


Figure 4.4 Sample images from the hyperspectral image series of uncontaminated and contaminated peppers for halogen and UV illuminations.

4.3 Preprocessing the Hyperspectral Images

Camera software by default applies histogram equalization to acquired images. Although, histogram equalization automatically controls oversaturation and under-saturation by applying adaptively changing exposure time, it also modifies original pixel values. To overcome this problem, one should fix the exposure time value of the camera and the camera gain parameter to a predefined value. On the other hand, single exposure time eventually leads to under saturated and over saturated regions in the hyperspectral image series. Therefore, we divided the spectral support into three spectral regions by manually changing exposure time values. **Table 4.3** depicts the prescribed spectral regions, exposure times and corresponding normalization coefficients. We normalized the images by their normalization coefficient before extracting the feature vector.

It should be noted that, exposure normalization leads some pixel values to exceed 255 pixel gray values and increase the dynamic range excessively. Therefore, we applied non-linear square root transformation in order to limit the range of the pixel value of the normalized images to 0-255 pixel gray value interval. With a maximum normalization coefficient of 9.0, a pixel value of 256 will yield corresponding maximum pixel gray value of 48 after the square root transformation. We used 48 levels to represent pixel gray values.

Table 4.3 Exposure normalization coefficients of the most informative regions of the spectral bands.

Illumination	Exposure time (s)	Normalization coefficient
Halogen (400-490) nm	4.5	1.0
Halogen (500-590) nm	2.4	1.875
Halogen (600-720) nm	0.5	9.0
UV (400-690) nm	9.6	1.0
UV (700-720) nm	3.1	3.09

In order to eliminate the dust particles and reduce sensor noise in the images we applied 3x3 median filtering.

CHAPTER 5

PROPOSED FEATURE EXTRACTION METHODS

5.1 Introduction

Classifier performance is strongly related to the relevance of the extracted features. In the ideal case, the feature vector should keep the most compact description of the desired function. In our problem this is the aflatoxin presence signature. Nevertheless, extracting meaningful and discriminative feature vector is not a straightforward and trivial process. It requires acquiring domain knowledge and underlying physical phenomena. In the hyperspectral images of chili pepper samples, shape and orientation of chili pepper flakes do not correlate with aflatoxin presence. Therefore, useful features should have weak relevance to the second order features like edges and orientation. Relying on solely spectral band mean intensity is not desirable either. There may be Afl- samples and Afl+ samples with nearly the same mean intensity value. In the previous studies, Kalkan et al., (2011) used wavelet based intensity features and achieved 79.2 % classification performance indicating that spatial and textural information contributes the prediction accuracy. Conversely, most of the researchers utilized spectral band energies as a feature vector in their studies and achieve reasonable accuracy rates (Hirano et al., 1998, Pearson et al., 2001, Yao et al., 2006). In this study, we propose a novel approach for feature extraction by analyzing the pixel gray value intervals in more detail. We will compare the performance of different types of features.

5.2 Energy Features of Individual Spectral Bands

Let us assume the pixel gray value located at *x*, *y* of the *k* 'th spectral band is denoted by $I_k(x, y)$. Then, individual spectral band energy (ISBE) features can be expressed as:

$$e_k = \sum_x \sum_y I_k(x, y) \quad k = 1,2, \dots \dots ,33 \qquad \text{(Equation 5.1)}$$

where, *k* denotes the spectral bands. **Figure 5.1 and Figure 5.2** demonstrate, Boxplot, Fisher discrimination power and the correlation coefficient matrix of the hyperspectral images of the 53 chili pepper samples under the halogen and UV illuminations after the Z-Score normalization. From now on, we will call the set of Boxplot, Fisher and correlation coefficient matrix figure as a COMPOSITE_VIEW. The X axis of the COMPOSITE_VIEW corresponds to the spectral bands. Boxplot demonstrates the normalized contaminated (dark color) and uncontaminated (light color) chili pepper samples distribution for the associated spectral bands from 400 to 720 nm. Plus (+) and circle signs (o) depict outliers of contaminated and uncontaminated chili pepper samples, respectively. In the middle graph, the Fisher discrimination power of each feature is plotted as a bar graph and the Fisher discrimination power and its rank is indicated at the top and the center of each bar, respectively. Finally, the Pearson correlation coefficient between all features is presented in a matrix form.

5.3 Energy Features of Absolute Difference of Consecutive Spectral Bands

Absolute difference of consecutive spectral band energy (ABSDIF_CSBE) features can be defined as

$$e_k = \sum_x \sum_y |I_{k+1}(x, y) - I_k(x, y)| \quad k = 1,2, \dots \dots ,32 \qquad \text{(Equation 5.2)}$$

where, k denotes the spectral band. As the acquisition system is kept stationary, there is no need for extra image registration. In this case, we expect to extract informative spectral signature using absolute difference of consecutive spectral bands. **Figure 5.3** and **Figure 5.4** depict COMPOSITE_VIEW of the ABSDIF_CSBE features under the halogen and UV illuminations.

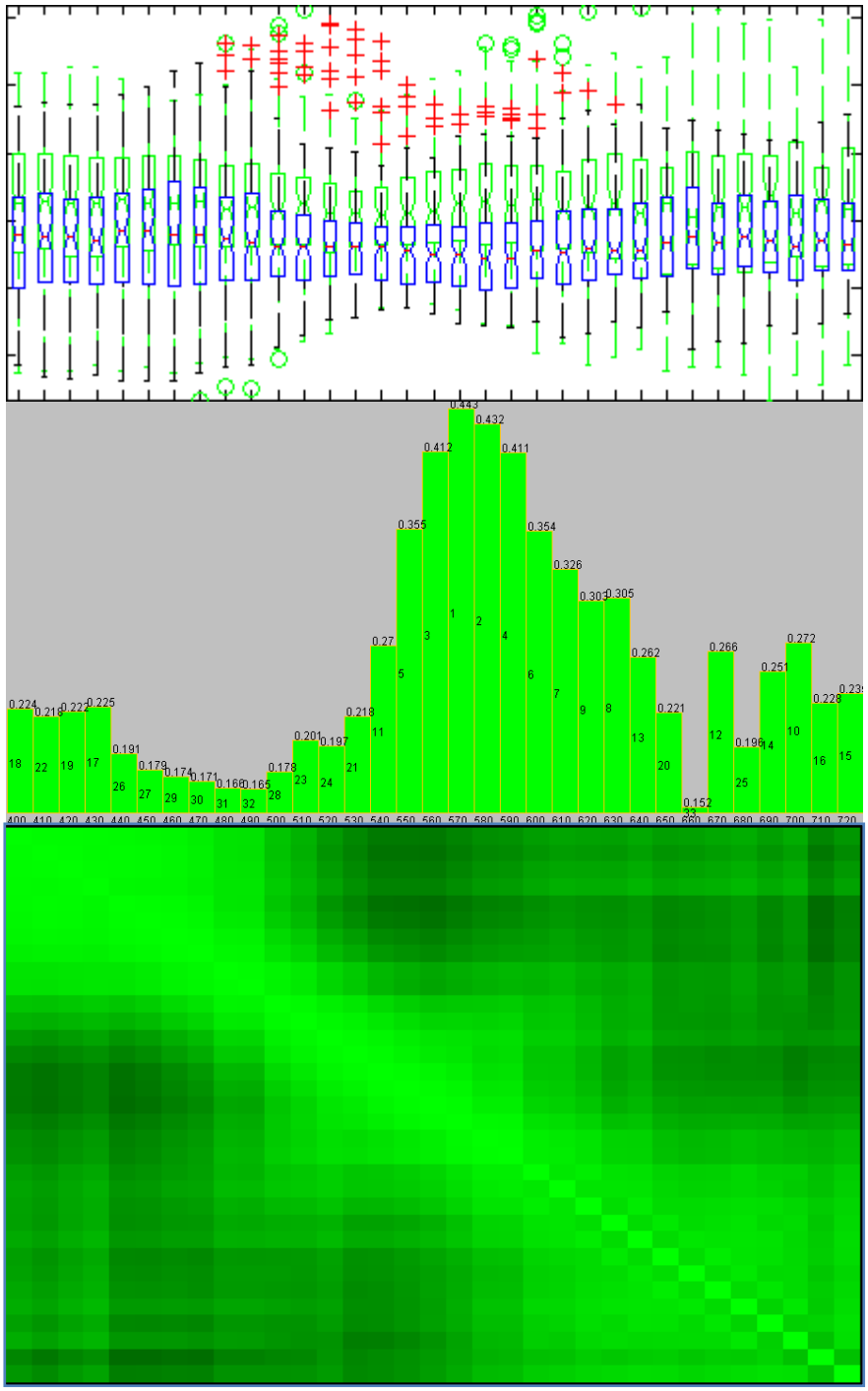


Figure 5.1 Boxplot (top), Fisher discrimination power (middle) and the correlation coefficient map (bottom) of the ISBE features for the halogen excitation.

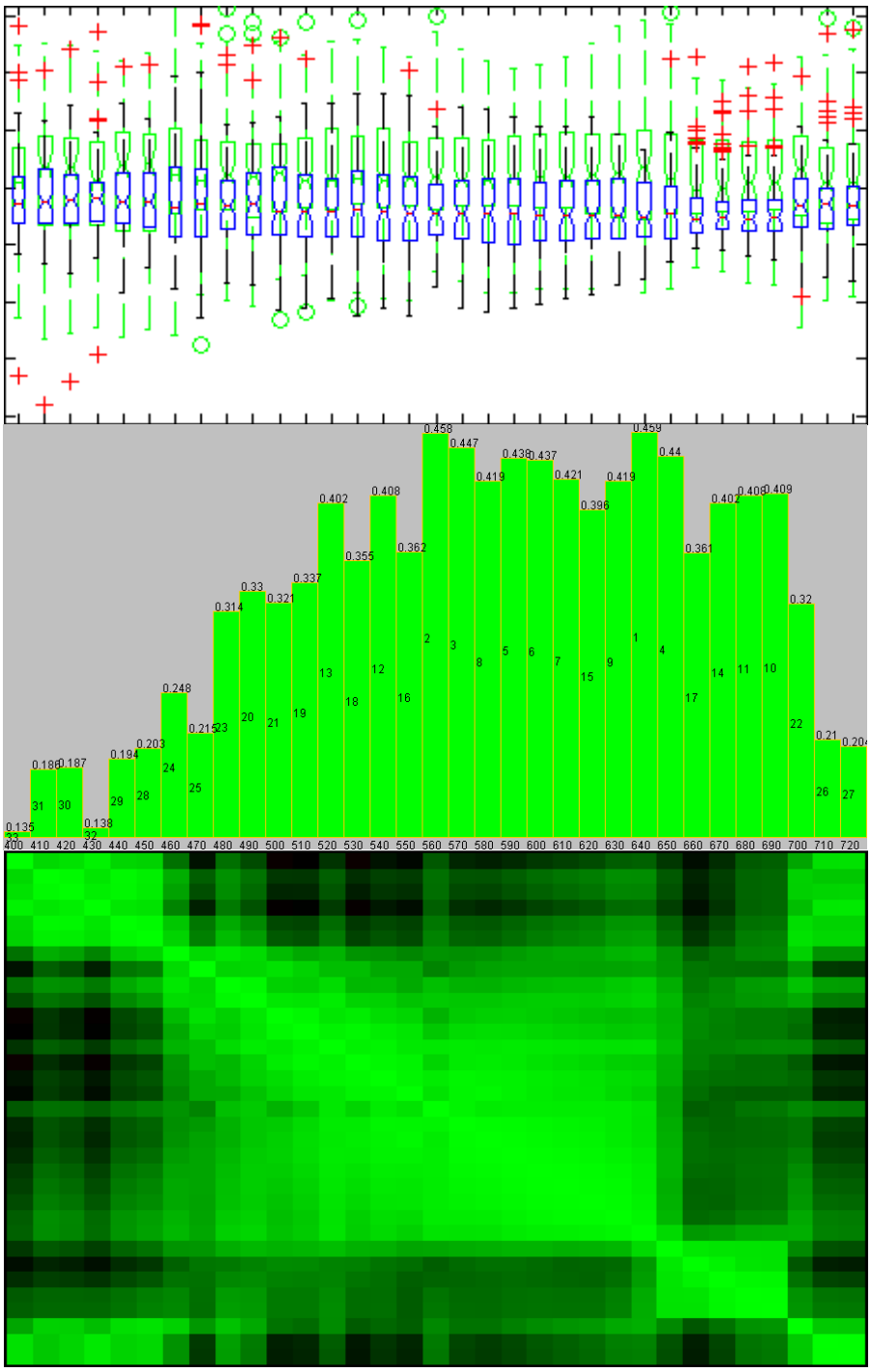


Figure 5.2 Boxplot (top), Fisher discrimination power (middle) and the correlation coefficient map (bottom) of the ISBE features for the UV excitation.

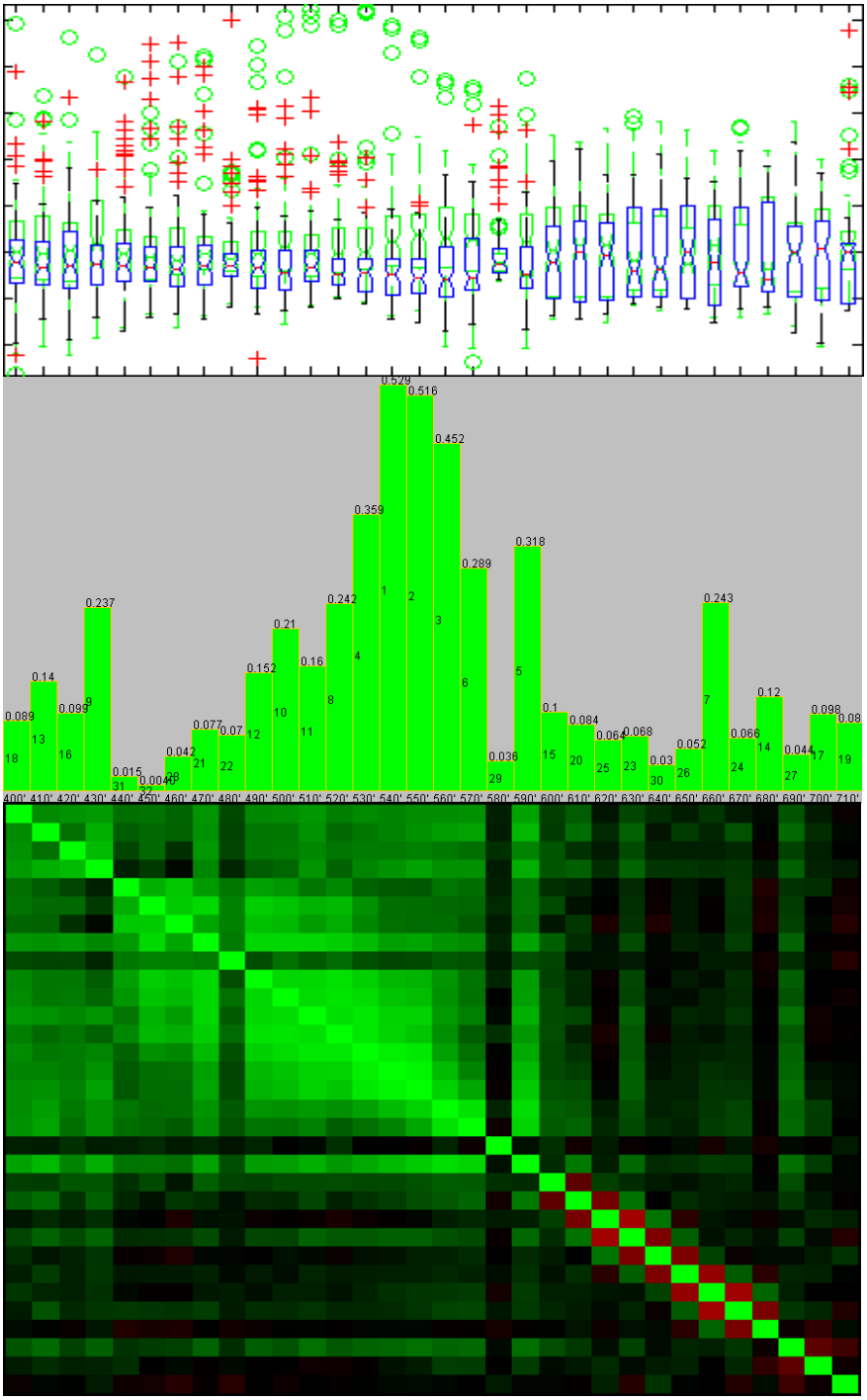


Figure 5.3 Boxplot (top), Fisher discrimination power (middle) and the correlation coefficient map (bottom) of the ABSDIF_CSBE features for the halogen excitation.

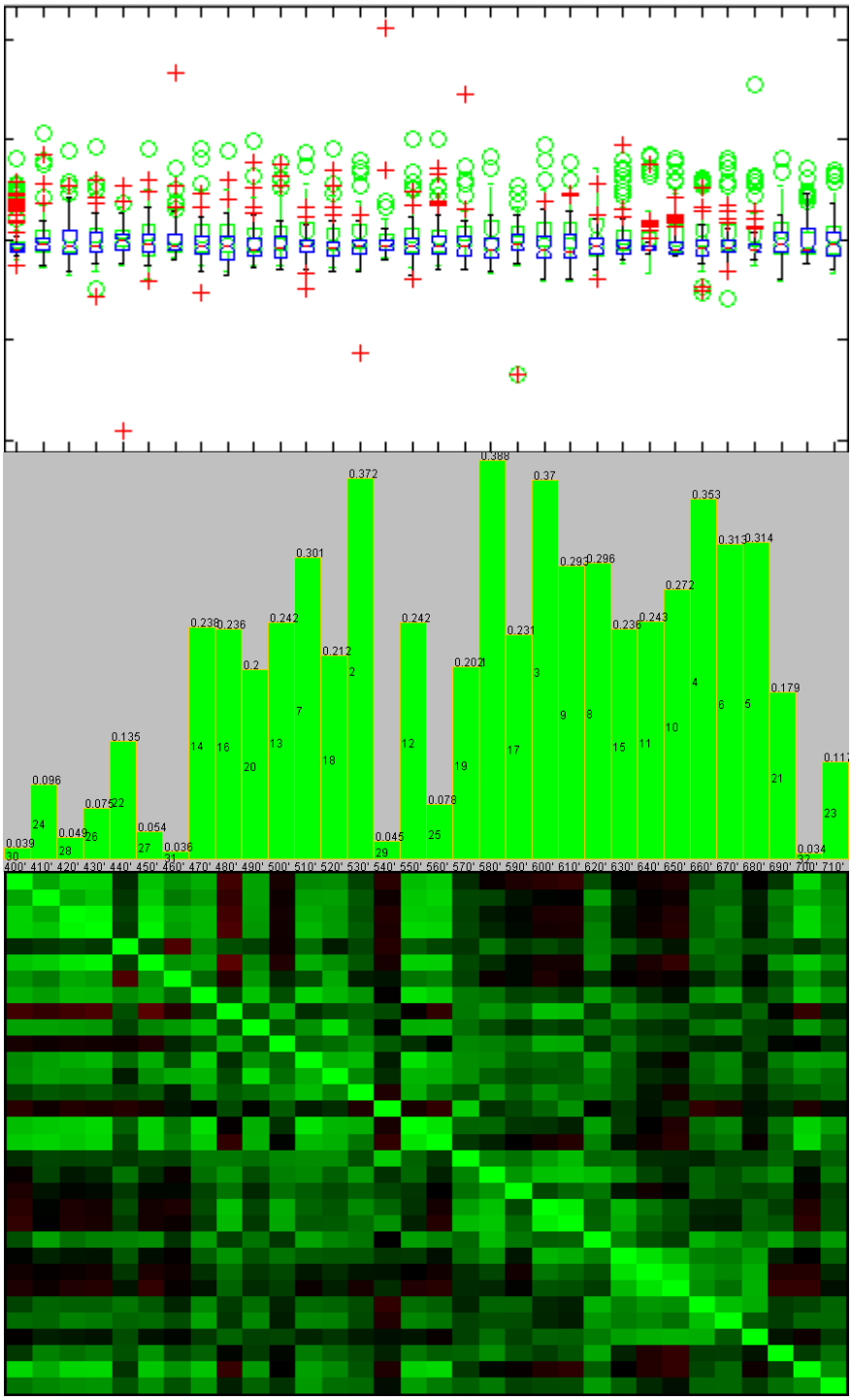
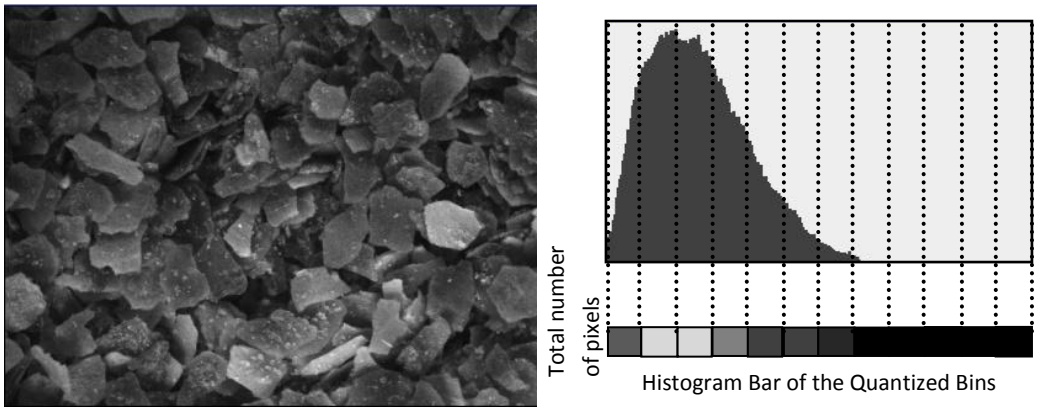


Figure 5.4 Boxplot (top), Fisher discrimination power (middle) and the correlation coefficient map (bottom) of the ABSDIF_CSBE features for the UV excitation.

When we look at the overall fisher scores and correlation coefficient maps in **Figure 5.1** to **Figure 5.4**, the class separation ability of the extracted ABSDIF_CSBE features is higher than the class separation ability of the ISBE features. We inferred from **Figure 5.1** that, 560, 570, 580 and 590 nm spectral bands have higher class separation power for ISBE features and halogen illumination although there are some outliers. For ISBE features under UV excitation shown in **Figure 5.2**, we see that the class separation potential is dispersed to a wider area from 520 nm to 700 nm at the expense of increased redundancy in consecutive spectral bands. When we compare correlation coefficient maps of **Figure 5.1** and **Figure 5.3** we observed that ABSDIF_CSBE features not only have less redundancy through its features but also have higher Fisher scores than the ISBE features. From the **Figure 5.3** we see that 540', 550' and 560' ABSDIF_CSBE features provide relatively higher class separability than the ISBE features. In the case of UV illumination, 520', 590', 600', 660' ABSDIF_CSBE features have a good class separation power as shown in **Figure 5.2** and **Figure 5.4**. They yield relatively less redundancy when we compare the correlation coefficient maps of the ISBE and ABSDIF_CSBE features.

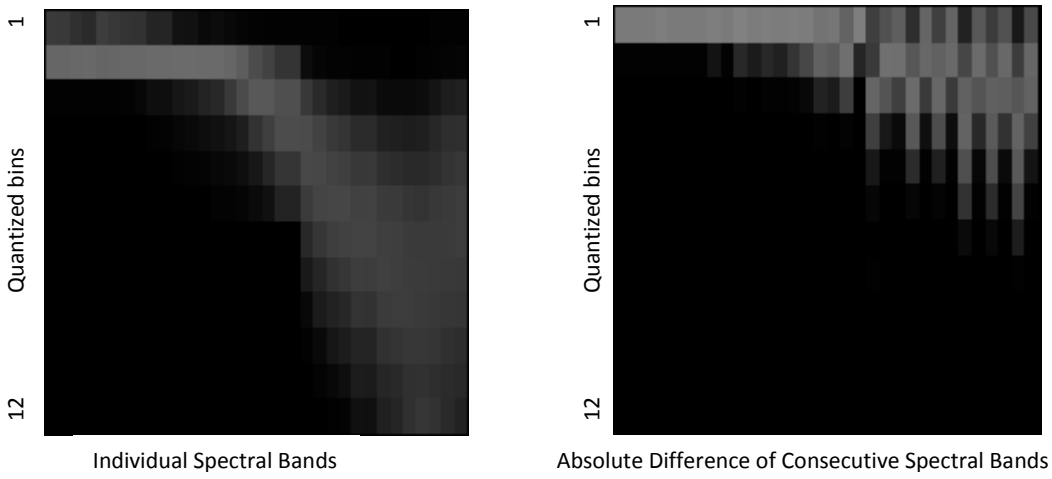
5.4 Quantized Histogram Matrix Features

The feature vectors described in **Equations 5.1 and 5.2** reduce the information in a given band to a single value. However, the frequency of a particular intensity value or the frequency of the difference of the intensity values may provide valuable information. This information can be extracted if the histogram of the intensity values or the difference of the intensity values for a given spectral band is used. **Figure 5.5** and **Figure 5.6** present extracting processes of the quantized histogram matrix features.



a b

Figure 5.5 a) A sample image (640 nm). b) Representative gray level histogram of the image in a) to 12 bins.



a b

Figure 5.6 Quantized histogram matrix (QHM) is composed of histogram bars of the a) ISBE b) ABSDIF_CSBE. X axis denotes the spectral bands (or band pairs in the case of absolute difference) and Y axis denotes histogram bar for that band.

As it is shown in **Figure 5.5**, the histogram of the spectral band image is first computed with predefined number of bins. This limits the feature vector size and also promotes that a reasonable number of pixels fall in each bin. Then, the total

number of pixels within the particular bin is used as the histogram feature. The color of the histogram bar at the bottom depicts the total number of pixels falling in each bin. By using all spectral bands we can construct the quantized histogram matrix (QHM) as depicted in **Figure 5.6**.

QHM feature set is expected to contain the aflatoxin signature. This signature may be degraded if the mean intensities of the spectral bands are used instead. We tried 6, 12, 24 bins feature set and achieved best discrimination power with 12 bins. The QHM features are computed both for individual spectral bands (feature size=33x12) and also for the absolute difference of consecutive spectral bands (feature size=32x12). Hence the QHM features can be expressed as

$$e_{k,b} = \sum_x \sum_y I_{k,b}(x,y) \qquad k = 1,2,\dots\dots,33 \;\; b = 1,2,\dots 12 \qquad \text{(Equation 5.3)}$$

where, *k* denotes index of spectral band and *b* denotes the bin number. As a result, $I_{k,b}(x,y)$ is the pixel gray value of the *k*'th spectral band or absolute difference of consecutive spectral bands of the *b*'th bin.

5.6 Teager Energy Features

Teager Energy Operator (TEO) is first proposed by Kaiser at 1990 (Kaiser, 1990). TEO reveals running estimate of energy as a function of amplitude and current frequency of the signal (Patil, 2008). Teager energy operator has been successfully utilized as a feature extraction technique in various speech processing studies (Teager, 1980, Teager, 1989, Kaiser, 1990, E. Erzin, 1995, Patil, 2008). TEO utilizes three consecutive samples in order to calculate Teager energy. In our problem, TEO can be adapted along with spectral band dimension. To do this, each spectral band and its left and right neighbor's energies were utilized. Teager energy can be computed as

$$\psi[I_k] = |{I_k}^2 - I_{k+1}I_{k-1}| \qquad k = 2,3,\ldots,32 \qquad \text{(Equation 5.4)}$$

Since TEO needs at least 3 consecutive bands, the Teager energy feature vector will be 31 dimensional. **Figure 5.7 and Figure 5.8** show COMPOSITE_VIEW of Teager energy features under the halogen and UV illuminations.

When we compare Teager energy features in **Figure 5.7** with ABSDIF_CSBE features in **Figure 5.3** we see that, class separation power is reduced and the number of outliers is increased. 560', 570', 580' and 590' features have relatively higher Fisher scores under halogen and 620', 590' and 690' features under UV illumination, respectively.

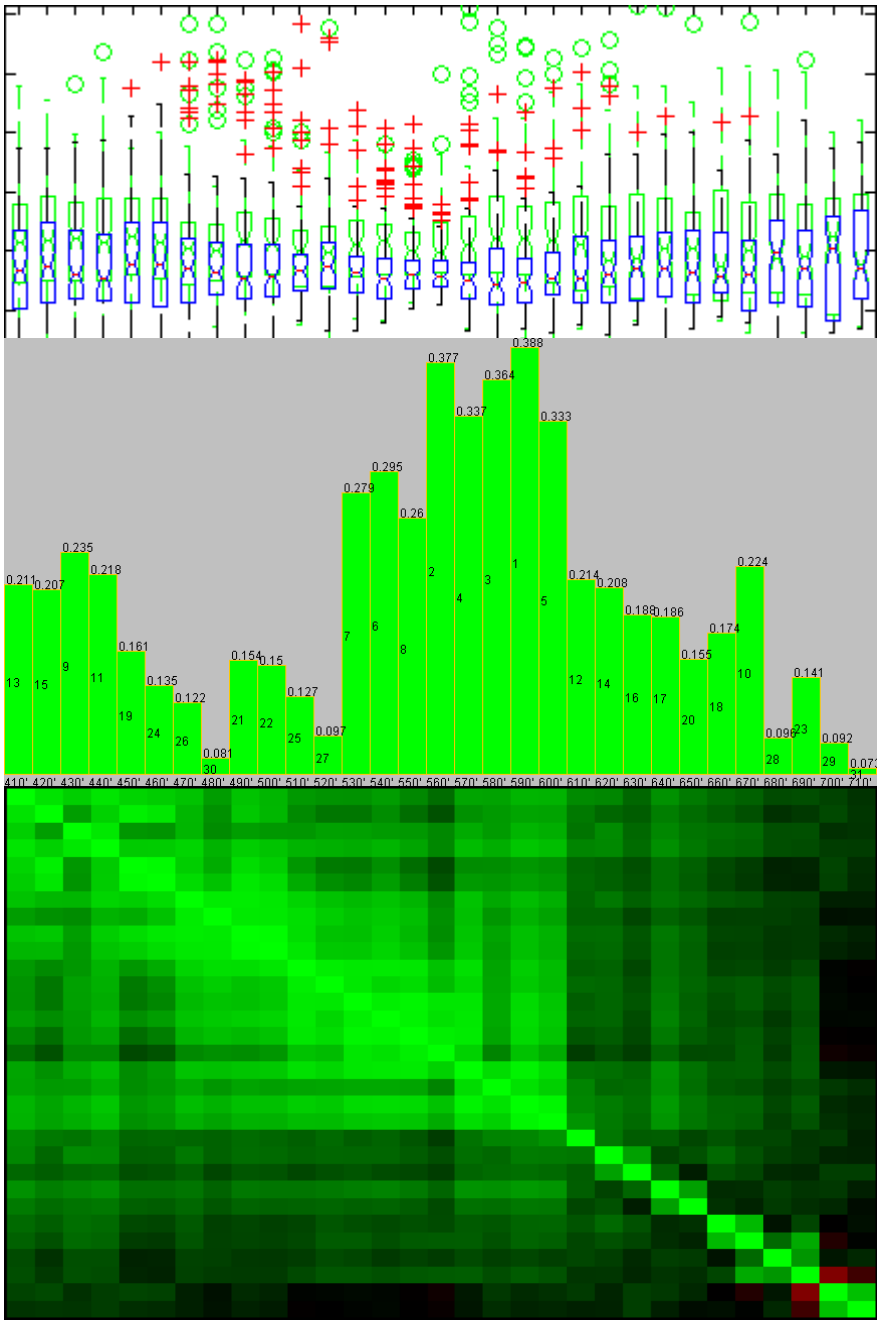


Figure 5.7 Boxplot (top), Fisher discrimination power (middle) and the correlation coefficient map (bottom) of the Teager Energy features for the halogen excitation.

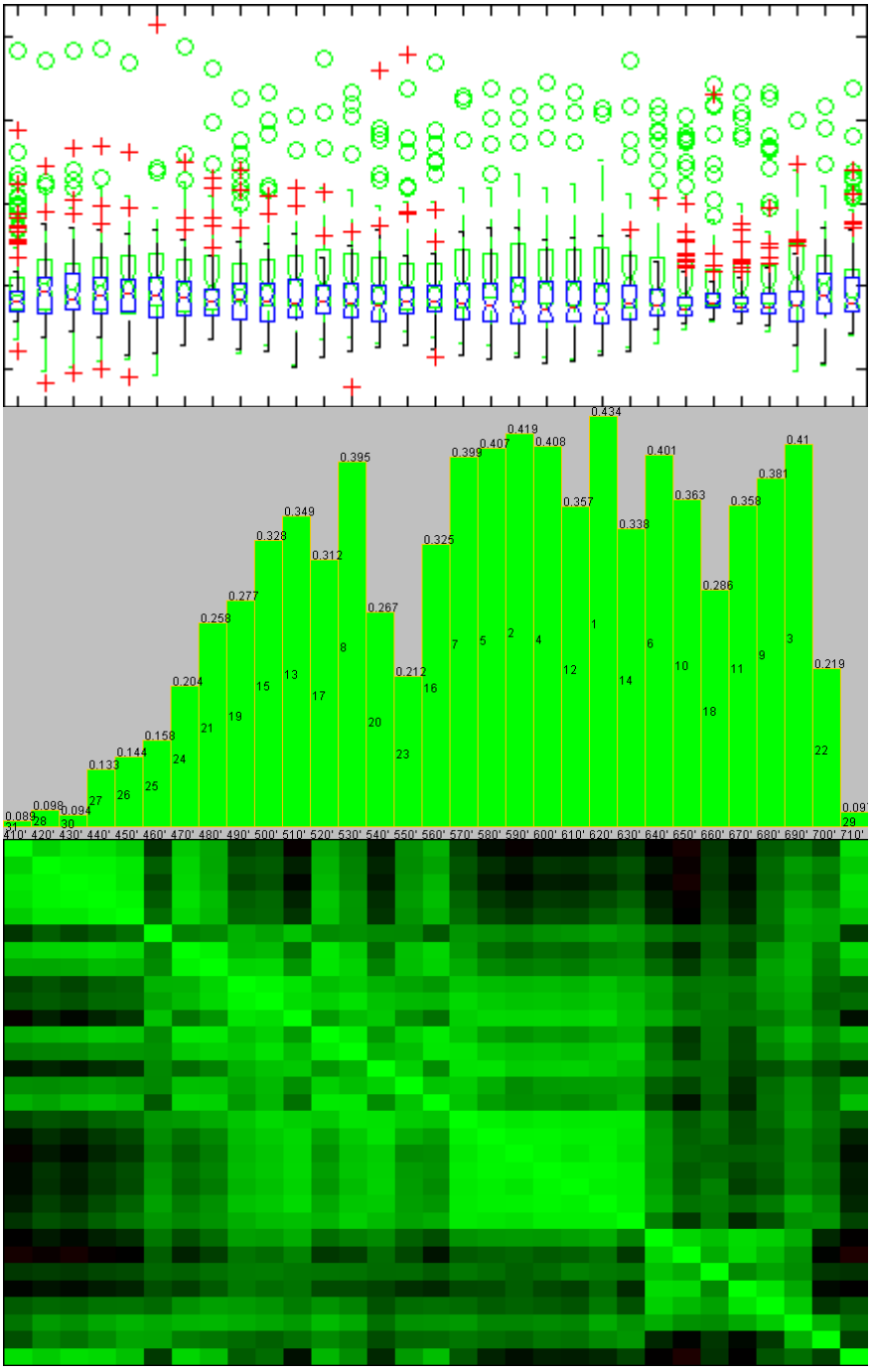


Figure 5.8 Boxplot (top), Fisher discrimination power (middle) and the correlation coefficient map (bottom) of the Teager Energy features for the UV excitation.

5.7 Wavelet LDB Based Features

Original LDB approach tries to get exact location of the discriminative features in a time-frequency domain. However, LDB was extended and used for classification of hyperspectral data just replacing time axis by spectral one dimensional signal (Kumar et al., 2001, Hsu and Tseng, 2002, Kalkan, 2008, Beriat, 2009). Kalkan, (2008) and Beriat, (2009) used LDB for hazelnuts and chili pepper classification problems. **Figure 5.9** shows workflow diagram of adapted LDB based feature extraction and selection approach.

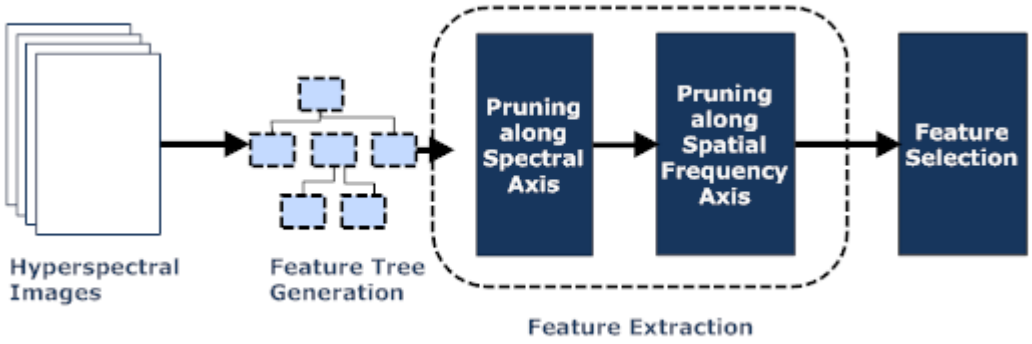


Figure 5.9 LDB based feature extraction and selection

Two different trees are constructed as shown in **Figure 5.10 and Figure 5.11**. The first is a binary tree of spectral images and the other is a quad tree of spatial-frequency (Wavelet features). Note that, one should keep the number of leaf nodes to 2^N where N designates the number of levels. In our case since we have 33 spectral bands (400 to 720 nm and 10 nm width), we ignored the last band in order to satisfy the 2^N constraint.

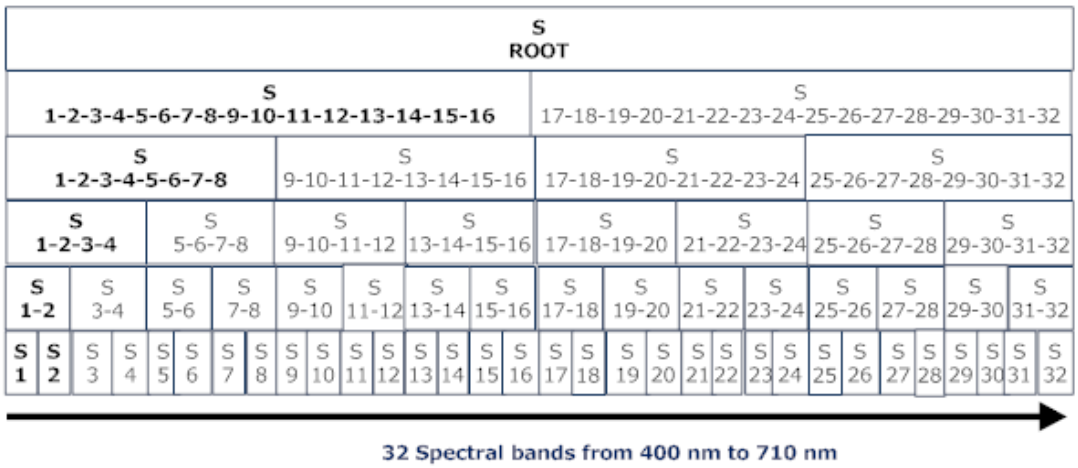


Figure 5.10 Binary tree of spectral images (performs on dyadic bands)

While constructing the spectral tree, we insert energies of every spectral image to the bottom layer of the tree. Then we calculate the average energy of the child nodes and insert them to parent nodes until reaching the root. The pruning algorithm in the spectral tree works as follows.

If $d_{parent} < \max(d_{child1}, d_{child2})$ then

Set $\max(d_{child1}, d_{child2})$ as d_{parent} node

Else remove child nodes

Here, *d* designates the discrimination power of the nodes in **Figure 5.10**. For pruning in both trees, Fisher discrimination metric is used for computing the discrimination power.

In spatial-frequency tree, each node contains four child nodes, called as quad tree. We applied db4 (Daubechies) discrete wavelet transform to each spectral images. In wavelet transform one can apply high pass filtering and low pass filtering along both of the *X* and *Y* axis. Hence, original image decomposes into four sub bands *LL, HL, LH,* and *HH*. First character depicts wavelet transform along *X* axes and the second

along *Y* axes. *HH* gives diagonal details, *LL* corresponds to approximate (low frequencies), *HL* depicts horizontal details and finally *LH* represents vertical details. Applying wavelet transform three times for each sub band, constructs quad spatial-frequency tree with three levels in depth as seen in the **Figure 5.11**.

Figure 5.11 Spatial-frequency quad tree.

Similar to the pruning of the spectral tree, spatial-frequency tree is also pruned according to the same algorithm. The main difference is that, spatial-frequency tree has four children for each node while spectral tree has two.

CHAPTER 6

PROPOSED FEATURE SELECTION METHODS

6.1 Introduction

The main objective of feature selection is to reduce the feature vector size without modifying the feature values. As the features are computed separately for each band in our problem, reducing the feature size removes the need for acquiring the corresponding image along the spectral axis. As a result, a more compact machine vision system can be established. Moreover, the dataset may contain redundant, irrelevant and/or noisy features. Removing these features is expected to yield higher and more robust classification rates. Feature selection can be divided into two main categories. First is feature ranking. The second is feature subset selection.

6.1 Feature Ranking

Feature ranking is the process of sorting the candidate features according to their predictive significance. Although this approach is computationally efficient, determining the number of features satisfying maximum classification accuracy is still an open problem. Determining the best subset after ranking the features proceeds as follows:

F=(f_1,f_2, … f_N) //F is the ranked feature set; where, *N* is the //total number of features

F_S, F_D=Ø //F_S is the selected feature set, F_D is the desired //Feature set

Max, Accuracy=0 //Max and Accuracy are the maximum and the //current accuracy rates, respectively.

For each feature f_i in the ranked feature set do

  F_S=F_S U f_i //concatenate current feature to F_S

  Accuracy ← Evaluate(Fs) //compute Fs accuracy rate and assign it to //current accuracy

  If (Max<Accuracy) then

    Max ←Accuracy, F_D ←F_S //return F_S

End-for

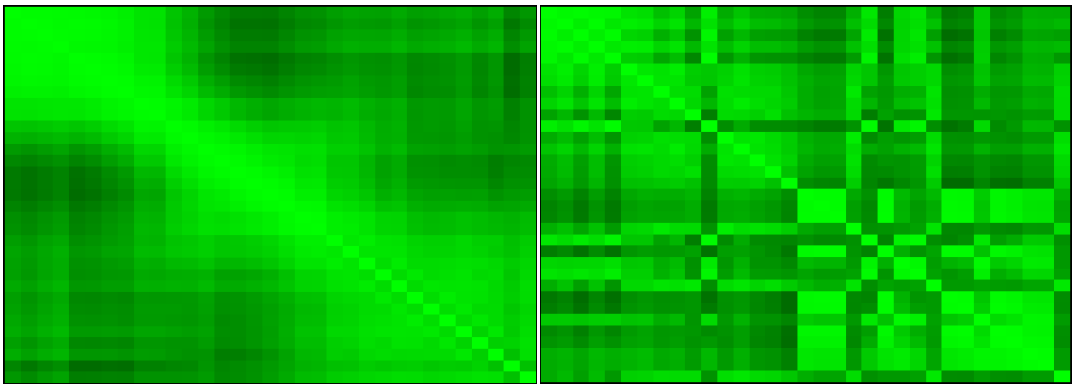
Feature selection scheme above is called as Incremental Feature Selection (IFS) (Niu et al., 2010). Selecting the best *N* features is intuitive but there is a possibility of retaining highly correlated features as in the hyperspectral imaging domain. In this thesis we used various feature ranking schemes. These are Fisher, mRMR and MLP based ranking methods.

6.1.1 Fisher Based Ranking

As the first ranking scheme we utilized Fisher discrimination power. Fisher discriminant was first proposed by Fisher, (1936) and is utilized in various studies (Beriat, 2009, Ataş et al., 2011a, Kalkan et al., 2011). Fisher discrimination projects data from n-dimensional space to a one-dimensional space where between class-scatter is maximum and within class-scatter is minimum. It can be computed as;

$$F_{dp} = \frac{|\mu_1 - \mu_2|^2}{\sigma_1^2 + \sigma_2^2} \qquad \text{(Equation 6.1)}$$

The effectiveness of the method can be compared in **Figure 6.1** in terms of correlation coefficient matrix of the individual spectral bands with the original and Fisher based ranking.



a b

Figure 6.1 Ranking the halogen individual spectral band energy features a) without b) with based on Fisher discrimination power.

As it is seen in **Figure 6.1 a**, there are high redundancies amongst features. Lighter colors indicate higher correlation between the features. As the color becomes darker lower correlation is observed. Likewise in **Figure 6.1 b** the ranking based on Fisher discrimination power yields lower correlation (less redundancies) amongst the features than the original ranked features.

6.1.2 mRMR Based Ranking

An Information theoretic approach, mRMR stands for minimum redundancy and maximum relevance and first proposed by Peng, (2005). Fisher based or Pearson correlation coefficient based feature ranking have the weakness of analyzing at most two features at the same time. However, mRMR assesses the correlation of the given feature with other features while preserving high correlation between current feature and the class labels. It penalizes the inter-feature correlation whilst

favoring the class label dependencies. mRMR approach is a filtering method and it is reported as computationally efficient (Peng et al., 2005b, Peng et al., 2005a).

6.1.3 MLP Based Ranking

Features can be ranked based on MLP connection weight as well. By MLP, we mean the special case of one hidden-layer, feed-forward neural network trained by the well known back propagation algorithm (Rumelhart et al., 1986). MLP has a large number of free parameters which may render the interpretation of its internal dynamics during the training and the validation stages difficult. To address these challenges, and provide adequate robustness and reliability for the classification process, dozens of trials have to be performed. These free parameters are learning rate, momentum, number of epochs, the number of nodes in the hidden layer and initialization scheme of the connection weights. Exhaustive search for the optimal parameter set requires an exponentially increasing number of trials which makes search unfeasible. We proposed some heuristics for tuning the free parameters of the MLP to overcome this challenge.

6.1.3.1 Adaptively adjusted learning parameter, momentum and number of epochs

The learning rate, momentum coefficient and the number of epochs were adjusted adaptively with the rate of convergence. The learning rate and the momentum value were initialized as best practices to 0.1 (Cravener and Roush, 1999, Rajanayaka et al., 2003), and decayed during the learning phase. Decaying procedure is summarized at **Figure 6.2**. By this way it is expected to prevent the classifier from over-fitting and under-fitting.

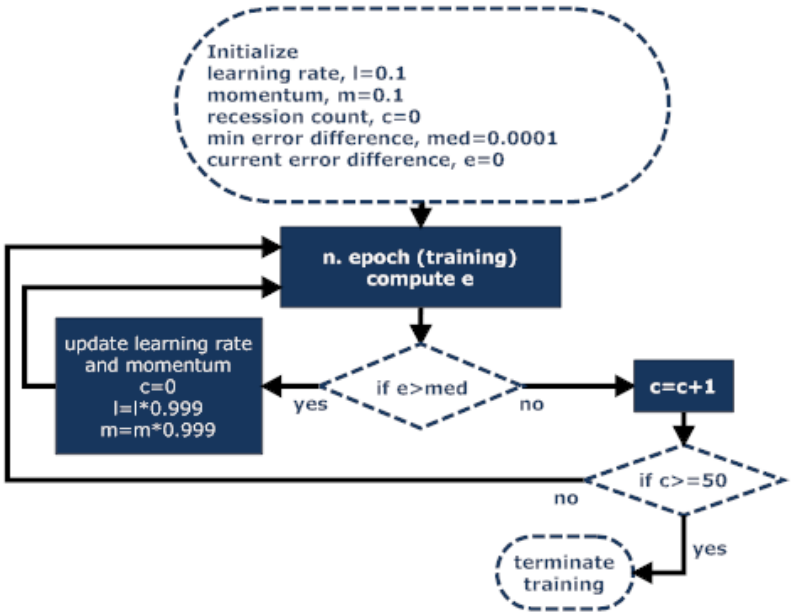


Figure 6.2 Schematic diagram of the decaying procedure of the learning rate and momentum values.

6.1.3.2 Adjusting number of neurons in the hidden layer

As Hornik et al. (1989) and Malek et al. (2000) pointed out MLP with single hidden layer is adequate as a universal approximator. For the sake of simplicity, we employed a single hidden layered network topology in our study. Determining the optimal number of neurons in the hidden layer is still an open problem in the ANN domain. Suggested data size should exceed ten times the total connections in the neural network to achieve 90% classification accuracy (Baum and Haussler, 1989). There are some rules of thumb which are extensively used by the researchers (Blum, 1992, Berry and Linoff, 1997, Boger and Guterman, 1997, Rapid Miner, 2009). Specifically, the Rapid Miner team (2009) suggested number of neurons in the hidden layer to be:

$$N_{nod} = \frac{(N_{features}+N_{classes})}{2} + 1 \qquad \text{(Equation 6.2)}$$

In our trials, this approach gave satisfactory results so we decided to utilize it. Empirical results are presented in **Table 6.1.** Here we used MLP as the classifier to assess the effectiveness of the **Equation 6.2** against three different datasets. N_{nod} in **Table 6.1** corresponds to the number of nodes given by **Equation 6.2.**

Table 6.1 Benchmark tests for best practice used in this study.

Dataset	Feature Size	N_{nod}	Generalization error % based on LOO-CV				
			$N_{nod}-2$	$N_{nod}-1$	N_{nod}	$N_{nod}+1$	$N_{nod}+2$
AbsDif Halogen	32	18	43%	41%	32%	36%	38%
Teager Energy	31	17	41%	39%	39%	40%	40%
Heart Disease	10	7	27%	26%	25%	27%	29%

6.1.3.3 A new weight initialization scheme

A common approach to initialize the artificial neural network connection weights is generating random numbers from a uniform distribution within a small interval centered around zero. Random connection weights lead the network to initialize at different points in the weight space. If this initialization is close to the global minimum the network often converges to the global solution. Unfortunately, it may get trapped in a local minimum when the initial point is far away. Hence multiple initial points should be tried in order to get consistent results. The number of trials depends on the confidence level and the computational cost constraints. Especially for combinatorial problems, utilizing random initialization is not preferred. There are various studies (Nguyen and Widrow, 1990, Shimodaira, 1994, Yam and Chow, 1997) that investigate efficient weight initialization techniques.

According to Haykin (1999), usage of both small and large values for initializing connection weights should be avoided for two reasons. First if you choose large initial values, it may lead to saturation of the neurons in the network, causing the training process to slow down. Conversely, small values results in the back propagation algorithm to operate on a relatively flat area around the origin of the error surface. This is particularly true for the activation function using the hyperbolic tangent function (Haykin, 1999).

In general, the connection weights between the input and the hidden layer are assigned randomly at startup. Studies (Keeni et al., 1999, Russo, 1991), indicate that random numbers are often generated at the interval of [-0.5, 0.5]. Hence, we used that interval for our random initialization.

Another alternative is using the Nguyen-Widrow weight initialization method (Nguyen and Widrow, 1990). With Nguyen-Widrow method, initial connection weights of the hidden nodes are distributed so that outputs of the hidden neurons are always in their active regions which make learning process more efficient in terms of convergence speed. Nguyen-Widrow method works as follows:

- Generate random numbers for connection weights from input node *i* to *p* hidden nodes w_{ij} in a specific range.
- Calculate the β value for *n* inputs and *p* hidden nodes as

$$\beta = 0.7\sqrt[n]{p} \qquad \text{(Equation 6.3)}$$

- Recalculate connection weight as

$$\frac{\beta w_j}{\|w_j\|} \rightarrow w_j \quad j = 1 \dots p \qquad \text{(Equation 6.4)}$$

where, $\|w_j\|$ is the Euclidean norm of the weights and can be calculated as

$$\|w_i\| = \sqrt{\sum_{i=0}^{p} w_i^2} \qquad \text{(Equation 6.5)}$$

Apart from these we utilized several weight initialization strategies. With *p* hidden nodes

- **Zero Weight:**

$$w_j = 0 \quad j = 1 \dots p \qquad \text{(Equation 6.6)}$$

- **Constant Value:**

$$w_j = 1000 \quad j = 1 \dots p \qquad \text{(Equation 6.7)}$$

- **Random**:

$$w_j \sim U[-0.5, 0.5] \qquad \text{(Equation 6.8)}$$

- **Symmetric Random**:

$$w_j = \begin{cases} U[-0.5, 0.5] & if\ j \leq p/2 \\ -w_{j-p/2} & if\ j > p/2 \end{cases} \qquad \text{(Equation 6.9)}$$

- **Linearly increasing weights**:

$$w_j = -0.5 + \frac{j}{p} \qquad j = 1 \dots p \qquad \text{(Equation 6.10)}$$

- **Linearly decreasing weights:**

$$w_j = +0.5 - \frac{j}{p} \quad j = 1 \dots p \qquad \text{(Equation 6.11)}$$

- **V-Shape**:

  As **Figure 6.3** indicates, weight function is similar to the V shape that starts from +0.5 to -0.5 and returns to +0.5 again.

$$w_j = \begin{cases} +0.5 - \frac{2j}{p} & if\ j \leq p/2 \\ -0.5 + \frac{2j}{p} & if\ j > p/2 \end{cases} \qquad \text{(Equation 6.12)}$$

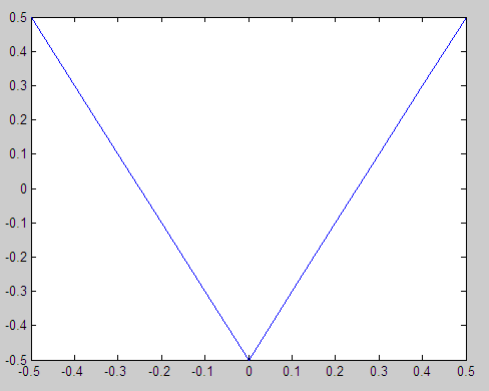


Figure 6.3 V-Shape weight initialization scheme.

- **Inverse V-Shape**:

$$w_j = \begin{cases} -0.5 + \frac{2j}{p} & if\ j \leq p/2 \\ +0.5 - \frac{2j}{p} & if\ j > p/2 \end{cases} \qquad \text{(Equation 6.13)}$$

- **Gaussian Shape:**

  Given, *μ=0, σ=0.3* and constant *0.83*, Gaussian function can be plotted as shown in **Figure 6.4**.

$$w = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{(j-\mu)^2}{2\sigma^2}} - 0.83 \quad j = 1 \ldots p \qquad \text{(Equation 6.14)}$$

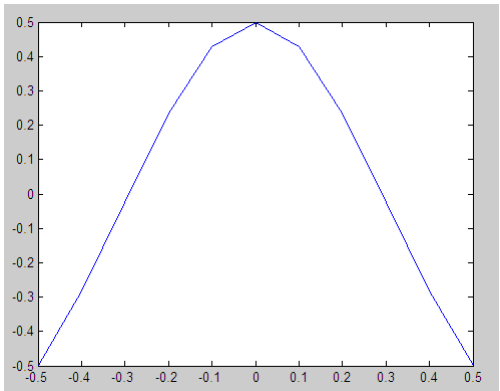


Figure 6.4 Gaussian-Shape weight initialization scheme.

- **Sigmoid Logistic Shape:**

$$w_j = \frac{1}{1+e^{-jp}} \qquad j = 1 \ldots p \qquad \text{(Equation 6.15)}$$

- **Hyperbolic Tangent:**

  Given β=*0.5* function can be drawn as

$$w_j = \beta tanh\,(\frac{p}{3}j) \quad j = 1 \ldots p \qquad \text{(Equation 6.16)}$$

- **Rectangular pulse approximation: Figure 6.5** shows the shape of the function.

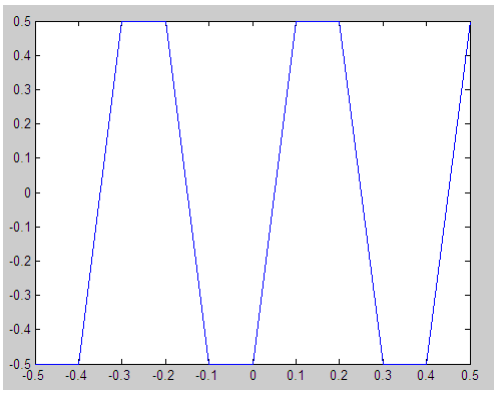


Figure 6.5 Rectangular pulse approximation weight initialization scheme.

- **Alternate: Figure 6.6** depicts Alternate shape.

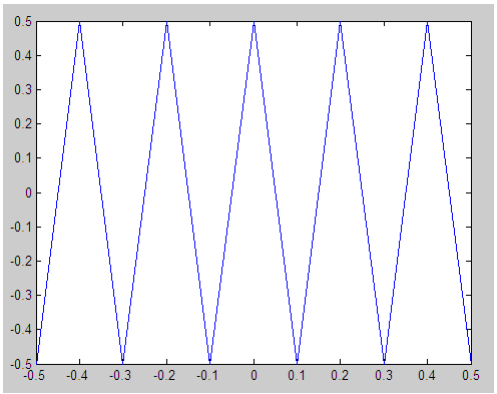


Figure 6.6 Alternate-Shape weight initialization scheme.

V-Shape produced best generalization performance among all weight initialization methods as shown in **Table 8.10**. Throughout this study, we utilized V-Shape weight initialization strategy for MLP.

Garson, (1991) stated that one can define feature saliency metric as:

$$\tau_i = \sum_{j=1}^{N} \left| w_{ji} w_j \right| \qquad i = 1,2,..M \qquad \text{(Equation 6.17)}$$

Here, $\boldsymbol{\tau_i}$ denotes saliency metric of the *i*'th feature and *M* and *N* designate the number of input and hidden layer nodes, respectively (Garson, 1991). In addition to this, Olden and Jackson, (2002) used Garson's algorithm and sensitivity analysis by applying randomization approach. They showed that MLP based variable selection can be applied and interpreted successively in the ecological domain (Olden and Jackson, 2002). Each feature/variable is represented as an input node in the neural network topology. According to **Equation 6.17**, the *i*'th feature's saliency metric can be computed as the sum of the absolute product of the connection weights from the *i*'th input node through the hidden nodes to the output node as indicated in **Figure 6.7**.

The rationale behind MLP based feature saliency metric is as follows. Connection weights are continuously updated in the training phase so that significant input nodes have strong connections in the network topology which means those input nodes have higher contribution to the output. Likewise, connection weights of the irrelevant features tend to vanish. Hence, MLP based feature saliency metric can be used as a dimensionality reduction technique by decaying the connection weights of the insignificant features. Moreover, the saliency can also be instrumental in ranking the features based on their discrimination power.

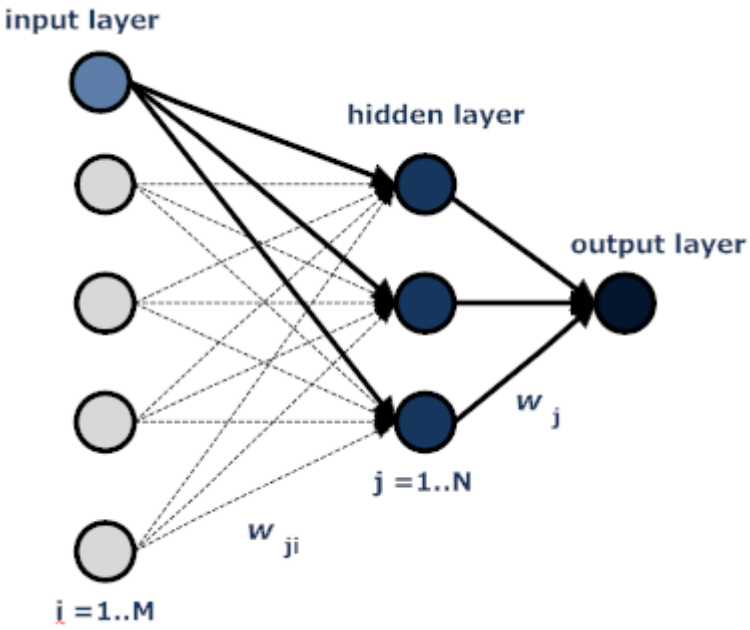


Figure 6.7 MLP with input, hidden and output layer. w_{ji} is the connection weight between *i*'th input node and *j*'th hidden node. Similarly, w_j is the connection weight between *j*'th hidden node and the output node.

Figure 6.8 presents the correlation coefficient maps of the original, Fisher, mRMR and MLP based ranking for individual spectral band energy features under the halogen illumination. As it is seen in **Figure 6.8**, MLP and mRMR rank the features in such a way that correlations amongst features are small.

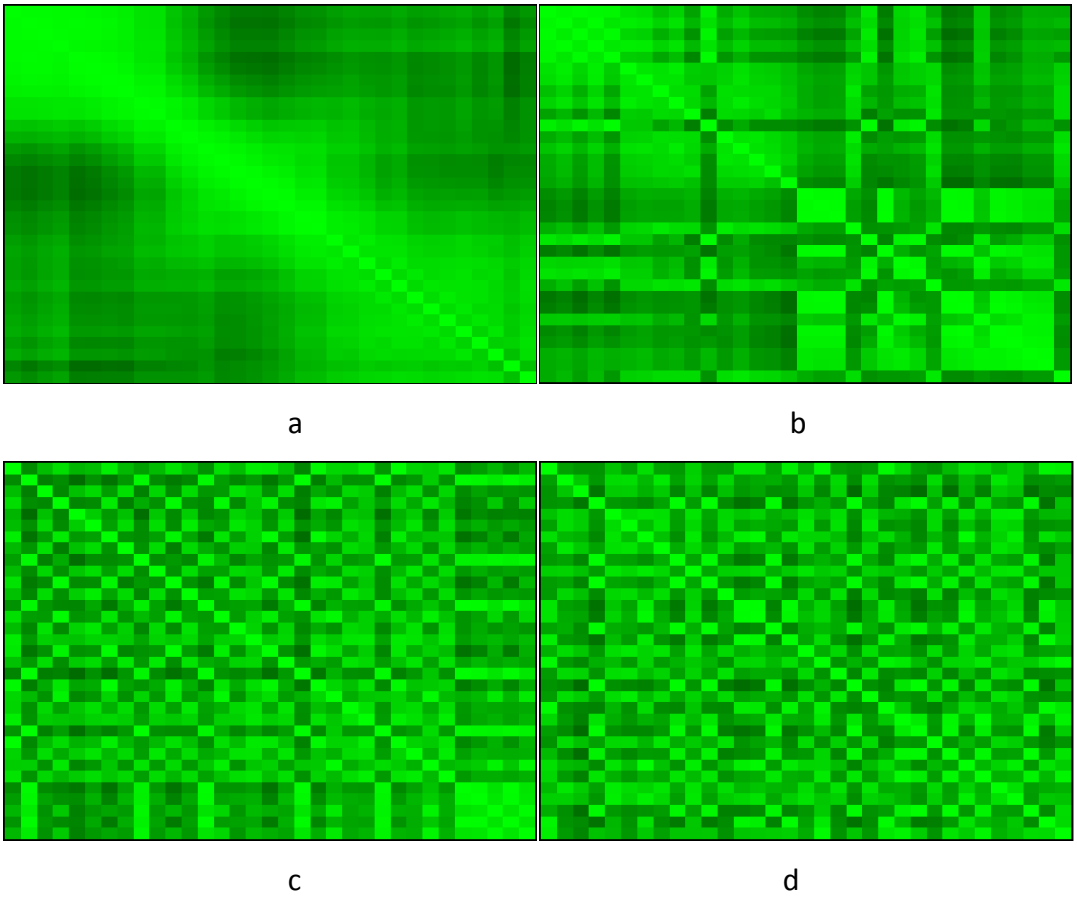


Figure 6.8 Correlation coefficient maps of the a) original, b) Fisher, c) mRMR and d) MLP based ranking for the individual spectral band energy features under the halogen illumination.

Throughout the thesis we used MLP as a specialized learning model for feature ranking (Section 6.1.3), feature subset selection (Section 6.2) and evaluation purposes (Chapter 8). In all these applications of the MLP we use the same approach to connection weights initialization, selection of number of hidden nodes, learning rate, momentum constant and number of epochs as described in the present section.

6.2 Feature Subset Selection

Feature subset selection can be described formally as follows: Given a set of *Y* features, if X_d is all possible subsets of size *d* then, the optimal feature subset is the one that provides the best generalization performance based on a criterion function *J(X)*. Explicit formulation of the desired feature subset $\check{X}_d$ can be written as (Somol et al., 2010).

$$J(\check{X}_d) = max_{X \in X_d} J(X) \qquad \text{(Equation 6.18)}$$

The simplest approach to find efficient feature subset is the exhaustive search method. For large number of features this method is intractable due to high computational cost. Stochastic SA method can be used as an alternative and give some preliminary intuition for the best feature subset. Due to time constraints, in this thesis, we employed exhaustive and SA searches when they were feasible. More specifically, we investigated ISBE and ABSDIF_CSBE features up to five features with exhaustive search strategy. This yields C(33,5)=237,336 and C(32,5)=201,376 different feature subsets for ISBE and ABSDIF_CSBE features, respectively. Beyond this limit, exhaustive search was unfeasible with our computational resources. Therefore, we utilized SA search for feature sizes 6 to 10. No search was attempted in higher dimensional feature subsets. Since quantized ISBE and quantized ABSDIF_CSBE features dimension is one order of magnitude higher than those of ISBE and ABSDIF_CSBE features we could not employ exhaustive and SA search, instead we used proposed the HBBE method described in Section 6.1.1.

Figures 6.9 and 6.10 demonstrate a sample of exhaustive and SA search graphical outputs.

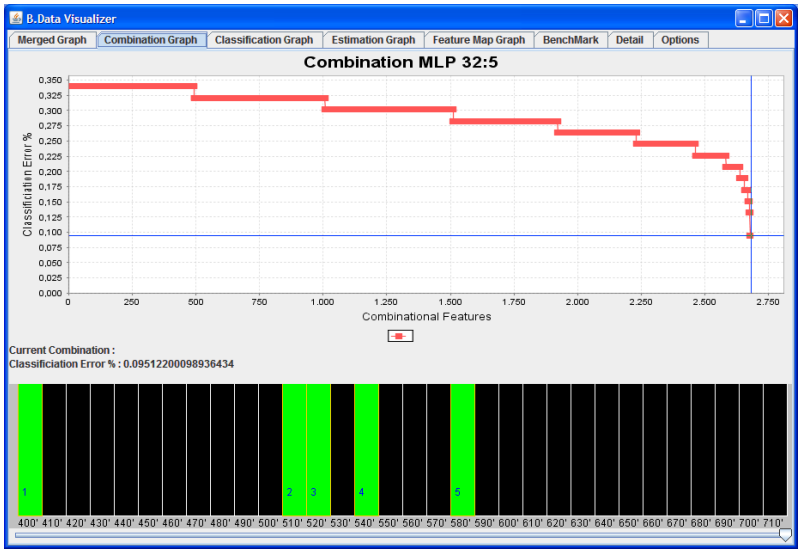


Figure 6.9 A typical illustration of the exhaustive search for the halogen ABSDIF_CSBE features for the subset size of 5 in the 32 features.

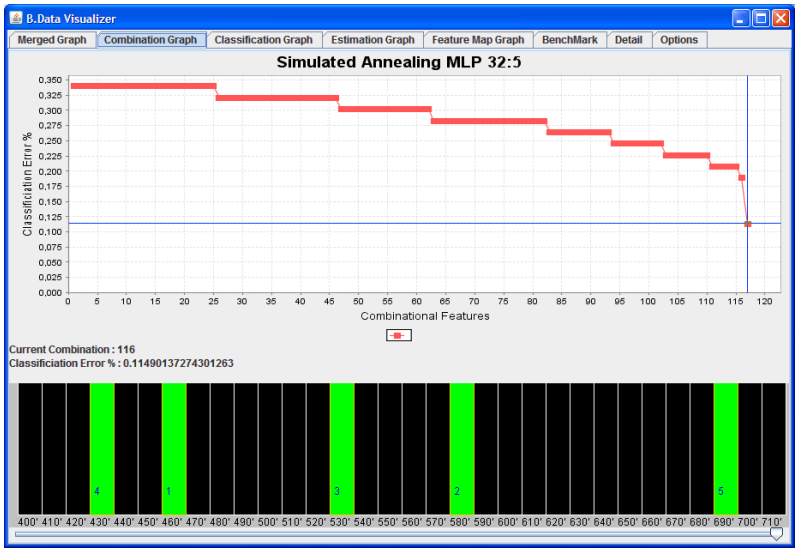


Figure 6.10 A typical illustration of the simulated annealing search for the halogen ABSDIF_CSBE features for the subset size of 5 in the 32 features.

6.1.1 Proposed HBBE Method

We propose a novel feature subset selection technique based on the MLP feature saliency metric. Forward selection and backward elimination were extensively used subset selection methods and regarded as robust to over fitting (Guyon and Elisseeff, 2003). Again, Guyon et al, (2002) proposed a state of the art support vector machine based recursive feature elimination method (RFE). In the RFE approach, the search starts with complete feature set. At every iteration, the features are ranked according to their predictive significance and the least significant one is removed from the feature set. This procedure proceeds until some convergence criterion is met. Order of complexity of the RFE method can be expressed as **O(n)**. One can remove a single or last *N* features. Eliminating one feature per iteration is computationally costly, especially if we have large number of features. Thus, we modified the RFE method of Guyon et al, (2002) by replacing the central classifier SVM by MLP and we accelerated the process by removing last *N* features at each iteration where *N* is the number of hidden neurons in the MLP network. The procedure consists of two main stages: Backward elimination and subset verification. At the backward elimination stage, the candidate feature set is ranked based on the MLP saliency metric and the last N_{nod} as in **Equation 6.2** are eliminated. This process is repeated until only one feature remains. The number of steps, *M*, required for *P* original features is given by

$$M \cong log_2 P \qquad \text{(Equation 6.19)}$$

Order of complexity of the proposed HBBE method can be described as **O(logn)** which is superior to RFE in terms of computational cost. The candidate feature subset at every step is recorded in an array list data structure as shown in **Figure 6.11**. After the feature elimination process is completed, subset verification based on IFS search strategy is initialized. At this stage, starting from the lowest feature

subset (which typically consists of a single feature) generalization errors for each feature subset are computed. If there are *L* ranked features at a given subset, the generalization error for i) only one, ii) first two, iii) first three, etc. features are computed using LOO-CV validation strategy. The computational cost associated with this step increases exponentially with increasing the number of ranked features *L*. Therefore, we only computed generalization errors for the first five levels from the bottom of the triangle in **Figure 6.11**. We chose the feature subset with the minimum generalization error. In the case of similar generalization errors we preferred the smaller set of features amongst the candidate feature subsets.

Proposed HBBE algorithm can also be summarized as in the following simplified programming code:

```
OFS = [f1, f2 , … fn]                                   //original feature set
S = []                                                  //ranked feature list
While (OFS .length>1) do
        Rank OFS by MLP
        NH=number of hidden nodes in the MLP
        Eliminate last (OFS.length-NF) features from OFS
        S [i++] = OFS
EndWHILE
DS = [], HBBE_F= []                                     //desired subset for HBBE
For i=S.length-1 to 1 do                                //bottom-up approach
        RF=S[i]                                         //current ranked feature set
        CS = []                                         //current subset
        Add CS to list DS                               //expand DS list
        For k=0 to RF.length-1 do                       //search for the lowest error
                CS = concatenate (CS, RF[k])            //concatenate subsequent feature
                If LOO-CV (CS) < error then             //error evaluation by LOO-CV
                        Error = LOO_CV (CS)
                        DS[i] = CS                      // i'th desired subset
                EndIF
```

```
        EndFOR
EndFOR

For k=0 to DS.length-1 do
        If LOO-CV (DS[k]) < error then          //error evaluation by LOO-CV
                Update error
                HBEE_F = DS[k]                  //get the HBBE feature subset
        EndIF
EndFOR
```

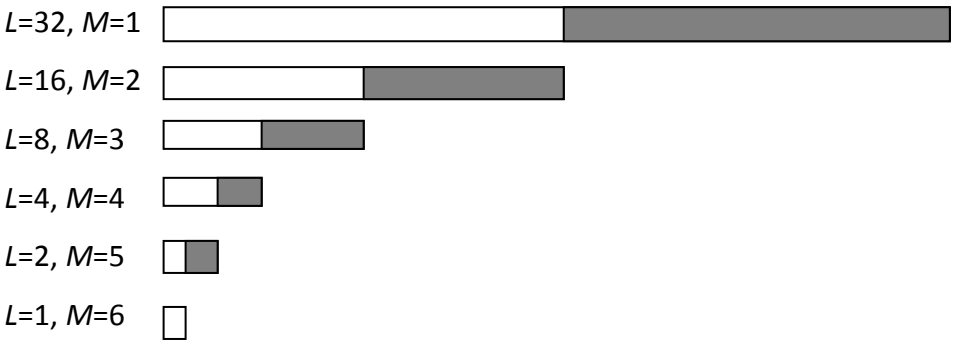


Figure 6.11 Proposed HBBE method. *L* and *M* designate the number of features and the number of steps, respectively.

6.1.2 Complementary Subset Search by Dependency Graph (CSS_DG) Method

We propose another feature subset selection approach which we named as complementary subset search by dependency graph (CSS_DG). The rationale behind this method is that, the contribution of a feature can be evaluated by observing the effect of its introduction to or removal from the set. We analyzed the relationships between features when a single feature is removed. The algorithm used in this method can be summarized as

F= (f_1, f_2, .. f_n) // *F* is a set of *N* original features

F' = rank (F) //F' ranked feature set based on MLP

For each feature *fi* in the set of F

- Remove *i'*th feature from the F and rank the set as F''
- Compute the dislocation of each feature with respect to original index as

$$d_x^i = \nabla(F'_x, F''_x) \quad x = 1,2,..n-1 \quad where\ x \neq i$$

- If current location of the *x*'th feature is moved to the *right* then this indicates that removing *i*'th feature *adversely* affects the *x*'th feature so *i*'th feature and *x*'th features can be regarded as coherent features with the degree of the d_x^i.
- Else If current location of the *x*'th feature is moved to the *left* then this indicates that removing *i*'th feature *favorably* affects the *x*'th feature so *i*'th feature and *x*'th features can be regarded as counterproductive features with the degree of the d_x^i.

End.

This CSS_DG process for the ISBE features under the halogen illumination is demonstrated below:

Ranked features

12;26;31;1;22;25;18;28;0;21;2;32;16;17;3;15;23;11;29;4;27;13;9;5;30;7;19;8;24;20;6;14;10;

Ranked features after removing i'th feature

0 1;12;31;22;21;26;2;32;16;18;28;3;15;11;9;25;8;13;4;23;20;29;17;27;19;7;24;5;30;14;10;6;

1 12;26;22;0;31;28;18;27;32;16;21;25;13;2;15;11;3;23;17;9;29;7;4;19;5;20;10;24;8;14;30;6;

2 12;1;22;31;21;26;0;28;18;32;16;15;3;11;4;13;17;23;9;25;27;29;5;8;20;19;7;10;30;14;24;6;

.

.

Dependency Scores

0 25:10;17:9;26:4;5:4;30:4;29:3;28:3;27:3;18:3;23:3;12:1;6:1;7:0;31:0;22:-1;4:-1;24:-2;14:-2;10:-2;19:-2;15:-3;1:-3;3:-3;32:-4;13:-4;16:-4;2:-4;11:-4;21:-5;9:-8;20:-9;8:-11;

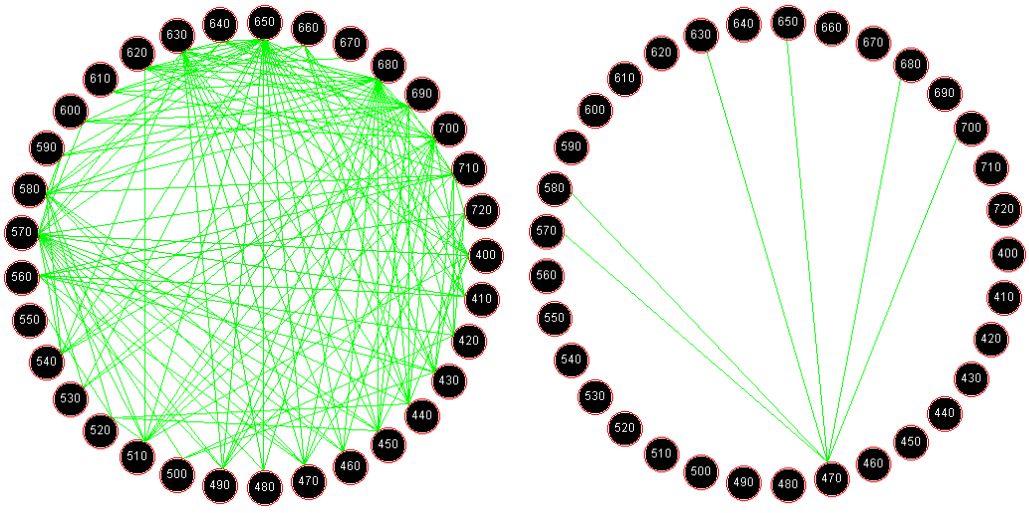
1 25:6;30:6;17:5;4:3;2:3;29:2;3:2;31:2;21:1;5:1;8:1;23:1;6:1;26:0;12:0;18:0;24:-1;15:-1;22:-2;14:-2;11:-2;28:-2;16:-3;19:-3;9:-3;32:-3;20:-4;7:-4;0:-5;10:-6;13:-9;27:-13;

2 25:14;26:4;30:4;29:3;17:3;18:2;24:2;7:1;31:1;23:1;6:1;27:0;12:0;28:0;5:-1;19:-1;3:-2;0:-2;1:-2;22:-2;14:-2;32:-2;16:-2;15:-4;8:-4;11:-4;9:-4;20:-5;21:-5;4:-5;10:-5;13:-6;

.

Note that, at the given example, the proposed method only analyzes the dyadic feature dependencies but it can be extended to examine ternary and higher degree relationship as well. Furthermore, the algorithm above can be extended by evaluating the ranking weights instead of feature index. We implemented this approach as well but we did not see any significant improvement.

Figure 6.12, 6.13, and 6.14 demonstrate the subset selection dependency graphs for coherent (productive), neutral, and counterproductive feature subsets, respectively.



a b

Figure 6.12 Coherent feature subsets amongst ISBE features a) considering all features and b) considering a sample feature (470 nm).

Figure 6.12 b indicates that when we remove 470 nm feature from the original feature set, 570, 580, 630, 680 and 700 nm features are affected adversely. This implies 570, 580, 630, 680 and 700 nm features are complementary to 470 nm feature (highly coherent subset).

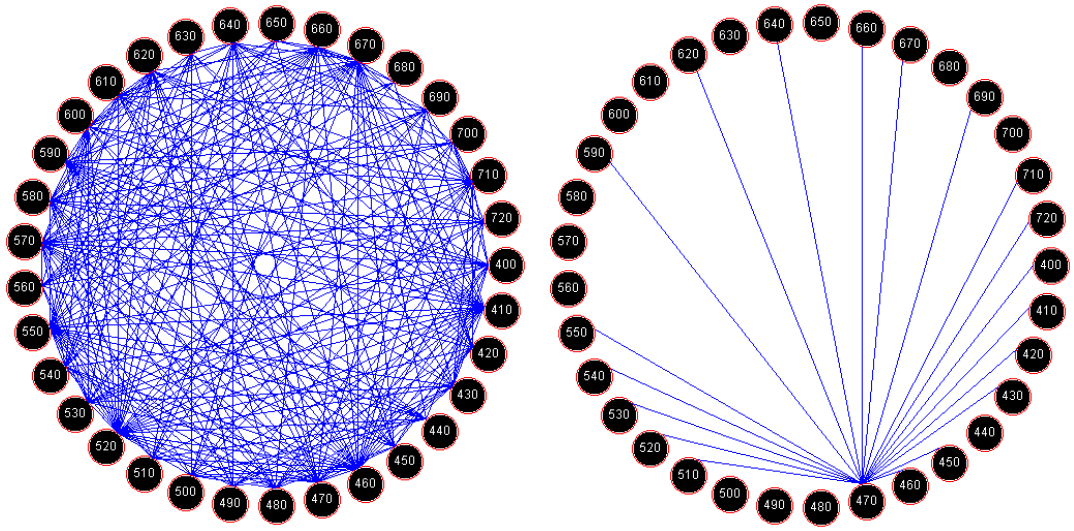
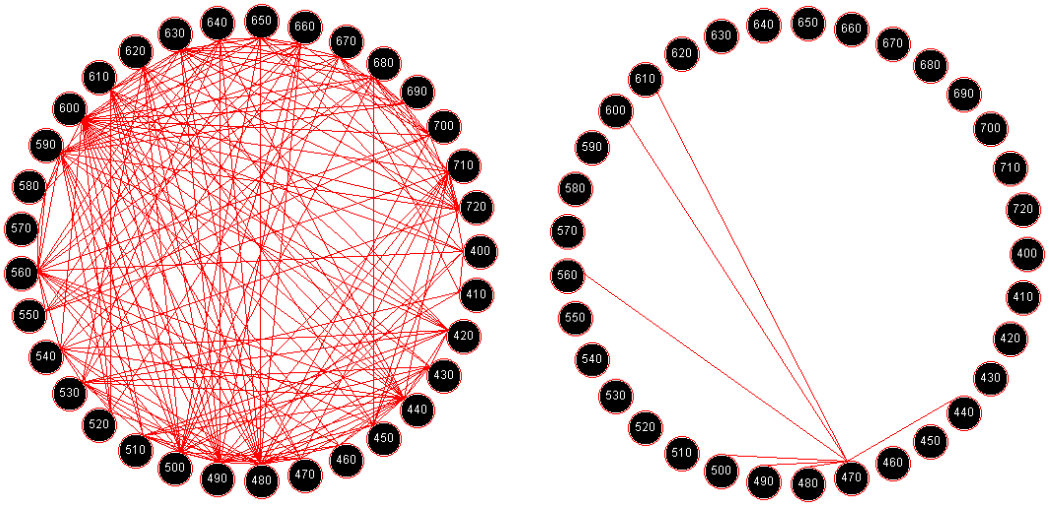


Figure 6.13 Neutral feature subsets amongst ISBE features a) considering all features and b) considering a sample feature (470 nm).



a b

Figure 6.14 Counterproductive feature subsets amongst ISBE features a) considering all features and b) considering a sample feature (470 nm).

In our trials we did not achieve reliable and consistent results by applying complementary subset search. Therefore we did not include detailed analysis and results in Chapter 8: Results and Discussion.

CHAPTER 7

MULTIVARIATE FEATURE EXTRACTION AND SELECTION TOOL

In this thesis, we developed multivariate feature extraction and selection (MFES) tool which enabled us to investigate feature extraction and selection more effectively. This tool has been instrumental for evaluation and analysis of different algorithms on different datasets. MFES tool was developed with JAVA programming language. Eclipse framework was chosen as the integrated development environment (IDE) due to its efficient coding and development functionalities. MFES tool comprises our original codes along with some third party libraries including XStream (object serialization and de-serialization), JAMA (Matrix Computation Library), libSVM (SVM classifier), jfreeChart (graph and chart library), UJMP (Data Matrix Visualization Tool), LDA.java and PCA.java source codes. Our novel code comprises approximately 35 classes with almost 30,000 lines of code. Broadly speaking MFES tool consists of two applications. FramePepperDataSetBuilder.java is for feature extraction purposes. FramePepperClassifier.java is on the other hand is our feature selection tool and contains various classification and analysis modules and components. **Figure 7.1** and **Figure 7.2** illustrate the general view of the main interfaces of the extraction and selection applications. In addition to this, **Figure 7.3** depicts the taxonomy of the MFES tool. Each main component/module will be presented in detail at the following sections.

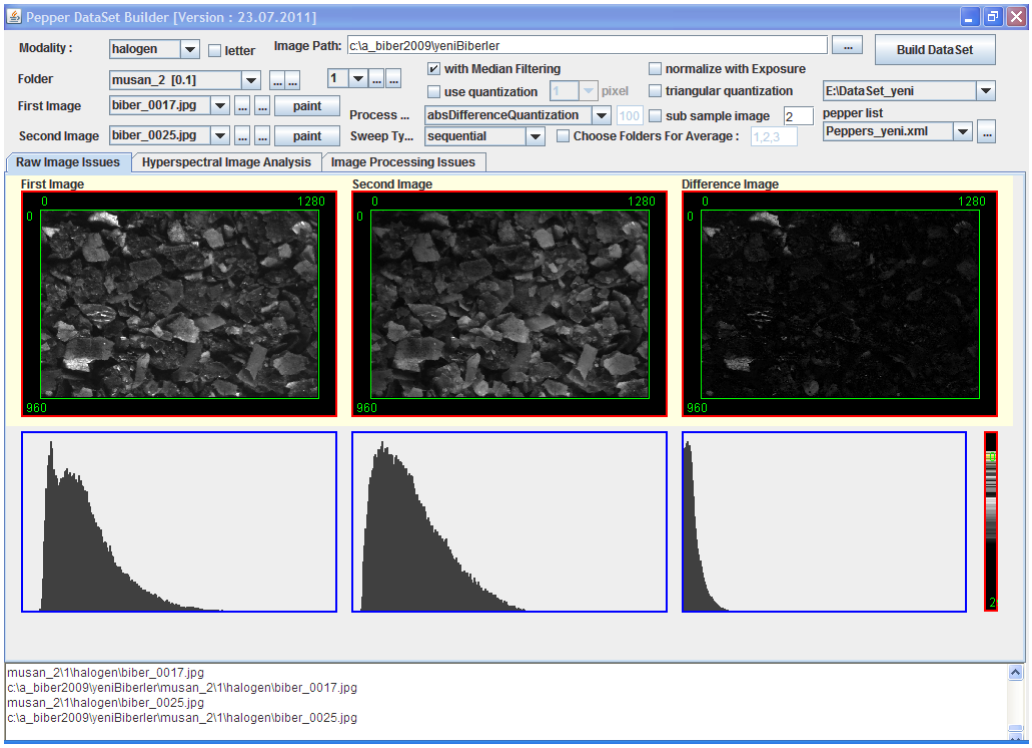


Figure 7.1 Main interface of the feature extraction application.

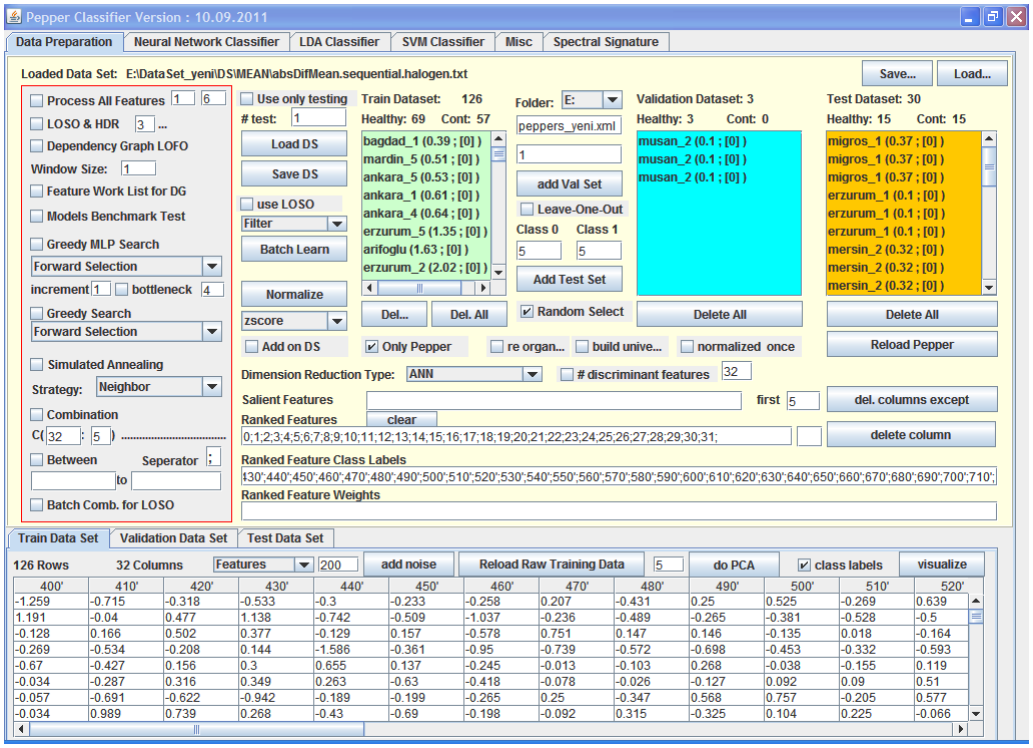


Figure 7.2 Main interface of the feature selection application.

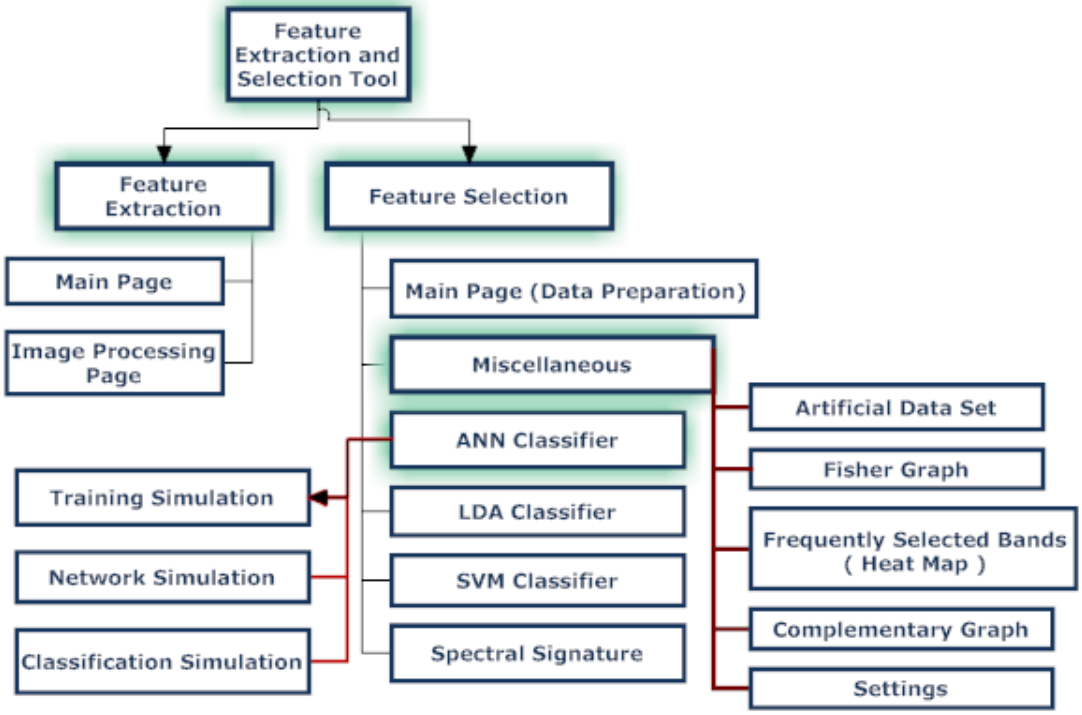


Figure 7.3 Some modules of the feature extraction and selection tool.

7.1 Feature Extraction Interface

Feature extraction interface comprises two pages as follows.

7.1.1 Main Page

The main page is shown in **Figure 7.1**. In this page, hyperspectral images can be loaded and demonstrated to the user. It is also possible to view the histogram of consecutive spectral bands along with absolute difference image and quantized histogram bars. The user can set various parameters to produce different types of features that we used in this dissertation.

7.1.1 Image Processing Page

User can apply several image processing operations on the hyperspectral images including histogram equalization, median filtering and region of interest selection in order to analyze specific image in detail.

7.2 Feature Selection Interface

7.2.1 Main Page

As **Figure 7.2** indicates, this page is used for data preparation and manipulation purposes. Several data centric operations such as choosing normalization and cross validation types and adjusting feature set size are handled on this page. In addition to these, feature selection strategies are also determined and necessary parameters are set. The classifiers (LDA, SVM, and MLP) then utilize those settings for searching the most discriminative feature subsets.

7.2.2 ANN Classifier

It is also known as the MLP classifier. We developed an efficient, flexible but extendible software module. One drawback of MLP is its black-box nature. The user can only observe the input and the output but the internal dynamics cannot be monitored. Our tool enables the user to observe the change of synaptic connection weights throughout the training phase. We developed three simulation components which are training, network and classification simulations and **Figure 7.4, 7.5** and **7.6** demonstrate them, respectively. As **Figure 7.4** (red frame on the left) indicates, free parameters of the MLP can be set before starting the training process. Some of them are as follows (from top to bottom)

- Setting the number of nodes in the first and second hidden layer. The default is Rapid Miner tool's best practice selected if auto is checked.
- Setting the number of iterations, learning rate and momentum can be done manually. These values are selected automatically managed by the heuristics mentioned in Section 6.3.3.1.
- Setting the weight initialization scheme. Default method is V-Shape but other selection schemes described in Section 6.1.3.3 is possible as well.

As it is shown in **Figure 7.4**, in the ANN classifier page, instantaneous training error curve and confusion matrix of the test results can be monitored at the central frame. In Figure 7.5, Network Simulation illuminating the MLP internal structure is called as Neural Interpretation Diagram (NID) as well (Özesmi and Özesmi, 1999).

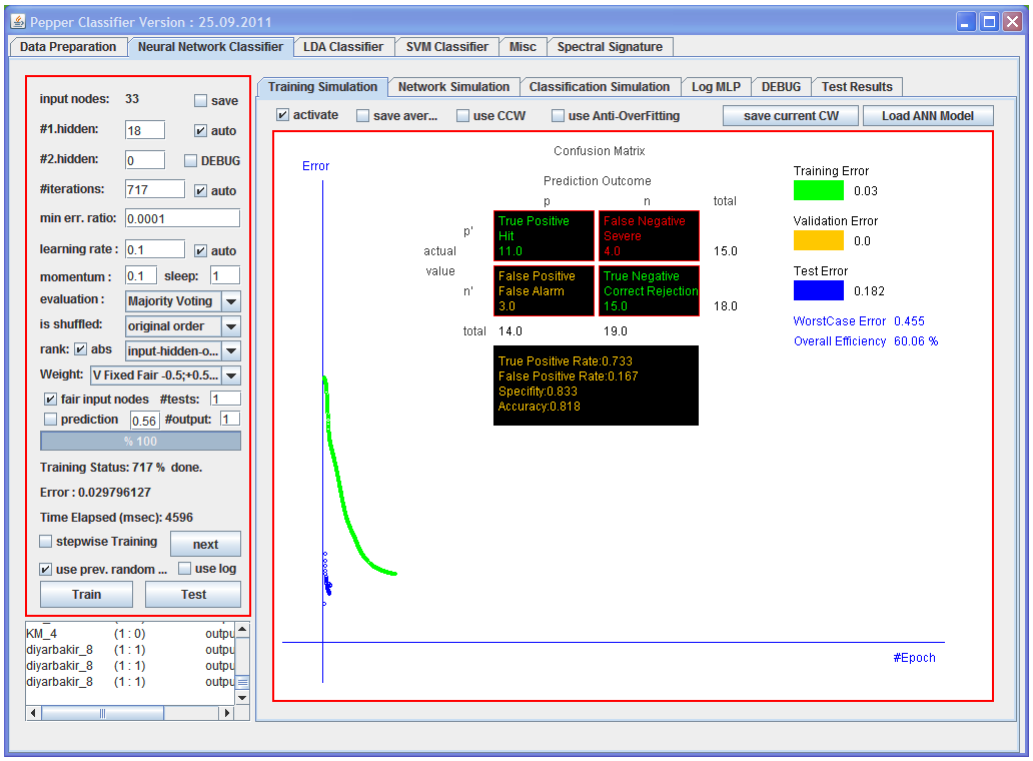


Figure 7.4 ANN classifier page.

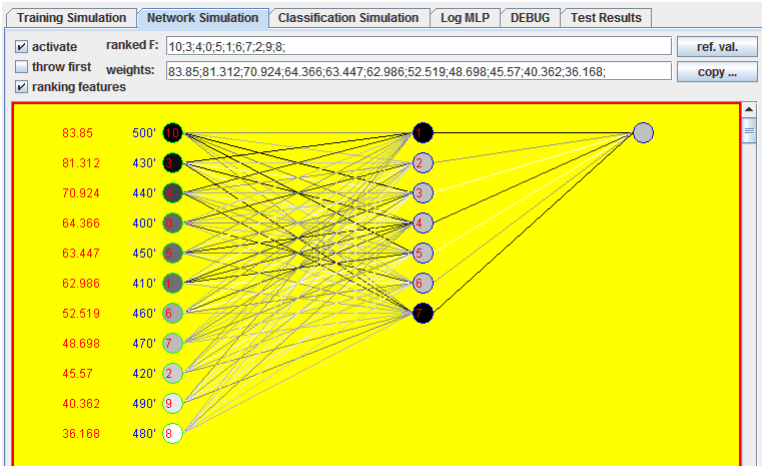


Figure 7.5 Network Simulation to illuminate the MLP internal structure. It is also known as Neural Interpretation Diagram (NID).

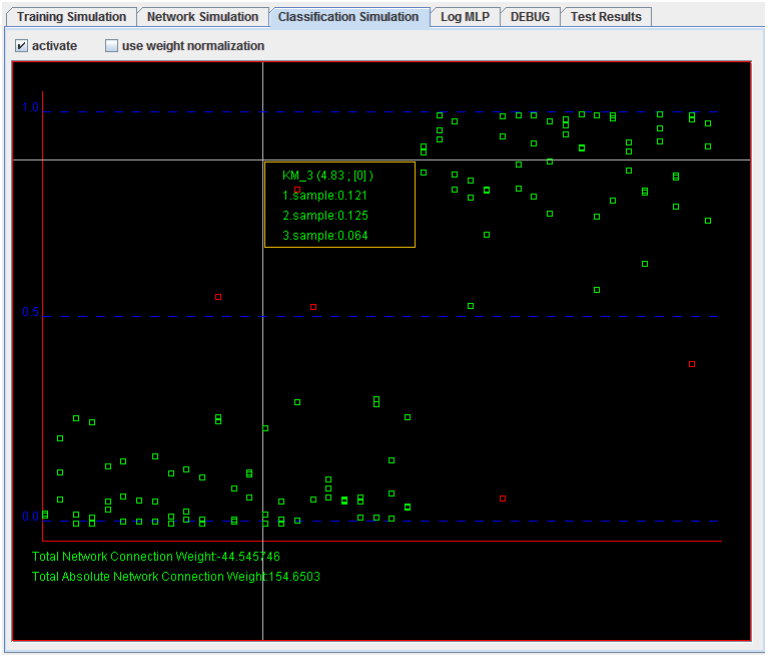


Figure 7.6 Simulation of the classification process.

LDA and SVM pages are similar to MLP main page where we can set the parameters and start to classification process.

7.2.3 Miscellaneous Page

In this page we developed several modules as follows:

- *Artificial Dataset Module:* For comparison and evaluation of the feature selection and MLP weight initialization schemes, various artificial datasets from different types of distributions with different number of features are generated as shown in Section 6.1.3.3.
- *Fisher Graph:* As it is demonstrated in **Figure 5.1** in Section 5.3 we plot the individual Fisher discrimination power values of features based on **Equation 3.11.**
- *Heat Map:* For visualizing the most frequently selected spectral bands and bins, we developed a heat map tool. A sample is given in **Figure 8.1** and **Figure 8.4**.
- *Complementary Dependency Graph:* Relationships amongst the features can be visualized and interpreted as in **Figure 6.12.**

CHAPTER 8

RESULTS AND DISCUSSIONS

In this thesis we utilized two chili pepper datasets. First one, we called as Dataset-1, comprises 53 new chili peppers and the second one (Dataset-2) is the dataset which were acquired by the previous researchers in our machine vision group (Kalkan et al., 2011). It contains 40 chili pepper samples that were screened under the UV illumination sources. Note that, they utilized 12 optical filters with 10 nm FWHM for the 400 nm to 510 nm spectral bands and 70 nm and 40 nm for the 550 and 600 nm spectral bands respectively (Kalkan et al., 2011). Images of three different locations of the same chili pepper samples were acquired to increase the training set. For Dataset-1, Hyperspectral image series with 33 spectral bands at two different illumination modes (halogen or UV) of 53 chili pepper samples were acquired. Images of three different locations of each chili pepper sample yielded a total of 10494 images with 1280x960 resolution.

It should be noted that, during the evaluation stage, three images of each chili pepper sample were isolated from the training data so that unbiased accuracy results can be achieved. The classification was performed based on majority voting over these three samples. By using Equation 5.1 and Equation 5.2, feature vector with size of 33 individual spectral band energy (ISBE) features and size of 32 absolute differences of consecutive spectral band energy (ABSDIF_CSBE) features

were extracted from Dataset-1. Similarly, 14 ISBE features and 13 ABSDIF_CSBE features from Dataset-2. The other two types of feature sets were extracted according to Equation 5.3. These are, quantized individual spectral band energy (Q_ISBE) features and quantized absolute difference of consecutive spectral band energy (Q_ABSDIF_CSBE) features. The total number of features in the quantized individual spectral band was 33 (spectral bands) x 12 (quantization bins) = 396. Similarly for the quantized absolute difference of consecutive spectral band, originally we had 32 (difference spectral band pairs) x 12 (quantization bins) which resulted in 384 features. As it is seen in **Figure 5.6**, there exists high number of zero value features in the feature set. MLP typically discard those features in the first step.

K-fold cross validation technique was utilized for the evaluation of generalization performance. In the machine learning community *K* is commonly selected as 10 or 5 (Breiman and Spector, 1992). Therefore in this study we used *K* as 5. We partitioned our data set randomly into five disjoint folds. Four folds were used for training and validation purposes and the remaining fold was utilized as the unseen test data for our predictive model. Since our data is limited, we would like to exploit all the data at the training and validation set. Thus, we preferred to employ leave one out cross validation (LOO-CV) technique for training and validation purposes. The final decision on aflatoxin presence is made using majority voting on the three images of the same chili pepper. This process was repeated for each fold and average accuracy rate was computed from the five folds test results. All the data were normalized according to Z-Score normalization yielding a data distribution with zero mean and unit variance. Throughout the thesis, all the trials were run on an Intel Core i7 2.79 GHZ four-core CPU desktop PC with 4 GB RAM at the Informatics Institute's Virtual Reality and Computer Vision Laboratory. Six applications could be run simultaneously. One complete experiment consists of 53 training and test processes. The time required is dependent on the number of features used and the

type of the classifier. More precisely, LDA is very efficient in terms of convergence speed and least susceptible to the size of the feature set. On the other hand, SVM performs optimization process and therefore estimating the time duration is very difficult. In the case of MLP, convergence speed is directly related to the feature set size but still it can be estimated. **Table 8.1** demonstrates average durations needed for one experiment against the classifier used.

Table 8.1 Approximate durations of the classifiers for a single experiment.

Classifiers	Time needed for a single experiment (LOO-CV) in seconds
MLP	50 - 120 s, average : 85 s
LDA	1 - 3 s, average : 1.5 s
SVM	1 - 50 s, average: hard to estimate

Since one of the main objectives of this thesis is to find the discriminative spectral bands, we performed three different types of trials. The experimental results are given in the upcoming sections.

8.1 Exhaustive Search Results

Exhaustive search is a combinatorial problem. It has potential to converge to the global optima but computational cost grows exponentially with the dimension of the search space which makes search intractable for large number of features. For *N* number of features, exhaustive search theoretically investigates all subsets. Suppose we have 33 features, *8589934591* feature subsets (trials) need to be performed in order to find optimal solution. Even for the fastest classifier LDA it would take about 45 years in order to complete all the combinations. We limited the feature size so that experiments should last at most one week. **Table 8.2** and **Table 8.3** show best generalization accuracy rates of the classifiers for Dataset-1 and for Dataset-2, respectively. Exhaustive search proceeded up to 5 features for

DataSet-1. For DataSet-1, all combinations were investigated due to its low feature size. Note that QHM features were not investigated because exhaustive search was unfeasible. All the evaluations have been performed based on LOO-CV.

Table 8.2 Exhaustive search results of various feature sets up to 5 features and corresponding spectral bands for Dataset-1.

DataSet-1 Extracted Features	**Illumination Type**	**Classifier Type**	**Subset Size**	**Accuracy Rate %**	**Optimal Spectral Bands**
Individual Band Energies 33 features	UV	LDA	3	71.70	600;690;710; 600;690;720;
		SVM	4	73.60	490;560;700;720; 490;540;560;700;
		MLP	5	**79.20**	420;580;660;670;690;
	Halogen	LDA	5	77.40	610;620;650;660;690;
		SVM	5	81.10	420;570;590;620;680;
		MLP	3	**84.90**	530;570;680;
Absolute Difference Band Energies 32 features	**UV**	LDA	5	75.50	420';450';520';540';570';
		SVM	5	75.40	430';480';600';630';650';
		MLP	5	**77.40**	410';460';510';620';650';
	Halogen	LDA	5	84.90	430';440';490';530';640';
		SVM	4	84.90	420';440';540';680';
		MLP	5	**90.60**	400';520';540';550';580'; 430';440';480';540';690';
Teager Energies 31 features	**UV**	LDA	5	75.50	420';430';560';640';680';
		SVM	5	77.40	490';520';530';580';690';
		MLP	5	**77.40**	550';560';640';660';700';
	Halogen	LDA	4	77.40	470';490';530';560';
		SVM	4	75.50	420';560';660';680';
		MLP	5	**81.10**	430';470';520';530';580';

Table 8.3 Complete exhaustive search results of various feature set and corresponding spectral bands for Dataset-2.

Dataset-2 Extracted Features	Illumination Type	Classifier Type	Subset Size	Accuracy Rate %	Optimal Spectral Bands
Individual Band Energies 14 features	UV	LDA	4	82.50;	410;420;490;600;
		SVM	2	80.00	400;550;
		MLP	4	**87.50**	430;440;490;510;
Absolute Difference Band Energies 13 features	UV	LDA	3	75.00	420';470';550';
		SVM	2	72.50	420';470';
		MLP	4	**85.00**	400';460';470';490';
Teager Energies 12 features	UV	LDA	1	80.00	490';
		SVM	4	**85.00**	420';450';490';550';
		MLP	6	**85.00**	410';460';490';500';510';550';

Superscripts on the spectral bands are inserted to remind that the computation of each feature at a given spectral band uses multiple neighboring bands for absolute difference and Teager energy features as in **Equations 5.2** and **5.4**.

Table 8.4 Comparison of the exhaustive search results of Dataset-1 and Dataset-2 for 12 spectral bands (400-510) under the UV illumination.

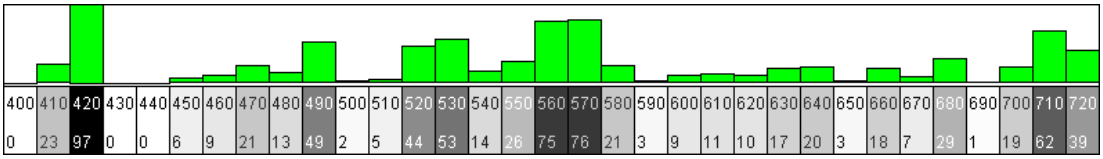
Extracted Features	Dataset-1			Dataset-2		
	Classifier Type	Accuracy Rate %	Optimal Spectral Bands	Classifier Type	Accuracy Rate %	Optimal Spectral Bands
ISBE 12 features	LDA	70.00	**410**;**420**;440;500; 510	LDA	80.00	**410**;**420**;**480**;490; 510
	SVM	62.50	**410;**420;**430;480; 510;**	SVM	77.50	400;**410**;**430**;450; 490;510;
	MLP	**75.00**	410;**430**;**450**;470;	MLP	**87.50**	**430**;**450**;**490**;510;
ABSDIF CSBE 11 features	LDA	75.00	**410';440';490';500'**	LDA	75.00	**400';420';470';490'**
	SVM	72.50	**420'**;460';500';	SVM	72.50	420';**470'**;
	MLP	**77.50**	**400';410';480';490' ;500';**	MLP	**85.00**	**400'**;**450'**;470';490';
Teager Energy 10 features	LDA	70.00	410',440',460',**490'**	LDA	80.00	**490'**;
	SVM	62.50	410',**420'**,470',**490'**	SVM	**85.00**	**420'**;450';**490'**;
	MLP	**72.50**	**410', 450', 470'**	MLP	**85.00**	**410'**;460';490'; **500'**;

Table 8.2 reveals that ABSDIF_CSBE features are superior to the Teager energy and ISBE features. Furthermore, according to **Table 8.3** still ABSDIF_CSBE features can

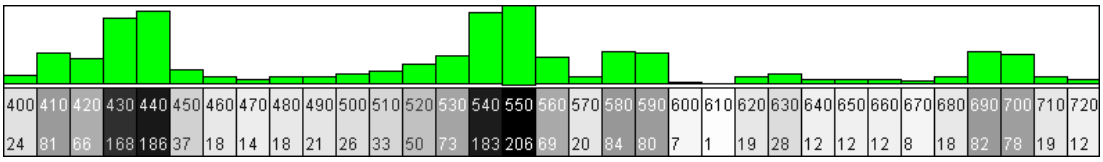
compete with the Teager energy features but, both of them are inferior to ISBE features. This difference could be due to the different acquisition procedures of both datasets. Detailed discussion will be presented after presenting the HBBE results at Section 8.3. We further investigated the improvement provided by additional features on the feature subsets of exhaustive search results for Dataset-1 (**Table 8.2)**. Up to two new features were added to the five features from the exhaustive search by applying forward selection strategy. Significant improvement for halogen ABSDIF_CSBE features were obtained by using MLP as the classifier with these seven features. Both seven feature sets 400';520';540';550';580';680';710'; and 400';520';530';540';550';580';610'; yielded 94.30% classification accuracy rate. Since Dataset-2 originally has 14 features which are 400-510 nm, 550 and 600 nm under the UV illumination, we chose the spectral bands between 400-510 nm and the UV illumination modality from Dataset-1 as well. 550 and 600 nm were not added since FWHM of them are greater than 10 nm. Moreover, chili pepper samples around 10 ppb are removed so that both datasets have 40 samples. **Table 8.4** reveals that, for almost the same conditions, for ISBE features, 410, 430, 450, 480 and 490 nm spectral bands were frequently selected spectral bands for both datasets. For ABSDIF_CSBE features, 400', 420', 450', 470', and 490' features were commonly selected. In addition, for Teager energy features the mostly selected features were 410', 420', 490' and 510'.

As it is seen in both **Table 8.2** and **Table 8.3**, the best accuracy rates were achieved by MLP; therefore we introduced a map of the most frequently selected spectral bands (heat map) over the exhaustive search for feature subsets based on MLP. Specifically, whenever a spectral band is selected throughout the exhaustive search procedure its score is incremented. **Figure 8.1** and **Figure 8.2** depict the heat maps of different feature sets for two datasets. Each heat map is given both as a bar graph and intensity depiction. There are three heat maps in each figure, ISBE feature set is located at the top, ABSDIF_CSBE feature set is in the middle and

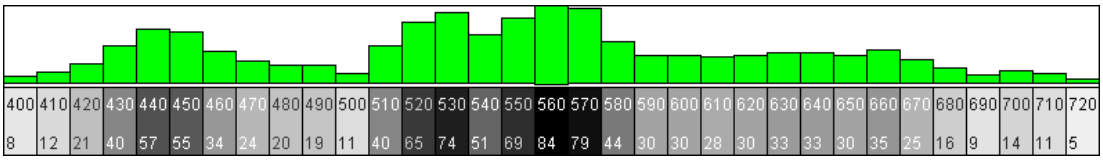
Teager energy feature set is at the bottom. The spectral band and its corresponding vote are indicated at the top and the bottom of the cells, respectively. Darker color indicates higher scores whereas lighter color implies lower scores.



Individual (accuracy rate >=75%)

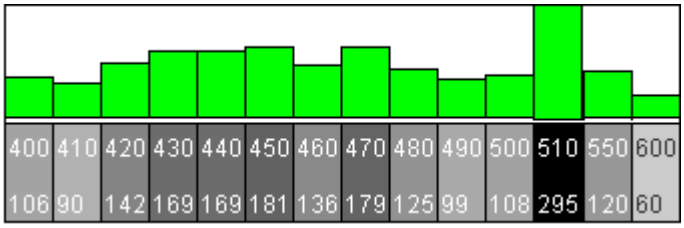


AbsDif (accuracy rate >= 85%)

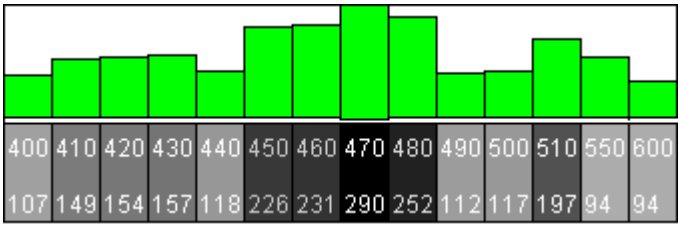


Teager (accuracy rate >= 75%)

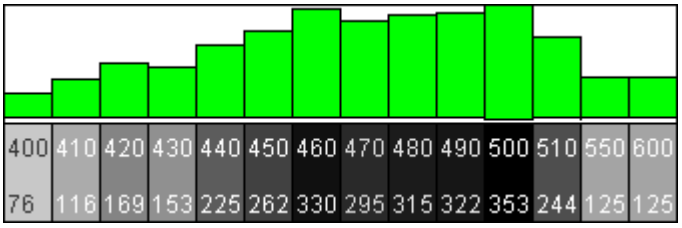
Figure 8.1 Most frequently selected spectral bands of exhaustive search for Dataset-1.



Individual (accuracy rate >=85%)



AbsDif (accuracy rate >= 75%)



Teager (accuracy rate >=80%)

Figure 8.2 Most frequently selected spectral bands of exhaustive search for Dataset-2.

8.2 Stochastic Simulated Annealing Search Results

As mentioned in the previous section, for Dataset-1 we could not perform exhaustive search beyond five features for ISBE, ABSDIF_CSBE and Teager energy features due to the time constraint. For investigating feature size further, we continued with the simulated annealing method. Since we already completed examining all subset combinations of Dataset-2 there was no need to use SA method. Again QHM features were not examined thorough exhaustive search and SA due to high number of features.

Table 8.5 demonstrates SA results of various types of feature set in Dataset-1 for different classifiers and corresponding sub-optimal results of the most discriminative spectral band as in **Table 8.2**. Note that, SA method utilized LOO-CV for evaluation criterion as well.

Table 8.5 SA search results of various feature sets up to 10 features and corresponding sub-optimal results of spectral bands for Dataset-1.

DataSet-1 Extracted Features	Illumination Type	Classifier Type	Subset Size	Accuracy Rate %	Sub-Optimal Spectral Bands
Individual Band Energies 33 features	UV	LDA	7	73.60	490;500;570;580;620;650;670;
		SVM	6	75.50	490;520;530;540;560;700;
		MLP	6	**83.00**	410;440;450;500;530;710;
	Halogen	LDA	6	75.50	420;460;500;570;590;620;
		SVM	7	81.10	420;450;490;570;620;680;710;
		MLP	7	**86.80**	410;470;490;560;580;590;720;
Absolute Difference Band Energies 32 features	**UV**	LDA	6	75.50	530'; 610';630'; 640';650';670';
		SVM	9	81.10	400';550'; 600';630';650';670';680';690';
		MLP	6	**86.80**	450';460';610'; 630';640';650';
	Halogen	LDA	8	79.20	440';460';480';520';540';560';580';680';
		SVM	6	83.00	400'; 490';540';550';640';680';
		MLP	9	**92.50**	**420';430';440';470';500';610';640';650'; 670';**
Teager Energies 31 features	**UV**	LDA	6	75.50	490';500';600';620';680';690';
		SVM	6	75.50	420';460';490';560';660';680';
		MLP	7	**83.00**	430';500';510';530'; 650';660';670';
	Halogen	LDA	6	75.50	540';560'; 570';580';640';650';
		SVM	6	79.20	510'; 520';530';580'; 680';700';
		MLP	7	**85.00**	420';520';530';590';640';560';710';

When we compare **Table 8.2** and **Table 8.5** we observed that, stochastic search yields slightly higher accuracies. One interesting point is that better results were obtained around 6 features and further increasing the number of features decreased the classification accuracy. This outcome might also be due to the fact that at the higher feature size search space grows exponentially and the SA algorithm cannot sample the search space effectively.

Figure 8.3 shows heat maps of the most frequently selected spectral bands by applying SA in Dataset-1. We did not apply SA search on Dataset-2 since exhaustive search has already covered all the combinations.

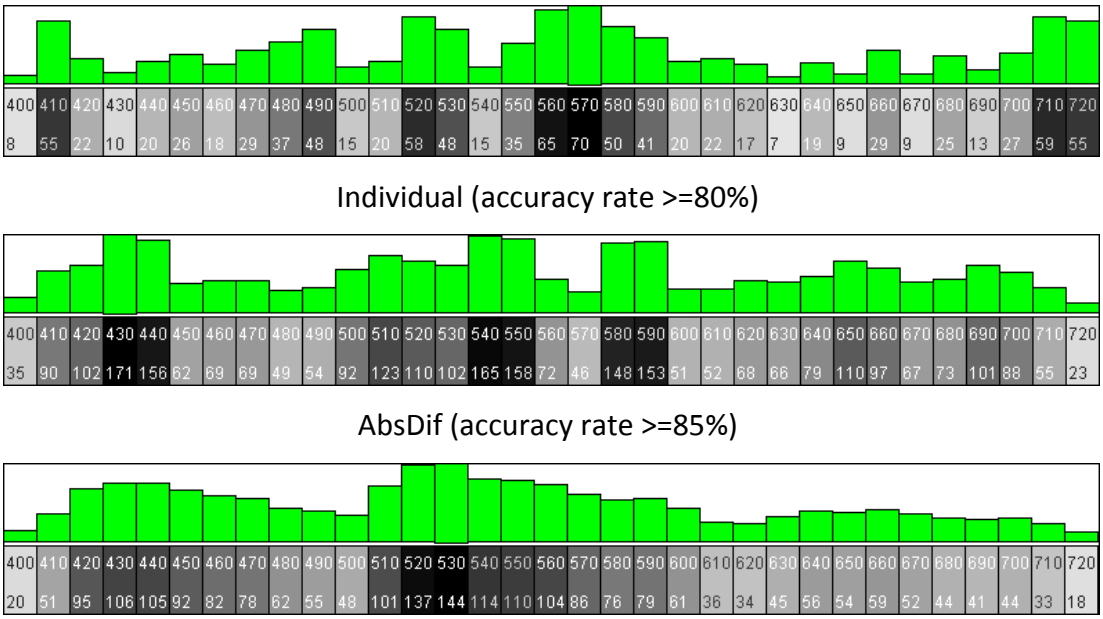


Figure 8.3 Most frequently selected spectral bands of SA search for Dataset-1.

8.3 Proposed HBBE Feature Selection Results

In order to assess the effectiveness of our proposed method we compared its classification accuracy rates with those of the original features and reduced features by applying PCA, Fisher and mRMR methods. We utilized 5-Fold cross validation technique for the evaluation purposes. **Table 8.6** shows overall accuracy rates of several feature sets with various feature selection methods under the halogen and UV illuminations. Note that, classification was performed based on majority voting over three different samples of the same chili pepper. As it is observed in **Table 8.6**,

no performance improvement is observed by applying PCA to original features using MLP.

Table 8.6 Classification accuracy rates for Dataset-1 with the extracted features versus classifiers and feature selection methods based on 5-Fold CV.

Dataset-1	Feature Selection Method	ISBE 33 f		Q_ISBE 396 f 12 bins		ABSDIF_CSBE 32 f		Q_ABSDIF_CSBE 384 f 12 bins	
		Halogen	UV	Halogen	UV	Halogen	UV	Halogen	UV
	Original	37.64	33.84	48.88	48.88	50.70	48.90	50.88	48.90
LDA	PCA	41.24	45.44	58.52	45.06	51.08	43.46	45.06	45.06
Classification	Fisher	35.64	41.82	56.16	56.90	37.26	38.00	52.14	52.54
Accuracy %	mRMR	41.26	37.46	60.16	58.16	46.70	39.64	56.52	50.07
	HBBE	34.00	35.82	60.70	50.90	43.64	49.44	61.98	52.90
	Original	64.16	53.10	69.80	52.70	62.16	65.80	64.52	61.98
SVM	PCA	56.70	47.44	60.16	50.88	58.34	65.98	56.88	58.34
Classification	Fisher	64.34	56.88	60.16	51.08	60.16	60.16	64.16	52.88
Accuracy %	mRMR	60.34	58.52	59.98	56.16	65.62	67.46	68.16	58.88
	HBBE	65.80	62.16	61.98	55.08	67.62	**68.52**	69.98	65.80
	Original	62.34	63.80	61.98	58.34	62.52	62.52	60.34	58.34
MLP	PCA	43.24	50.70	52.70	51.06	53.08	50.72	46.54	49.06
Classification	Fisher	58.70	61.98	60.16	58.52	62.34	54.16	69.98	54.16
Accuracy %	mRMR	58.52	62.16	67.98	52.90	62.70	57.98	71.80	55.98
	HBBE	**68.16**	**69.80**	**81.26**	**62.14**	**71.62**	65.98	**83.26**	**67.98**

It can be seen from **Table 8.6** that, taking the absolute difference of consecutive spectral bands improves the classification performance for both halogen and UV excitations. Moreover, performance improvement is consistent for MLP, SVM and LDA classifiers. Similarly quantization process increases the accuracy rate for all feature sets and classifiers.

Table 8.7 compares the best results obtained by our proposed methods in both Dataset-1 and Dataset-2 to wavelet LDB method of Kalkan et al (2011). As it is seen in the table, our proposed methods outperform wavelet LDB method.

Table 8.7 Benchmark of the proposed method against wavelet LDB method for Dataset-1 and Dataset-2 based on 5-Fold CV.

Dataset	Illumination Source	Feature Extraction Type	Feature Selection	Classifier Type	Accuracy Rate%
Dataset-1	Halogen	QABSDIF_CSBE (12 bins)	HBBE	MLP	**83.26**
			HBBE	LDA	61.98
		Wavelet Features	LDB	MLP	63.64
			LDB	LDA	62.44
	UV	QABSDIF_CSBE (12 bins)	HBBE	MLP	**67.98**
			HBBE	LDA	52.90
		Wavelet Features	LDB	MLP	61.82
			LDB	LDA	57.10
Dataset-2	UV	QISBE (25 bins)	HBBE	MLP	**87.50**
			HBBE	LDA	67.50
		Wavelet Features	LDB	MLP	77.50
			LDB	LDA	74.65

In the case of Dataset-2, the quantized individual band energy features yield better accuracy rates than the quantized absolute difference energy features. This may be due to the fact that for Dataset-1, we fixed the camera gain parameter to 850 electron/CCD and manually changed the exposure time instead of using the histogram equalization feature of the camera. On the other hand, Dataset-2 was acquired under the automatic gain parameters enabling automatic histogram equalization which may modify the original spectral signal. Although histogram equalization aims to enhance the image quality and yields visually appealing images, it will also modify the spectral signature. This may degrade the features based on absolute difference of consecutive spectral bands more than individual band energy features. Therefore individual spectral bands may contain more informative pattern than the absolute difference of consecutive spectral bands. As a result, wavelet features and quantized individual spectral band energy features produced relatively better results than the absolute difference features under the UV excitation.

Table 8.6 and **Table 8.7** reveal that aflatoxin detection in chili pepper problem is not a linearly separable problem because almost in all cases MLP outperforms both LDA and linear SVM in terms of classification accuracy.

Figure 8.4 illustrates the two dimensional heat map of most frequently selected features by 5 fold cross validation of QHM features based on HBBE feature selection under the halogen illumination for Dataset-1. The features selected at each fold are added to the corresponding bin. As the QHM features require two consecutive bands to be used, each feature contributes to the tally in two bins. In **Figure 8.4**, the most discriminative spectral bands and features can be seen. According to this figure, the most informative bands are 430, 440, 540, 550, 560, 580, 590, 640, 650 and 660 nm.

As it seen in **Table 8.6**, the highest classification accuracy on Dataset-1 was obtained with absolute difference of QHM features selected by the HBBE feature subset selection using MLP classifier as 83.26% under halogen illumination. Similarly, the highest classification accuracy on the Dataset-2 was obtained with individual band of QHM features selected by the HBBE feature subset selection using MLP classifier as 87.5% under UV illumination.

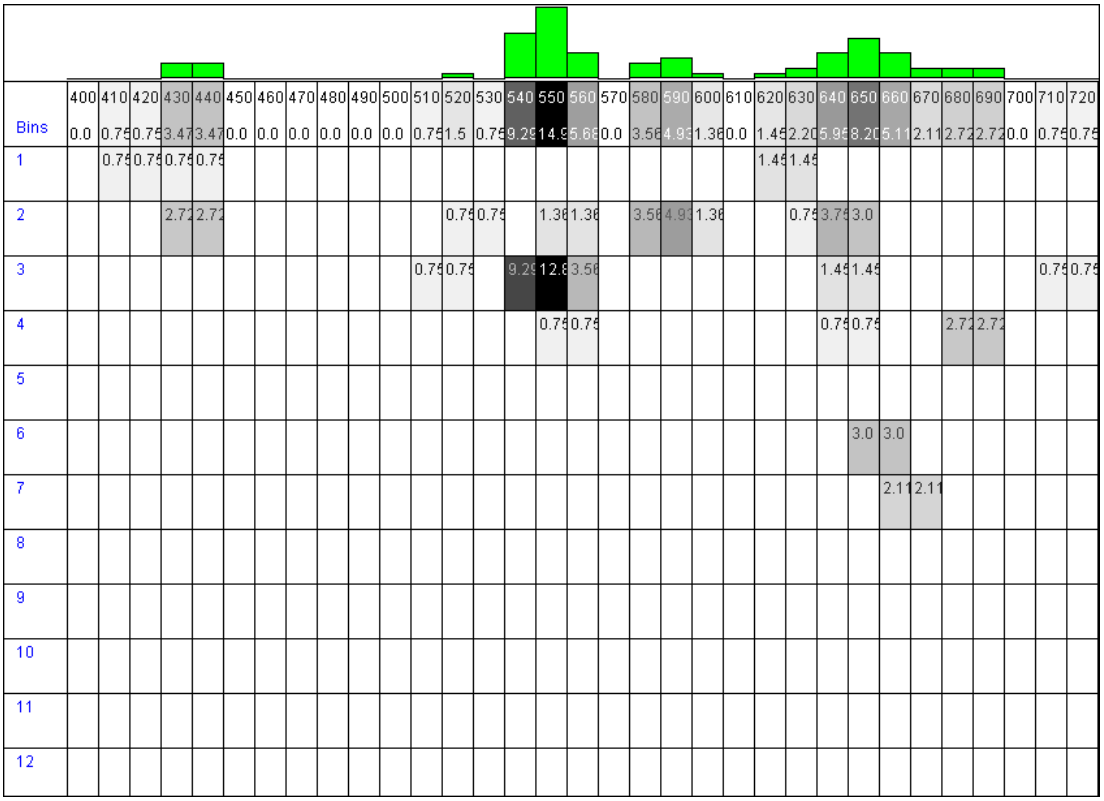


Figure 8.4 Dataset-1 heat map for visualizing the most frequently selected spectral bands with the associated bin numbers.

In **Figure 8.4**, the votes for each feature are indicated on the corresponding feature cell. The frequency of a particular spectral band is the total of the votes of all bins in that spectral band and it is located at the first row of the heat map. Darker colors indicate high counts whereas lighter colors correspond to low counts. Each cell below the first row indicates quantized histogram features that were utilized in **Table 8.6** and **Table 8.7.** Darker cells are the most discriminative features with higher number of votes. They are obtained by the proposed HBBE feature selection method. Likewise, empty cells are the least discriminative ones. We can extract the series of most discriminative spectral bands by applying different threshold values. **Table 8.8** lists such discriminative spectral bands based on different threshold values associated with LOO-CV accuracy rates.

Table 8.8 Most discriminative spectral bands based on different threshold values associated with LOO-CV accuracy rates for the Dataset-1.

Threshold	Selected Bands (nm)	Feature size	LOO-CV Accuracy Rate %
0.5	410-440, 510-560, 580-600, 620-690, 710,720	20	85
1.5	430,440,520,540,550,560,580,590,630-690,	17	81
3.5	540,550,560,580,590,640,650,660	12	78
4.5	540,550,560,590,640,650,660	10	68
6	540,550,650	4	62

Table 8.8 indicates that there is a tradeoff between the accuracy rate and number of features selected which will determine the design of the machine vision system. Higher classification performance may require excessive numbers of spectral filters which will increase the complexity of the overall machine vision system. On the other hand, establishing a relatively simpler machine vision system can be realized at the expense of lowering generalization performance.

We repeated the same scenarios for the Dataset-2 as well. **Figure 8.5** and **Table 8.9** demonstrate the heat map of the most frequently selected spectral bands and threshold features. It is seen in **Table 8.9**, for the Dataset-2, a simpler machine vision system can be established with single spectral band of 420 nm with 85% classification accuracy. On the other hand, if 400 nm is also used in the system, 90% classification accuracy would be possible.

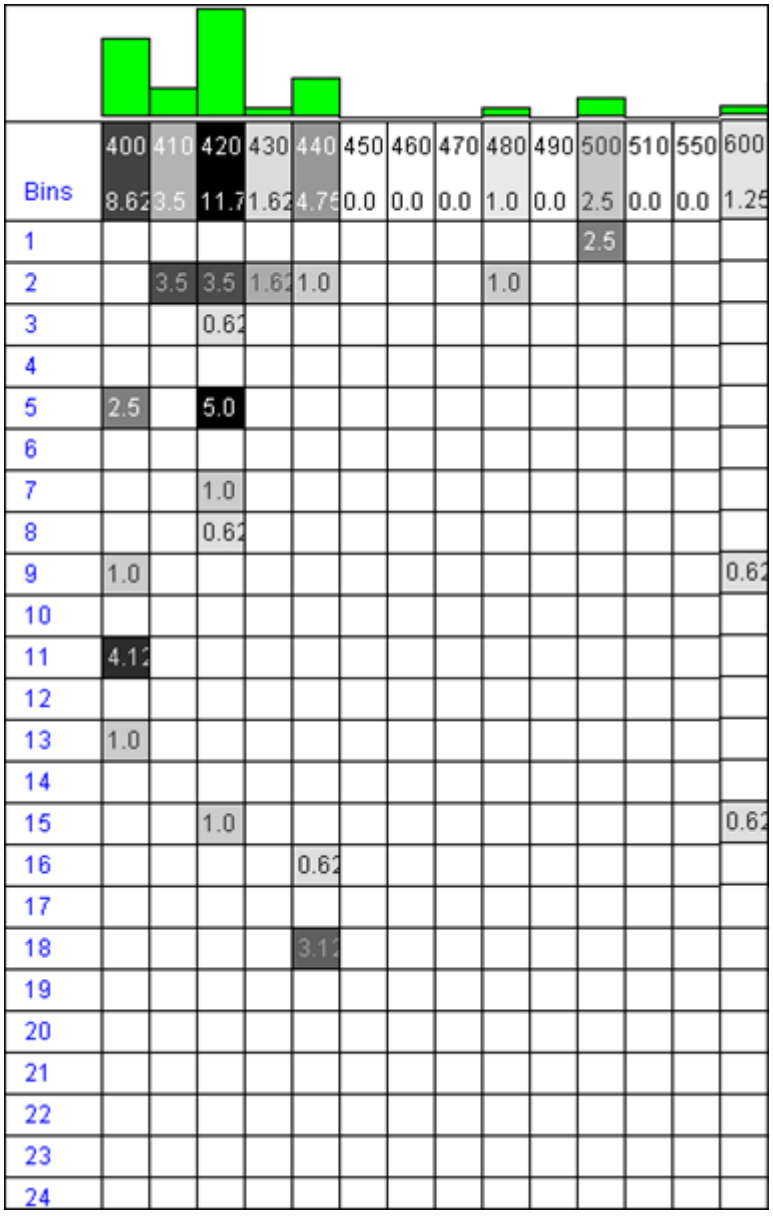


Figure 8.5 Dataset-2 heat map for visualizing the most frequently selected spectral bands with the associated bin numbers.

Table 8.9 Most discriminative spectral bands based on different threshold values associated with LOO-CV accuracy rates for the Dataset-2.

Threshold	Selected Bands (nm)	Feature size	LOO-CV Accuracy Rate %
1	400,410,420,430,440,480,500,600	19	85
2	400,410,420,430,440,500,600	17	85
4	400,420,440	13	90
6	400, 420	10	90
9	420	6	85

8.4 Comparison of the Weight Initialization Schemes

In order to assess the effectiveness of different weight initialization schemes, we performed several benchmark tests on 10 different datasets downloaded from UCI machine learning repository (Frank and Asuncion, 2011). In addition, four artificial datasets with different types of distribution and feature sizes and the chili pepper dataset were employed. **Table 8.10** lists the generalization errors of MLP classifier on different datasets with different weight initialization schemes. Note that, free parameters of the MLP other than the initialization of the connection weights were adjusted based on the principles mentioned in **Section 6.1.3**. Best classification accuracies for each dataset were indicated in bold. Then, the number bold values for each weight initialization scheme were summed up and indicated in the last row (Vote). V-Shape weight initialization scheme has the highest number of votes as shown in the last row of **Table 8.10**; therefore we utilized this scheme for MLP connection weight initialization method throughout the dissertation.

Table 8.10 Benchmark of the various weight initialization schemes against different dataset for MLP classifier.

Dataset	Features & Samples	Accuracy Rate % LOO-CV													
		Nguyen	Zero	Constant (0.25)	Random	Sym.Random	+Linear	- Linear	V Shape	^ Shape	Gaussian	Sigmoid	Hyperbolic	Rectangular	Alternate
Breast Cancer	10 f 699 s	94.9	94.5	94.7	94.7	94.8	94.7	94.7	94.9	94.8	94.9	**95.4**	94.7	94.7	94.9
Heart Disease	13 f 303 s	78.7	70.3	78.6	78.5	78.6	78.6	78.6	**79.1**	78.9	**79.1**	78.3	78.6	78.8	78.6
Ozone	73 f 911 s	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**
Iris	4 f 100 s	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**
Parkinson	23 f 195 s	97.0	94.0	97.0	97.0	96.0	**99.0**	**99.0**	**99.0**	98.0	98.0	93.0	**99.0**	91.0	91.0
Spam	57 f 1905s	94.9	95.5	96.0	95.7	95.7	96.0	96.0	**96.1**	**96.1**	95.9	95.1	96.0	94.5	94.7
Steel Fault-1	27 f 403 s	93.0	92.0	93.5	93.7	93.7	93.8	93.7	93.7	**94.0**	93.8	93.8	**94.0**	92.5	92.0
Steel Fault-2	27 f 935 s	67.2	63.6	67.1	67.3	67.2	**67.5**	**67.5**	67.3	67.1	67.1	66.4	**67.5**	63.8	57.1
Spect Heart	22 f 80 s	67.1	67.1	66.9	67.1	67.1	**67.9**	**67.9**	67.6	67.7	66.4	66.4	67.1	67.9	66.5
Sonar	60 f 208 s	77.4	75.3	78.0	77.5	77.3	**78.3**	**78.3**	**78.3**	**78.3**	78.0	78.2	77.5	77.1	76.0
Artificial-1	5 f 1500s	78.1	77.9	78.4	78.3	78.2	78.1	78.1	**78.9**	**78.9**	78.4	77.9	78.1	78.1	78.1
Artificial-2	11 f 1500s	71.7	71.6	71.7	71.6	71.6	71.7	71.7	**72.0**	**72.0**	**72.0**	71.6	71.7	71.8	71.8
Artificial-3	5 f 200 s	**97.0**	**97.0**	96.9	96.9	96.9	96.9	96.9	96.9	96.9	**97.0**	**97.0**	96.9	96.9	**97.0**
Artificial-4	11 f 200 s	79.2	77.9	78.3	77.9	78.1	78.5	78.5	**79.0**	78.9	**79.0**	78.7	78.6	78.5	78.5
Pepper	5 f 159 s	87.6	71.7	86.7	87.4	87.4	86.8	86.8	**90.6**	**90.6**	88.7	88.7	**90.6**	88.7	71.7
Vote		**3**	**3**	**2**	**2**	**2**	**6**	**6**	**10**	**8**	**6**	**4**	**6**	**3**	**3**

8.5 Comparison of the Proposed HBBE method on Real World Datasets

We performed several benchmark tests to show the effectiveness of our proposed HBBE feature selection method. For comparison purposes, Fisher and MRMR methods were utilized. Throughout the experiment, we used 10 real world datasets downloaded from UCI machine learning repository (Frank and Asuncion, 2011) and two artificial datasets generated by two Gaussian distribution with $\mu_1 = 0.9, \sigma_1 = 2.5$ and $\mu_2 = 1.1, \sigma_2 = 3.5$, respectively. We employed MLP as a complex and non-linear classifier. The LDA and linear SVM are selected as simpler and linear classifiers. 5-Fold cross validation technique was utilized for evaluation purposes. **Table 8.11** lists overall classification accuracies of several dimensionality reduction methods versus different real world datasets and artificial datasets for LDA, SVM and MLP classifiers. In addition, we tabulated the accuracy rates of previous studies performed by other researchers on the same datasets. The accuracy rate provided for each dataset is the highest (to the best of our knowledge) among the studies reported in the literature. In this way, the accuracy of our proposed HBBE method can be compared to those studies. As it is seen in **Table 8.11**, HBBE feature subset selection method usually outperformed other feature subset selection methods for all classifiers. The difference was more prominent for the MLP classifier. This might be due to the nonlinear characteristics of MLP. In addition, when MLP is employed for feature selection using wrapper approach and then for classification, the classifier will be easily tuned to the remaining feature set. This phenomenon was also observed by (Guyon and Elisseeff, 2003).

Table 8.11 Benchmark of the various dimension reduction approaches against different datasets for LDA, SVM and MLP classifiers.

Dataset	Features & Samples	Accuracy Rate %, 5-Fold CV												Other Studies		
		LDA				SVM				MLP						
		Before FS	Fisher	MRMR	HBBE	Before FS	Fisher	MRMR	HBBE	Before FS	Fisher	MRMR	HBBE	Accuracy	Method	Reference
Sonar	60 f, 208 s	46.9	65.6	66.3	67.4	64.5	68.1	70.4	70.2	82.4	72.2	80.3	**86.2**	77.4	RBF Net	(Liang and Abernethy, 2011)
Parkinson	22 f, 195 s	53.4	50.5	51.4	55.0	63.2	65.4	71.1	73.0	80.6	84.9	86.8	**88.6**	86.2	Lib linear	(Liang and Abernethy, 2011)
Spam	166 f, 1904s	63.8	95.0	95.8	96.6	77.1	94.5	95.6	94.6	96.9	94.6	96.7	**97.3**	90.8	BP ANN	(Kiran et al., 2004)
Steel Fault	27 f, 865 s	69.8	78.6	79.7	83.2	74.5	82.2	84.5	86.0	95.4	89.8	91.2	**95.0**	84.5	BP ANN	(Buscema et al., 2010)
Ozone	72 f, 911 s	91.8	95.2	95.3	95.6	92.0	94.0	94.9	94.7	94.3	94.1	95.0	**96.5**	97.1	RBF Net	(Su, 2010)
Breast-1	10 f, 699 s	55.8	66.4	66.8	67.4	64.8	72.3	80.0	82.3	96.1	96.1	96.3	**97.4**	96.2	NN-Ruby	(Liang and Abernethy, 2011)
Breast-2	30 f, 569 s	34.2	40.4	41.1	42.2	54.5	62.3	70.4	74.5	95.7	95.8	97.0	**97.3**	97.0	Comitee	(Liang and Abernethy, 2011)
SPECT	22 f, 267 s	79.4	80.0	80.2	80.4	82.1	80.4	80.3	81.1	82.0	81.4	78.2	**83.6**	76.7	NaïveBayes	(Madden, 2002)
Ionosphere	33 f, 351 s	65.0	79.0	80.5	81.4	76.7	77.5	78.9	77.9	85.5	86.7	86.5	**87.5**	91.1	N.Ensemb	(Zhou and Jiang;, 2004)
Monks	6 f, 124 s	70.3	70.0	70.2	70.7	73.3	74.8	76.6	75.4	80.9	76.2	76.2	**83.7**	68.2	Ecob Web	(Thrun et al., 1991)
Artificial-1	110 f, 50 s	42.8	53.6	56.1	67.6	45.4	53.0	55.1	56.5	58.7	51.3	58.8	**80.0**			
Artificial-2	33 f, 150 s	68.0	71.2	73.2	77.6	71.5	72.5	73.0	72.7	76.4	76.8	72.8	**80.4**			

CHAPTER 9

CONCLUSION AND FUTURE WORK

In this dissertation, detection of aflatoxin contaminated chili pepper problem was investigated. Both UV and Halogen illuminations were used. Hyperspectral image series of 53 chili pepper samples (Dataset-1) from 400 nm to 720 nm with 10 nm bandwidth were acquired. In order to enrich the training set, images from three different locations of each pepper sample were obtained. Another chili pepper dataset (Dataset-2) provided by our previous machine vision group (Kalkan et al., 2011) was utilized for validation and comparison purposes. Experiments show that, utilizing halogen light for illumination is superior to UV in terms of classification accuracy.

Novel feature sets based on individual spectral band energies and absolute difference of consecutive spectral band energies were proposed. We observed absolute difference of consecutive spectral band energy features produced better accuracy rate than the individual spectral band energy features for Dataset-1. Another set of quantized histogram matrix features (QHM) were also proposed. The QHM features were extracted from both individual spectral bands and absolute difference of consecutive spectral bands by applying the quantization process. The most discriminative features were constructed with 12 bins quantization. Wavelet features were also extracted for wavelet LDB approach (Kalkan, 2008) for comparison.

Experiments showed that, QHM of absolute difference of consecutive spectral band energy features were superior to wavelet based and QHM of the individual spectral band energy features.

Three different types of classifiers were used for the feature selection and classification purposes. LDA and linear-SVM were employed as simpler and linear classifiers and MLP was utilized as non-linear classifier, respectively. Experimental results reveal that MLP outperforms LDA and linear-SVM in terms of classification accuracy rate. This indicates that the aflatoxin contaminated chili pepper classification is a non-linear problem.

Several non-random weight initialization schemes for MLP were proposed. Comparison tests containing real world datasets reveal that, amongst the weight initialization schemes, V-shape produced better accuracy rates than the other methods including random and well known Nguyen-Widrow weight initialization approaches. Non-random approaches are preferable because they provide consistent and reproducible results.

Feature selection (dimensionality reduction) is an important problem in machine learning domain. A novel feature selection method, hierarchical bottleneck backward elimination (HBBE), was proposed based on saliency metric of MLP connection weights. The performance of the HBBE method that was compared to other well known methods: PCA, Fisher discrimination and mRMR. 83.26% accuracy rate was achieved for Dataset-1 under halogen illumination with the proposed QABSDIF_CSBE features and the proposed HBBE feature selection method. The most frequently selected spectral bands for Dataset-1 are 540, 550, 560, 590, 640, 650 and 660 nm. The proposed HBBE method was also applied on 10 different real world datasets. It performed favorably to other dimensionality reduction methods. Since only 14 features were used in the case of Dataset-2, feature selection was

performed via exhaustive search. Thus, we achieved 87.5% classification accuracy rate for Dataset-2 under the UV radiation. 400 and 420 nm spectral bands were selected as the most discriminative bands. In order to compare Dataset-1 and Dataset-2, we equalized the number of the spectral bands by limiting the analysis to 400 to 510 nm and used only UV excitation. Experiments showed that, 410, 430, 450, 490 and 510 bands are chosen jointly in both datasets. However, accuracy rates of Dataset-2 were higher than Dataset-1. This could be due to better transmittance of the optical filters used for Dataset-2 acquisition process.

The proposed feature extraction and selection techniques achieved high classification rates with reduced number of spectral bands. Hence, it will be possible to construct a simple machine vision system for aflatoxin detection in chili pepper using these spectral bands only.

9.1 Future Work

In this thesis, aflatoxin detection using the visible band hyperspectral images of chili pepper samples under UV and halogen illuminations is considered. Wider spectral region containing IR bands may also be investigated. Sunlight, fluorescent, tungsten, neon and laser are other illumination alternatives. In this thesis, we only utilized reflectance; transmittance phenomenon may also be investigated.

In this dissertation, we employed absolute difference of consecutive spectral bands. Absolute difference of other spectral bands may also give plausible results. Robustness of the proposed feature extraction methods and feature selection methods can be verified on a larger chili pepper dataset as well.

As a nonlinear classifier, we used only MLP. Other non-linear classifiers such as SVM with non-linear kernels or quadratic LDA may result in performance improvement.

Proposed feature extraction and selection methods will be the foundation of a real time detection and inspection machine vision system for non destructive testing. We intend to investigate online machine learning techniques and mixture of experts for the decision fusion of classifiers. Exhaustive search or neural network algorithms can be realized by parallelization and implementation on a GPU which may accelerate the search process.

The proposed feature extraction and selection techniques can be applied to other domains including satellite imaging, medical applications, nondestructive testing images and high dimensional datasets.

REFERENCES

Alpaydın, E. 2004. *Introduction to machine learning,* Cambridge, Massachusetts London, England, MIT Press.

Amaldi, E. and Kann, V. 1998. On the approximation of minimizing non zero variables or unsatisfied relations in linear systems. *Theoretical Computer Science,* 209**,** 237-260.

Antkowiak, M. 2006. *Artificial Neural Networks vs. Support Vector Machines for Skin Diseases Recognition.* Master Thesis, Ume°a University.

Ataş, M., Temizel, A. and Yardımcı, Y. 2010a. Using Hyperspectral Imaging and Artificial Neural Network for Classification of Aflatoxin Contaminated Chili Pepper. *IEEE 18th Signal Processing, Communication and Applications Conference.* Diyarbakır: Dicle Üniversity SIU-2010.

Ataş, M., Yardımcı, Y. and Temizel, A. 2010b. Aflatoksinli Biberlerin Hiperspektral Görüntülerinin Siniflandirilmasi İçin Yeni Yaklaşimlar. *DEÜ Mühendislik Fakültesi Mühendislik Bilimleri Dergisi,* 12**,** 17-33.

Ataş, M., Yardımcı, Y. and Temizel, A. 2011a. Aflatoxin Contaminated Chili Pepper Detection by Hyperspectral Imaging and Machine Learning. *Sensing for Agriculture and Food Quality and Safety III* Proc. SPIE.

Ataş, M., Yardımcı, Y. and Temizel, A. 2011b. A novel approach for aflatoxin detection in chili pepper by machine vision, Unpublished Work

Baum, E. B. and Haussler, D. 1989. What size net gives valid generalization? *Neural Computation,* 1**,** 151-160.

Bekkerman, R., El-Yaniv, R., Tishby, N. and Winter, Y. 2003. Distributional word clusters vs. words for text categorization. *Journal of Machine Learning Research,* 3**,** 1183-1208.

Beriat, P. 2009. *Non-Destructive Testing Of Textured Foods by Machine Vision.* Master Thesis, Middle East Technical University.

Berry, M. J. A. and Linoff, G. 1997. *Data Mining Techniques,* NY, John Wiley & Sons.

Bishop, C. M. 1995. *Neural Networks for Pattern Recognition,* Birmingham, UK, clarendon press, OXFORD.

Blum, A. 1992. *Neural Networks in C++,* NY, Wiley.

Bochereau, L., Bourgine, P., Palagos, B. 1992. A method for prediction by combining data analysis and neural networks: application to prediction of apple quality using near infra-red spectra. *Journal of Agricultural Engineering Research,* 51**,** 207–216.

Boger, Z. and Guterman, H. 1997. Knowledge extraction from artificial neural network models. *IEEE Systems, Man, and Cybernetics Conference.* Orlando, FL, USA.

Boser, B., Guyon, I. and Vapnik, V. N. 1992. A training algorithm for optimal margin classifiers. *Fifth annual workshop on computational learning theory.* San Maeo: CA:Morgan Koufmann.

Bothast, R. J. and Hesseltine, C. W. 1975. Bright Greenish-Yellow Fluorescence and Aflatoxin in Agricultural Commodities. *Applied Microbiology,* 30**,** 337-338.

Bourlard, H. and Kamp, Y. 1988. Auto-Association by Multilayer Perceptrons and Singular Value Decomposition. *Biological Cybernetics,* 59**,** 291-294.

Breiman, L. and Spector, P. 1992. Submodel selection and evaluation in regression: The X-random case. *International Statistical Review***,** 291-319.

Buscema, M., Terzi, S. and Tastle, W. 2010. A new meta-classifier. *Fuzzy Information Processing Society (NAFIPS), 2010 Annual Meeting of the North American.* Toronto.

Cheriyadat, A. and L.M.Bruce. Why principal component analysis is not an appropriate feature extraction method for hyperspectral data. Proceedings of the IEEE IGARSS, 2003. 3420-3422.

Cover, T. M. and van Campenhout, J. M. 1977. On the possible ordering in the measurement selection problem. *IEEE Trans. Syst. Man Cybern.,* 7**,** 657-661.

Cravener, T. L. and Roush, W. B. 1999. Improving Neural Network Prediction of Amino.Acid Levels in Feed Ingredients. *Poultry Science* 78**,** 983-991.

Curtis, P. D., Iezekiel, S. and Miles, R. E. Using Simulated Annealing for Topological Optimisation of All-Optical Microwave Filters. URSI, 2002 Masstritch.

Dash, M. and Liu, H. 2011. *Feature Selection for Classification* [Online]. Available: http://xin.cz3.nus.edu.sg/group/personal/cx/modules/DM/FeatureSelection.ppt [Accessed 01.09.2011.

Doster, M. A., Michailides, T. J. and Morgan, D. P. 1996. Aspergillus Species and Mycotoxins in Figs from California Orchards. *Plant Disease,* 80**,** 484-489.

Duda, R. O., Hart, P. E. and Stork, D. G. 2001. *Pattern Classification*, John Wiley

E. Erzin, A. E. C., and Y. Yardimci. Subband analysis for robust speech recognition in the presence of car noise. Int. Conf. Acoustics, Speech, and Signal Processing, 1995.

Efron, B. 1983. Estimating the error rate of a prediction rule: improvement on cross-validation. *J. Am. Stat. Assoc.***,** 316–331.

Egmond, H. P. v. and Paulsch, W. H. 1986. Determination of mycotoxins. *Pure & Appl. Chem,* 58**,** 315-326.

ElMasrya, G., Wangb, N. and Vigneault, C. 2009. Detecting chilling injury in Red Delicious apple using hyperspectral imaging and neural networks. *Postharvest Biology and Technology,* 52**,** 1-8.

Environment, H. a. S. O. 2011. *Aflatoxins in Your Food and their Effect on Your Health* [Online]. Available: http://www.ehso.com/ehshome/aflatoxin.php [Accessed 08/29/2011.

European Commission Regulation, E. 2006. Setting maximum levels for certain contaminants in foodstuffs. Official Journal of the European Union.

FAO. 2009. *chillies and peppers* [Online]. FAO. Available: http://faostat.fao.org/site/567/DesktopDefault.aspx?PageID=567#ancor [Accessed 07.09.2011.

Fersaie, A., McClure, W. F. and Monroe, R. J. 1978. Development of Indices for Sorting Iranian Pistachio Nuts According to Fluorescence. *Journal of Food Science* 43**,** 1550-1552.

Fisher, R. A. 1936. The Use of Multiple Measurements in Taxonomic Probl. *Annals of Eugenics,* 7**,** 179-188.

Forman, G. 2003. An extensive empirical study of feature selection metrics for text classification. *Journal of Machine Learning Research,* 3**,** 1289-1306.

Frank, A. and Asuncion, A. 2011. *UCI Machine Learning Repository* [Online]. University of California, Irvine, School of Information and Computer Sciences. Available: http://archive.ics.uci.edu/ml/.

Fukunaga, K. and Koontz, W. 1970. Application of Karhunen-Loeve Expansion to Feature Selection and Ordering,. *IEEE Trans. Comput.,* 19**,** 311.

Garson, G. D. 1991. Interpreting Neural Network Connection Weights. *Artificial Intelligence Expert,* 6**,** 47–51.

Gonzales, R. and Woods, R. 2002. *Digital Image Processing*, Prentice Hall.

Gruna, R., Vieth, K.-U., Michelsburg, M. and León, F. P. 2010. Hyperspectral Imaging - From Laboratory to In-line Food Sorting. *CIGR Workshop on Image Analysis in Agriculture.* Budapest.

Guyon, I. and Elisseeff, A. 2003. An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research,* 3**,** 1157-1182.

Guyon, I., Gunn, S., Nikravesh, M. and A.Zadeh, L. 2006. *Feature Extraction Foundations and Applications,* Netherlands, Springer.

Günther, W. and W.S., S. 1982. *Color Science: Concepts and Methods, Quantitative Data and Formulae* New York, Wiley Series in Pure and Applied Optics.

Hall, M. A. 1998. *Correlation-based Feature Selection for Machine Learning.* PhD Thesis, Waikato University.

Harper, P. R. 2005. A review and comparison of classification algorithms for medical decision making. *Health Policy,* 71**,** 315-331.

Hastie, T., Tibshirani, R. and Friedman, J. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* New York, Springer.

Haykin, S. 1999. *Neural Networks a comprehensive foundation,* Hamilton, Ontario, Canada, Prentice Hall.

Herrman, T. 2002. Mycotoxins in Feed Grains and Ingredients, Feed Manufacturing. Kansas State University.

Hirano, S., Okawara, N. and Narazaki, S. 1998. Near Infra Red Detection of Internally Moldy Nuts, Bioscience. *Bioscience, Biotechnology and Biochemistry,* 62**,** 102-107.

Hornik, K., Stinchcombe, M., White, H 1989. Multilayer feedforward networks are universal approximators. *Neural Networks,* 2**,** 359–366.

Hsu, P. H. and Tseng, Y. H. 2002. Feature extraction of hyperspectral data using the Best Wavelet Packet Basis. *IEEE Geoscience and Remote Sensing Symposium*

Hunt, E. B., Martin, J. and Stone, P. J. 1966. *Experiments in Induction,* New York, Academic Press.

IARC 2002. Aflatoxins. In Traditional Herbal Medicines, Some Mycotoxins, Naphthalene and Styrene. *IARC Monographs on the Evaluation of Carcinogenic Risks to Humans,* 82**,** 171-366.

Jashki, M.-A., Makki, M., Bagheri, E. and Ghorbani, A. A. 2009. An Iterative Hybrid Filter-Wrapper Approach to Feature Selection for Document Clustering. *Proceedings of the 22nd Canadian Conference on Artificial Intelligence: Advances in Artificial Intelligence.* Kelowna, Canada: Springer-Verlag.

Jayas, D. S., Paliwal, J., Visen, N.S. 2000. Multi-layer neural networks for image analysis of agricultural products. *Journal of Agricultural Engineering Research* 77**,** 119–128.

Joback, K. G. 1984. *A Unified Approach t o Physical Property Estimation using Multivariate Statistical Techniques.* MS Thesis, Mass. Inst. of Tech.

Jollifie, I. 2002. *Principal Component Analysis*, Springer.

Kaiser, J. F. On a simple algorithm to calculate the energy of a signal,. Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing, 1990 Albuquerque, NM, USA. ICASSP '90, 381–384.

Kalkan, H. 2008. *Feature Extraction from Acoustic and Hyperspectral Data by 2D Local Discriminant Bases Search.* PhD Thesis, Middle East Technical University.

Kalkan, H., Beriat, P., Yardimci, Y. and Pearson, T. C. 2011. Detection of contaminated hazelnuts and ground red chili pepper flakes by multispectral imaging. *Computers and Electronics in Agriculture,* 77**,** 28-34.

Keeni, K., Nakayama, K. and Shimodaira, H. A training scheme for pattern classifcation using multi-layer feed-forward neural networks. In Proceedings of 3rd International Conference on Computational Intelligence and Multimedia Applications, 1999 New Delhi, India. 307-311.

Kim, J., Mowat, A., Poole, P., Kasabov, N. 2000. Linear and non-linear pattern recognition models for classification of fruit from visible–near infrared spectra. *Chemometrics and Intelligent Laboratory Systems,* 51**,** 201–216.

Kira, K. and Rendell, L. A. A practical approach to feature selection. Machine Learning: Proceedings of International Conference (ICML'92), 1992. 249-256.

Kiran, R., S, S. and Atmosukarto, I. 2004. *Spam or Not Spam That is the question* [Online]. Available: http://www.cs.washington.edu/education/courses/cse573/04au/Project/mini2/Ravi&Indri/spamfilter_ravi_indri.pdf.

Kirkpatrick, S., Gelatt, C. D. and Vecchi, M. P. 1983. Optimization by Simulated Annealing. *Science,* 220**,** 671-680.

Kithu, C. J. 2001. Issues on SPS and Environmental Standards for India. *UNCTAD/IDRC Conference.*

Kittler, J. 1978. *Pattern Recognition and Signal Processing,* The Netherlands, Sijhoff and Noordhoff.

Kohavi, R. and John, G. H. 1997. Wrappers for feature subset selection. *Artifical Intelligence,* 97**,** 273-324.

Korycinski, D., Crawford, M., Barnes, J. W. and Ghosh, J. 2003. Adaptive Feature Selection for Hyperspectral Data Analysis using a binary hierarchical classifier and tabu search. *Geoscience and Remote Sensing Symposium.* IEEE International.

Kumar, S., Ghosh, J. and Crawford, M. 2001. Best-Bases Feature Extraction Algorithms for Classification of Hyperspectral Data. *IEEE Trans Geoscience and Remote Sensing,* 39**,** 1368-1379.

LeCun, Y. 1989. Generalization and network design strategies. Depertment of Computer Science.

LeCun, Y. 1993. Efficient learning and second order methods. *A tutorial at NIPS 93.* Denver.

Lee, C. and Landgrebe, D. 1993. Analyzing High-dimensional Multispectral Data. *IEEE Trans Geoscience and Remote Sensing,* 31**,** 792-800.

Liang, P. and Abernethy, J. 2011. *MLComp* [Online]. Available: http://www.mlcomp.org [Accessed 25.10.2011 2011].

Lin, W.-C. 1998. *A case study on support vector machines versus artificial neural networks.* Master Thesis, University of Pittsburgh.

Lu, R. 2007. *Quality evaluation of Fruit by Hyperspectral Imaging.In Computer Vision Technology for Food Quality Evaluation,* New York, Springer.

MacGregor, J. 1989. Multivariate Statistical Methods for Monitoring Large Data Sets from Chemical Processes. *AIChE Meeting.* San Francisco.

Madden, M. G. 2002. Evaluation of the Performance of the Markov Blanket Bayesian Classifier Algorithm. *CoRR*.

Mantaras, R. L. 1989. ID3 revisited: A distance based criterion for attribute selection. *Proceedings of Int. Symp. Methodologies for Intelligent Systems.* Charlotte, North Carolina.

Mark, G. and G., B. B. 2003. *Machine Vision for the Inspection of Natural Products,* London, Springer.

Marona Noelia, S., Betanzos, A. A. and Castillo, E. 2005. A new Wrapper Method for Feature Subset Selection. *ESANN, Europian symposium of ANN.*

Marsh, P. B., Simpson, M. E., Ferretti, R. J., Merola, G. V., Donoso, J., Craig, G. O., Trucksess, M. W. and Work, P. S. 1969. Mechanism of formation of a fluorescence in cotton fiber associated with aflatoxins in the seeds at harvest. *J. Agric. Food Chem.,* 17**,** 468-472.

McCulloch, W. and Pitts, W. 1943. A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics,* 7**,** 115 - 133.

Meyer, P. E. 2008. *Information-Theoretic Variable Selection and Network Inference from Microarray Data.* PhD Thesis, lUniversité Libre de Bruxelles.

Michalak, K. and Snicka, H. 2006. Correlation based feature selection strategy in classificiation problems. *Int. J. Appl. Math. Comput. Sci,* 16**,** 503-511.

Michie, D., Spiegelhalter, D. J. and Taylor, C. C. 1994. *Machine Learning, Neural and Statistical Classification*.

Nguyen, D. and Widrow, B. Improving the learning speed of 2-layer neural Networks by choosing initial values of the adaptive weights. Proceedings of the International Joint Conference on Neural Networks, 1990. 21–26.

Niu, S., Huang, T., Feng, K., Cai, Y. and Li, Y. 2010. Prediction of tyrosine sulfation with mRMR feature selection and analysis. *Journal of Proteome Research,* 9**,** 6490–6497.

Nocedal, J. and Wright, S. J. 1999. *Numerical Optimization,* New York, Springer-Verlag.

Olden, J. D. and Jackson, D. A. 2002. Illuminating the Black Box: a Randomization Approach for Understanding Variable Contributions in Artificial Neural Networks. *Ecological Modeling,* 154**,** 135–150.

Oracle. 2008. *Oracle® Data Mining Concepts* [Online]. Available: http://download.oracle.com/docs/cd/B28359_01/datamine.111/b28129/feature_extr.htm [Accessed 31/08/2011.

Özesmi, S. L. and Özesmi, U. 1999. An artificial neural network approach to spatial habitat modeling with interspecific interaction. *Ecol. Model.,* 116**,** 15-31.

Park, B., Chen, Y.R. 1996. Multispectral image analysis using neural network algorithm. *ASAE,* 96**,** 3034.

Patil, V. T. a. H. A. On the development of variable length teager energy operator (VTEO). Interspeech, 2008 Brisbane, Australia. 1056–1059

Pearson, J. and Lippmann, R. 1988. *Darpa Neural Network Study*, AFCEA International Press.

Pearson, T., Wicklow, D., Maghirang, E., Xie, F. and Dowell, F. 2001. Detecting Aflatoxin in Single Corn Kernels by Using Transmittance and Reflectance Spectroscopy. *Transactions of the ASAE,* 44**,** 1247-1254.

Peng, H., Ding, C. and Long, F. 2005a. Minimum redundancy maximum relevance feature selection. *IEEE Intelligent Systems,* 20**,** 70-71.

Peng, H., Long, F. and Ding, C. 2005b. Feature Selection Based on Mutual Information:Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy. *IEEE Transactions on pattern analysis and machine intelligence,* 27**,** 1226-1238.

Peter, M. L. 2003. *Black Light Test for Aflatoxin Is Questionable Process* [Online]. Available: http://www.ksre.ksu.edu/news/sty/2003/blacklight_test082803.htm [Accessed 30/08/2011.

Pudil, P., Novovicova, J. and J., K. 1994. Floating search methods in feature selection. *Pattern Recogn. Lett.,* 15**,** 1119-1125.

Rajanayaka, C., Kulasiri, D. and Samarasinghe, S. A comparative study of parameter estimation in hydrology modelling: Artificial neural networks and curve fitting approaches. Proceeding of International conference on Modelling and Simulation, 2003.

Rapid Miner, T. 2009. *Neural Net Learner (Rapid Miner Class Documentation)* [Online]. Available: http://rapid-i.com/api/rapidminer-4.4/com/rapidminer/operator/learner/functions/neuralnet/NeuralNetLearner.html [Accessed 15 August 2011.

Refaeilzadeh, P., Tang, L. and Liu, H. 2009. Cross Validation. *In:* Liu, L. Ö., M. Tamer (ed.) *Encyclopedia of Database Systems.* Spring.

Robnik-Sikonja, M. and Kononenko, I. 2003. Theoretical and Empirical Analysis of ReliefF and RReliefF. *Machine Learning Journal,* 53**,** 23-69.

Rumelhart, D. E., Hinton, G. E. and Williams, R. J. 1986. Learning representations by back-propagation errors. *Nature,* 323**,** 533–536.

Russo, A. P. 1991. Constrained neural networks for recognition of passive sonar signals using shape. *IEEE Conference on Neural Network for Ocean Engineering.* Washington DC.

Shimodaira, H. A weight value initialization method for improving learning performance of the back propagation algorithm in neural networks. Proceedings of the International Conference on Tools with Artificial Intelligence, 1994 New Orleans, LA. 672-675.

Somol, P., Novovicová, J. and Pudil, P. 2010. Efficient Feature Subset Selection and Subset Size Optimization. *Pattern Recognition Recent Advances.* INTECH.

Starr, C., Evers, C. and Starr, L. 2005. *Biology: Concepts and Applications*, Thomson Brooks/Cole.

Stephanopoulos, G. N. and Guterman, H. 1989. Pattern Recognition in Fermentation Processes. *ACS Meeting.* Miami Beach, FL.

Su, H. 2010. A comparison of RBF networks and random forest in forecasting ozone day. *International journal of mathematics and computers in simulation,* 4**,** 59-66.

Svaetichin, G. 1956. Spectral response curves from single cones. *Actaphysiol. scand. 39,* 134**,** 17-46.

Talavera, L. 2005. *An Evaluation of Filter and Wrapper Methods for Feature Selection in Categorical Clustering, Lecture Notes in Computer Science,* Heidelberg, Springer.

Taydaş, E. E. and Aşkın, O. 1995. kırmızı biberlerde aflatoxin oluşumu. *Gıda,* 20**,** 3-8.

Teager, H. M. 1980. Some observations on oral air flow during phonation. *IEEE Trans. Acoust., Speech, Signal Processing,* 28**,** 599–601.

Teager, H. M. T. a. S. M. Evidence for nonlinear speech production mechanisms in the vocal tract. NATO Advanced Study Institute on Speech Production and Speech Modeling, 1989 Bonas, France. 241–261.

Thrun, S., Bala, J., Bloedorn, E., Bratko, I., Cestnik, B., Cheng, J., Jong, K. D., Dzeroski, S., Hamann, R., Kaufman, K., Keller, S., Kononenko, I., Kreuziger, J., Michalski, R. S., Mitchell, T., Pachowicz, P., Roger, B., Vafaie, H., Velde, W. V. d., Wenzel, W., Wnek, J. and Zhang, J. 1991. The MONK's Problems: A Performance Comparison of Different Learning Algorithms. Pittsburgh: Computer Science Department,Carnegie Mellon University.

Tyson, T. W. and Clark, R. L. 1974. An Investigation of the Fluorescent Properties of Aflatoxin Infected Pecans. *Transactions of the ASAE,* 17**,** 942-944.

Vapnik, V. N. 1995. *The Nature of Statistical Learning Theory,* New York, Springer-Verlag.

Vergopoulou, S., Galanopoulou, D. and Markaki, P. 2001. Methyl Jasmonate Stimulates Aflatoxin B1 Biosynthesis by Aspergillus parasiticus. *J. Agric. Food Chem.,* 49**,** 3494–3498.

Vijayakumar, S. 2011. *Lecture Notes Dim.Reduction* [Online]. Available: http://www.inf.ed.ac.uk/teaching/courses/mlsc/Notes/Lecture8/MLSC_Lec8.pdf [Accessed 10.09.2011.

Wagacha, J. M. and Muthomi, J. W. 2008. Mycotoxin problem in Africa: Current status, implications to food safety and health and possible management strategies. *International Journal of Food Microbiology,* 124**,** 1-12.

Wise, B. M. and L., R. N. 1989. Upset and Sensor Failure Detection in Multivariate Processes. *AIChE Meeting.* San Francisco.

Witten, I. H. and Frank, E. 2005. *Data Mining Practical Machine Learning Tools and Techniques,* San Francisco, Elsevier, morgan kaufman.

Yam, Y. F. and Chow, T. W. S. 1997. A new method in determining the initial weights of feedforward neural networks. *Neurocomputing,* 16**,** 23-32.

Yang, C.-H., Chuang, L.-Y. and Yang, C.-H. 2009. IG-GA: A Hybrid Filter/Wrapper Method for Feature Selection of Microarray Data. *Journal of Medical and Biological Engineering,* 30**,** 23-28.

Yao, H., Hruska, Z., Brown, R. L. and Cleveland, T. E. Hyperspectral Bright Greenish-Yellow Fluorescence (BGYF) Imaging of Aflatoxin Contaminated Corn Kernels. In Optic East, SPIE conference on non-destructive sensing for food safety, quality and natural resource 2006. Proceedings of SPIE.

Zeringue, H. J. and Shih, B. Y. 1998. Extraction and Separation of the Bright-Greenish-Yellow Fluorescent Material from Aflatoxigenic Aspergillus spp. Infected Cotton Lint by HPLC-UV/FL. *Journal of Agricultural and Food Chemistry* 46**,** 1071-1075.

Zeringue, H. J., Shih, B. Y., Maskos, K. and Grimm, D. 1999. Identification of the bright-greenish-yellow-fluorescence (BGY-F) compound on cotton lint associated with aflatoxin contamination in cottonseed. *Phytochemistry,* 52**,** 1391-1397.

Zhao, J., Wang, G., Wu, Z., Tang, H. and Li, H. 2002. The Study on Technologies for Feature Selection. *Proceedings of the International Machine Learning and Cybernetics.*

Zhou, Z.-H. and Jiang;, Y. 2004. NeC4.5: neural ensemble based C4.5. *Knowledge and Data Engineering,* 16**,** 770 - 773.

Zude, M. 2008. *Optical Monitoring of Fresh and Processed Agricultural Crops,* Boca Raton, FL, CRC Press.

VITA

Musa Ataş was born in Diyarbakır in 1972. He graduated from the Department of Metallurgical and Materials Engineering of METU in 1995. He received his M.S. degree also from METU, Informatics Institute in 2005. He has started his PhD study at the Informatics Institute in 2006. He worked as an instructor at Yüzüncü Yıl University Hakkari Vocational School and taught computer programming languages between 1998 and 2006. Since 2007, he has been working as a software engineer at METU Technopolis. His current research interests are in the field of digital image processing, hyperspectral imaging, machine vision, pattern recognition and machine learning.