Modelling FTIR Spectral Data with Type-I and Type-II Fuzzy Sets for Breast Cancer Grading

By

Shabbar Naqvi, B.E, M.E

Thesis submitted to The University of Nottingham

for the Degree of Doctor of Philosophy

School of Computer Science

The University of Nottingham

Nottingham, United Kingdom

June 2014

Modelling FTIR Spectral Data with Type-I and Type-II Fuzzy Sets for Breast Cancer Grading

Shabbar Naqvi

Submitted for the degree of Doctor of Philosophy

June 2014

Abstract

Breast cancer is one of the most frequently occurring cancers amongst women throughout the world. After the diagnosis of the disease, monitoring its progression is important in predicting the chances of long term survival of patients. The Nottingham Prognostic Index (NPI) is one of the most common indices used to categorise the patients into different groups depending upon the severity of the disease. One of the key factors of this index is cancer grade which is determined by pathologists who examine cell samples under a microscope. This manual method has a higher chance of false classification and may lead to incorrect treatment of patients. There is a need to develop automated methods that employ advanced computational methods to help pathologists in making a decision regarding the classification of breast cancer grade. Fourier transform infra-red spectroscopy (FTIR) is one of the relatively new techniques that has been used for diagnosis of various cancer types with advanced computational methods in the literature. In this thesis we examine the use of advanced fuzzy methods with the FTIR spectral data sets to develop a model prototype that can help clinicians with breast cancer grading.

Initial work is focussed on using the commonly used clustering algorithms k-means and fuzzy c-means with principal component analysis on different cancer spectral data sets to explore the complexities within them.

After that, a novel model based on Type-II fuzzy logic is developed for use on a complex breast cancer FTIR spectral data set that can help clinicians classify breast cancer

grades. The data set used for the purpose consists of multiple cases of each grade. We consider two types of uncertainty, one within the spectra of a single case of a grade (intra-case) and other when comparing it with other cases of same grade (inter-case). Features have been extracted in terms of interval data from various peaks and troughs. The interval data from the features has been used to create Type-I fuzzy sets for each case. After that the Type-I fuzzy sets are combined to create zSlices based General Type-II fuzzy sets for each feature for each grade. The created benchmark fuzzy sets are then used as prototypes for classification of unseen spectral data. Type-I fuzzy sets are created for unseen spectral data and then compared against the benchmark prototype Type-II fuzzy sets for each grade using a similarity measure. The best match based on the calculated similarity scores is assigned as the resultant grade.

The novel model is tested on an independent spectral data set of oral cancer patients. Results indicate that the model was able to successfully construct prototype fuzzy sets for the data set, and provide in-depth information regarding the complexities of the data set as well as helping in classification of the data.

Declaration

The work in this thesis is based on research carried out at the *Intelligent Modelling and Analysis* Research Group, the School of Computer Science, the University of Nottingham, England. No part of this thesis has been submitted elsewhere for any other degree or qualification and it all my own work unless referenced to the contrary in the text.

Acknowledgements

Firstly, I would like to thank my supervisor Prof. Jon Garibaldi for his immense guidance, support and supervision in making this research possible for me. Many thanks to my second supervisor Dr. Simon Miller for his invaluable support and help throughout the course of this PhD.

I would also like to thank Dr. Xiao Ying Wang for providing her data as part of my work. Thanks to Prof. Michael George and Dr. Chris Stapleton from the School of Chemistry, University of Nottingham for helping me with FTIR related problems.

Special thanks to my father, Jawaid Badar Naqvi and my brother Shabbir Naqvi for their encouragement throughout the course of this work. Many thanks to my in laws in Lahore (Pakistan) especially my mother in law Mrs. Nusrat Ayesha for their support and prayers for me. Thanks to my Uncles Dr. Sohail Zahid, Danish Fatmi, my friends Younus Suleman, Tahir Naeem, Dr. Salahuddin and to all my fellows in room B38 and all IMA group members for providing a friendly environment.

A very special thanks to my lovely wife, Sana Asif for standing with me shoulder to shoulder during my stay in the U.K and always becoming a motivating factor for me.

Lastly, I would like to dedicate this thesis to my beloved mother (Late) Prof. Shahana Badar Naqvi (Head of the Department of Philosophy, Government Girls Degree College, Quetta Cantt Pakistan) who died in 2011 and always motivated me towards achieving high standards in education and always wanted me to excel in education like her.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This thesis investigates the use of advanced models created from the complex data with various levels of variabilities and uncertainties for the classification of unseen biological spectral data. Classification of breast cancer grading with Fourier transform infrared spectroscopy (FTIR) based spectral data has been used as a test case. This chapter provides a background of the research, motivations behind the work and the aims and objectives of the work. Later on, an outline of the thesis is also reported.

1.1 Background

Breast Cancer, which has the highest incidence rate in women, is also the most common cancer in the UK. It is estimated that one in eight women in the UK is likely to develop breast cancer during their lives [105]. In the USA, more than one million people are diagnosed with breast cancer every year [98]. After the disease has been diagnosed, monitoring its progress with the passage of time and monitoring the re-occurrence of disease based on the complication of the disease for better prediction of survival of patients is very important [82]. This approximation is generally known as a prognosis which plays a vital role in predicting the survival of patients in the future. In estimating long term survival prognostic Indices have shown good performance [99]. One of the widely used indices is Nottingham Prognostic Index (NPI), which considers tumour diameter, lymph

node status and tumour grade as parameters for prognosis. Of these, grading is the most important parameter and is determined by the Nottingham Grading System (NGS) which is based on the microscopic evaluation of tumour cells by the histopathologist. They observe the morphological variations found in the cells considering form and shape of the cells [28, 92]. The breast cancer grades have been classified as either grade-1 (G-I), grade-2 (G-II) or grade-3 (G-III). G-I patients have more chance of long term survival where as G-III is the most severe and long term prognosis of such patients is poor. This microscopic evaluation is dependent upon the observer's decision to categorise the cancer sample and different experts may disagree on complex cases when it is difficult to predict the grade. This manual method involves the chance of incorrect diagnosis of grade which may result in variable prognosis and sub-optimal treatment [89].

To overcome this critical issue, various efforts have been made in the literature using advanced computational methods for the correct prediction of breast cancer grade but no universally accepted global method is found that classifies the cancer grade [3, 32]. For this research work, we have looked at Fourier Transform Infra-red Spectroscopy (FTIR) in combination with advanced computational methods to describe a model that can help experts in classifying the grade.

1.2 Motivations

We have been motivated by the fact that FTIR is a relatively new technique that has been frequently used in the literature in combination with various machine learning methods for differentiating different cancers [4, 6, 20, 57, 58, 64, 70, 83, 106, 129]. In FTIR, infra-red radiation is passed through a sample and wave lengths of various functional groups involved in the sample along with the intensities at which the sample absorbs radiation are measured. The quantity of absorption by sample depends upon the chemical bonds and molecular structures in the sample. It means that small changes in molecular structure are noted and reflected in the resultant FTIR spectra. That is why researchers have elaborated

the importance of FTIR in cancer histoptahology [11]. The spectra created from FTIR serve as a bench mark for a particular class or sample based on its chemical composition. If samples of three grades of cancer have different chemical characteristics then it is likely that this method may help in differentiating between these grades. If we have any bench mark or characteristics spectra saved in a data base then an unknown sample's spectra can be compared with the three grades bench mark or finger print spectra and its grade can be classified. In comparison with traditional histology techniques, the FTIR has major advantages, some of them are [109].

- It is very sensitive to any molecular changes found in the samples
- For a very high volume of samples, it works much quicker than traditional method
- It has more potential of creating a fully automated measurement and analysis as it detects smaller changes in cellular compositions before any other method

In this work, we have combined FTIR with clustering algorithms as well as advanced methods based on fuzzy logic to create a mechanism that may help in the automation of the breast cancer grading. Principal Component Analysis (PCA) has been used to reduce the dimensions of the data set as it is a commonly used method for this purpose [12, 35, 51, 52, 55, 58, 59, 64, 67, 85, 129]. The motivation behind using clustering is that it is an unsupervised learning method that does not require any priori knowledge about the structure of data and it clusters the data according to the number of groups required given as an input. We have looked at two of the commonly used clustering algorithms of k-means and fuzzy c-means (FCM) clustering algorithm. Fuzzy logic has been found to be useful in real world applications with a high level of complexity involved, for example differentiating between various breast cancer cells [5,33,46,87]. It is likely that for a complicated problem like breast cancer grading where a high level of uncertainty is involved, fuzzy logic has the potential to deal with such complicated uncertainties. We have initially used three different data sets with each data set increasing a level of complexity. The first two data sets have been only used with clustering algorithms. For the third data set, which

is the most complex one, besides clustering, we develop a fuzzy inferencing system (FIS) to investigate whether it can be helpful or not for breast cancer grade categorisation. Hill Climbing (HC) and Simulated Annealing (SA) algorithms have been used for tuning of different parameters of FIS. We have also created Type-II (T-II) fuzzy sets from spectral data by using interval data extracted from various spectral features to explore the use of T-II fuzzy sets for spectral data set as creation of such sets based on spectral data is an under explored area of research.

1.3 Aims and Objectives

The aim of the current research is to develop an automated method based on advanced machine intelligence computational methods to classify the grade of breast cancer. We aim to achieve the the following objectives.

1. Apply necessary pre-processing techniques to remove abnormalities from FTIR spectral data in order to use a standard data set

2. Compare various advanced machine learning methods and try to identify a method best suited for classification of breast cancer grading

3. Find methods to identify key features found across various regions of the spectral data set

4. Use these key features to make a prototype model for classification of breast cancer grading with advanced mathematical methods suitable for complex data sets with a high level of uncertainty

5. Evaluate the performance of the created model prototypes on different spectral data sets in order to create a general frame work

1.4 Thesis Layout

Chapter 2 includes a detailed literature review. The review starts with introduction to breast cancer, its prognosis and factors important in prognosis. We describe the universally accepted Nottingham Prognostic Index (NPI) and its parameters. We also describe the Nottingham Grading System (NGS) used frequently around the world for breast cancer grading and also the difficulties involved in it. FTIR, its pre-processing and a discussion on selecting specific spectral regions instead of using the whole spectral range is also part of this chapter. Next, we describe Principal Component Analysis (PCA) as a common method for reduction of dimensionality of data sets and two commonly used clustering algorithms k-means and fuzzy c-means clustering (FCM) with their merits and drawbacks. This chapter also includes a description of Type-I (T-I) and Type-II (T-II) fuzzy sets and a brief review of their applications for biological data sets. Finally, various similarity measures used to distinguish between fuzzy sets have also been described.

In Chapter 3, three data sets are used with PCA in combination with k-means and FCM clustering algorithms to classify different grades. The first data set is created by combining two cases of oral cancer patients used separately in another research. It consists of 33 patients and the aim is to distinguish cancer cells from stroma cells. The second data set is a real breast cancer spectra data set obtained with the help of Nottingham City Hospital and the School of Chemistry at the University of Nottingham. It consists of one case of breast cancer for each of the three cancer grades. The standard algorithms of k-means and FCM with PCA have been used to classify the three grades and their results are compared. Third data set has been obtained from the University of Illinois at Urbana Champaign, USA. It consists of 40 cases of cancer. It is a complex data set and initially k-means and FCM clustering algorithms with PCA have been used to distinguish between the three grades. This chapter focuses on aim numbers 1 and 2 of this research.

In Chapter 4, a FIS is created to classify breast cancer grades for the complex third data set. PCA has been used and first 3 Principal Components (PCs) have been selected as an input to the FIS. Three membership functions have been defined for each input

PC. Membership functions have been trained with the help of Hill Climbing (HC) and Simulated Annealing (SA) methods. Various approaches have been used and their results have been compared. This chapter also focuses on aim number 2 of this research.

Chapter 5 introduces the concept of extracting features from various regions of spectra data instead of using the whole spectral region under investigation. A step wise model has been created with Type-II (T-II) fuzzy sets for grade prediction. Five features based on peak heights and troughs have been selected from 3 regions. These features have been used to create a zSlices based Type-II fuzzy set (zGT-II) that includes variabilities of all the cases within a grade. These features have been used as benchmark prototype and unseen data in the form of Type-I (T-I) fuzzy sets has been compared with them. A weighted similarity measure has been selected and comparison of unseen T-I fuzzy sets with bench mark T-II fuzzy sets for each feature for each grade has been made by this criteria. Similarity scores have been recorded for each feature. Majority vote and summation of similarity methods have been used to classify the unseen grade. A detailed grade profile based on similarity scores has also been created that reflects upon the complexity of the data set. The results have also been compared with the standard clustering algorithms of k-means and FCM clustering algorithm. An alternative configuration has also been used to further investigate grade profiles for three grades. This chapter focuses on aim number 3 and 4 of this work.

Chapter 6 uses a new data set to evaluate the zGT-II fuzzy sets based model created in Chapter 5 and creates profiles for classification of unseen data. The data set is a FTIR data set for Oral cancer patients previously used in another research for differentiate between cancer and stroma cells. We use the same data set to create our zGT-II fuzzy sets based model. A selection of five features has been made for the data set. These features have been used to create the prototype model. The prototype model is then tested against the unseen data and profiles for classification have been created. This chapter focuses on aim number 5 of the work carried out.

Chapter 7 provides a conclusion of the work with a brief summary. It also includes

contribution to the knowledge and directions towards the future work. In the end, a list of published papers and papers in submission has also been provided coming out of the research carried out during this work.

Chapter 2

Literature Review

This chapter includes a detailed literature review carried out during the research. It starts with breast cancer, its diagnosis and prognosis followed by introduction to the Nottingham Prognostic Index (NPI) with emphasis on its grade parameter. Then we discuss the difficulties involved in grading, and the potential of FTIR for this purpose. The rest of the chapter provides a literature review on commonly used clustering algorithms, Type-I and Type-II fuzzy logic. Finally, a brief review of similarity measures for fuzzy sets is given with a focus on Type-II fuzzy sets.

2.1 Breast Cancer

Cancer is a disease that can be characterized by uncontrolled growth and spread of abnormal cells. The uncontrolled spread can result in a patient's death [98]. Breast Cancer is a cancer that forms in tissues of the breast commonly in the ducts (tubes that carry milk to the nipple) and lobules (glands that make milk) that can be seen in Figure 2.1.

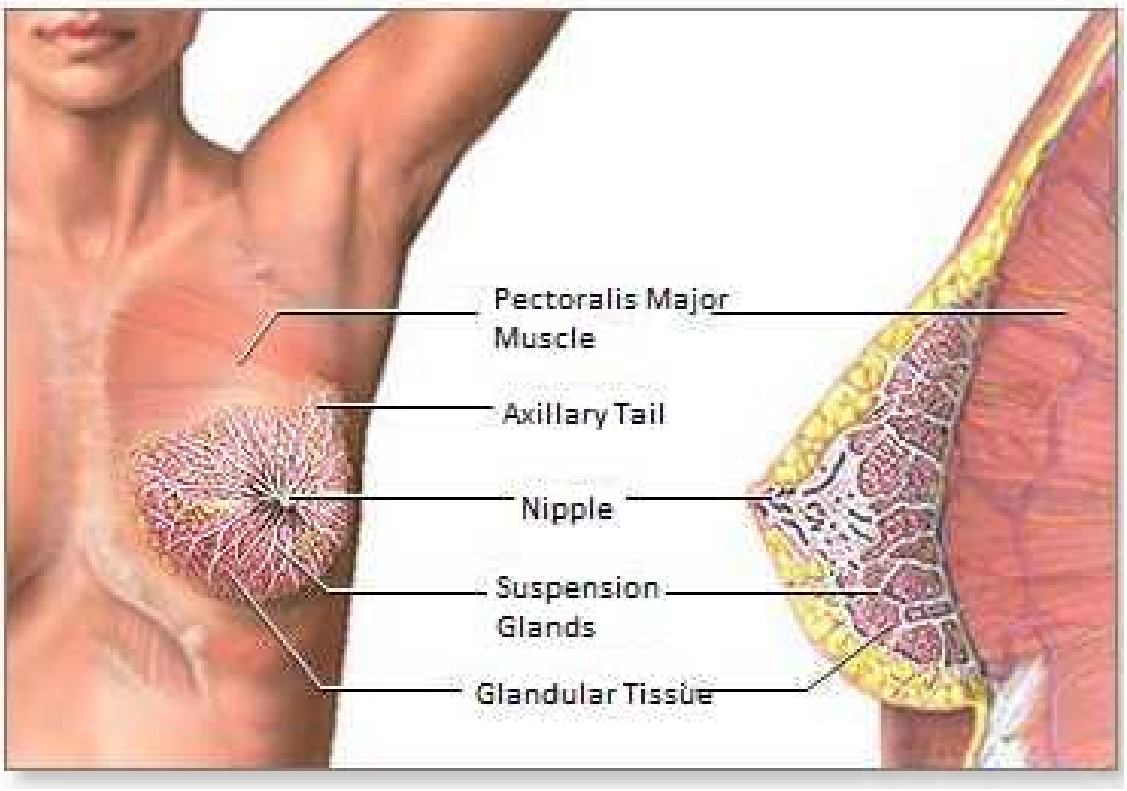


Figure 2.1: Female breast parts (taken from [1])

It can occur in both men and women, although male breast cancer is rare. Breast cancer is either non invasive (referred to as in situ, confined to the site of origin) or invasive (spreading). Metastasis is the term used to describe a phenomenon where cancer cells start to break away from their primary location and travel via the blood stream or lymphatic system. Breast cancer diagnosis can be achieved by imaging methodologies, such as X-ray, Mammography and Ultrasound. The result of mammography is a mammogram that includes additional x-ray views of areas of interest found by the physical examination of the patient to provide more information about the size and characteristics of the abnormality found in the cells. Biopsy is another technique that is used for cancer diagnosis. A biopsy involves the removal of a piece of tissue from the patient's suspected cancerous region and observing it under the microscope to assess the presence of cancerous cells. These techniques identify areas of tumour growth in the breast based on the identification of density changes within the tissue. These methods are not considered very reliable in

complex cases [28]. Diagnosis of breast cancer is also possible by examining the lymph nodes in the ipsilateral axoilla. Lymph nodes are oval shape organs and are part of human immune system. They become enlarged or inflamed in case of cancer. Figure 2.2 shows the position of lymph nodes in the female breast. Once the disease has been diagnosed, monitoring its progress with time is critical as it affects the medication and likelihood survival for a patient.

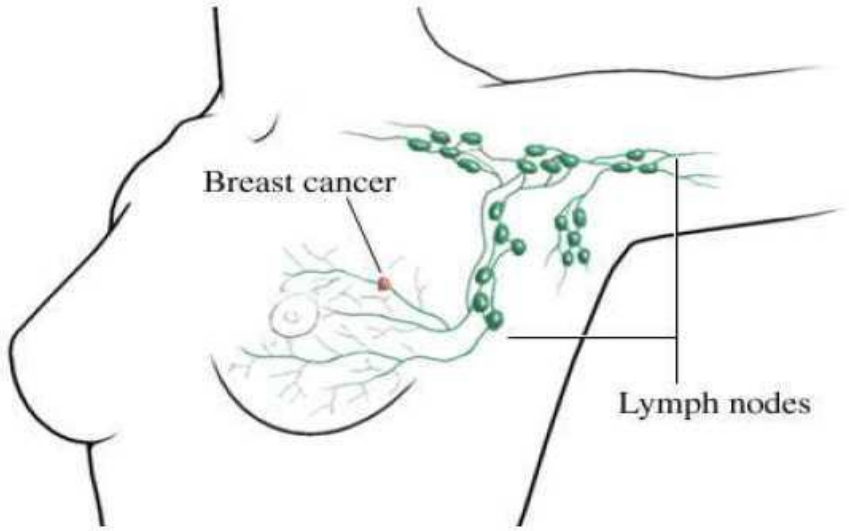


Figure 2.2: Location of lymph nodes in the breast (taken from [109])

2.2 Cancer Prognosis

Medical prognosis is a field in Medicine which deals with assessing the chances of the re-occurrence of the disease based on the complication of the disease for better prediction of survival chances of patients [82]. Prognostic factors in breast Cancer can be separated into two categories called Chronological and Biological. Chronological category is based on the amount of time the tumour has been present where as the Biological category is

based on the potential behaviour of the tumour [15]. The factors commonly in use today for prognosis of breast cancer include lymph node status, tumour size and histological grade [26]. Lymph node status is a time dependent factor and when the number of nodes involved increases, it results in poor prognosis. Tumour size is also a time dependent factor and a small tumour results in longer term survival of the patients and vice versa. Histological grade is a biological factor and is strongly correlated with long term survival. Patients with G-III tumours have the least chance of survival and patients with G-I tumours have the highest chance of survival [26, 93].

2.3 Nottingham Prognostic Index

In addition to the prognostic factors mentioned above, indices have also been developed for the prognosis of breast cancer. In the 1970's and 1980's, a team of clinicians from the Nottingham City Hospital developed a prognostic index for breast cancer based on tumour diameter, lymph node status and tumour grade [41]. It was subsequently validated and called the Nottingham Prognostic Index (NPI) [28]. It is calculated as:

$$NPI = (0.2 * tumour\,diameter\,in\,cm) + lymph\,node\,stage + tumour\,grade \qquad (2.1)$$

The possible values of $lymph\,node\,stage$ are:

1: no nodes affected

2: up to 3 nodes are affected

3: more than three nodes are affected

The possible values of $tumour\,grade$ are:

G-I: less aggressive appearance of tumour

G-II: intermediate appearance of tumour

G-III: more aggressive appearance of tumour

The generally accepted interpretation of the values is: Good (NPI $<$3.4), Intermediate (3.4 $<=$ NPI $<=$ 5.4) and Poor (NPI $>$5.4). The higher the value of NPI is, the lower

the chances of survival of patients are. The NPI has been recognized as the only properly (externally) validated prognostic index for breast cancer [92]. Figure 2.3 describes the link between different parameters of NPI and possible values.

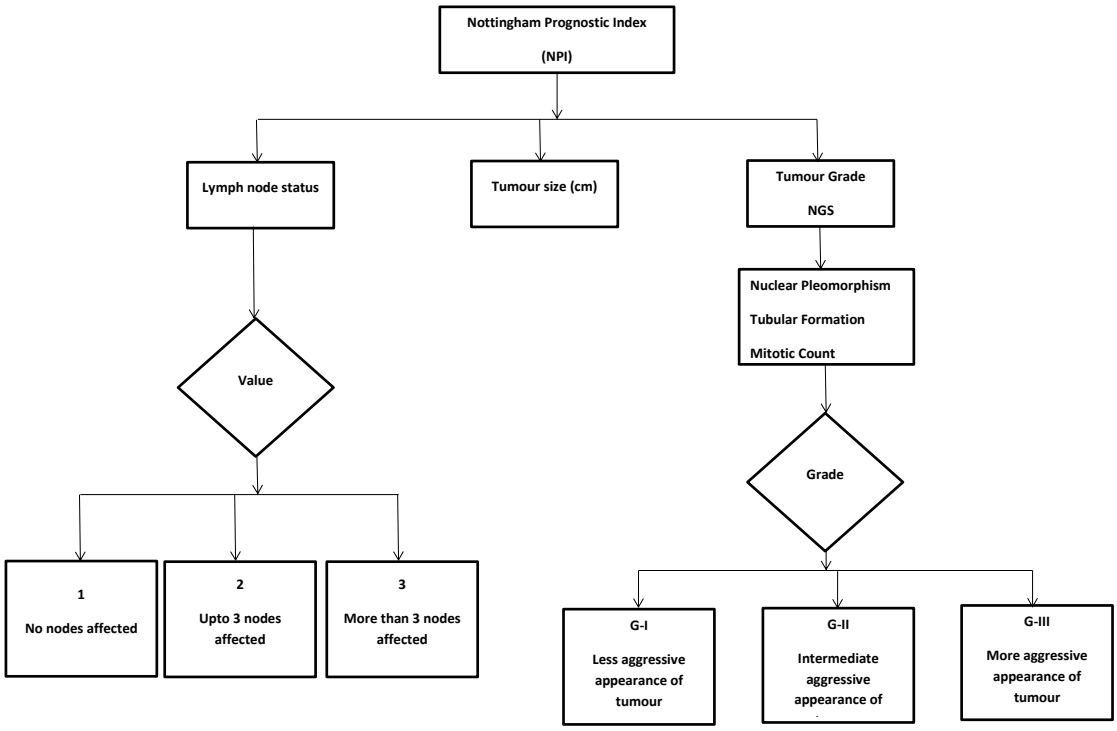


Figure 2.3: NPI and its parameters

2.4 Nottingham Grading System

The Nottingham grading system (NGS) is widely used around the world for grading breast cancer tumours and is also a component of the NPI [80, 92]. It is based on Mitotic Count, Tubule Formation and Nuclear pleomorphism medical parameters.

Tubule Formation parameter relates to the percentage of a tumour in the normal duct structures. A higher percentage is given fewer points and a lower percentage is given more points. A break down of points can be seen in Table 2.1.

The Mitotic Count parameter indicates how many mitotic figures (dividing cells) a histopathologist sees in 10 microscope fields. In cancers, cells divide uncontrollably. The

higher the number of mitotic count is, the more severe the cancer is. It is essential to calibrate the microscope before a mitotic count. A minimum of 10 hpf (high power field) should be counted at the periphery of the lesion and for an attempt to be made to seek out mitoses. Generally accepted values are given in Table 2.2.

Nuclear pleomorphism is a parameter used to find distinction between cell nuclei of a normal breast duct epethelial cells and larger, darker irregular (pleomorphoic) cells. In cancer, changes in genes and chromosomes in the nuclei and pleomorphic changes are considered signs of abnormal cell growth. Generally accepted values are given in Table 2.3.

If TF is the Tubule formation, MC is the Mitotic count and NP is the Nuclear pleomorphism then overall grade can be found with the help of the following equation.

$$Grade = TF + MC + NP \tag{2.2}$$

Where *Grade* is the overall grade. The grade is categorised either as Grade-I (G-I), Grade-II (G-II) or Grade-III (G-III) as shown in Table 2.4. This overall grade score is then used within the NPI [28]. The higher the grade is, the more severe it is so the lower grade patients have more chance of survival than higher grade patients.

Table 2.1: The values for Tubule Formation parameter of NGS

Criteria	Score
Majority of Tumour (>75%)	1
Moderate Degree (10-75%)	2
Little or None (<10%)	3

Table 2.2: The values for Mitotic Count parameter of NGS

Criteria	Score
0-9 Mitoses /10 hpf	1
10-19 Mitoses /10hpf	2
20 or more Mitoses /10 hpf	3

Table 2.3: Values for Nuclear Pleomorphism parameter of NGS

Criteria	Score
Small regular uniform cells	1
Moderate Nuclear size and variation	2
Marked Nuclear variation	3

Table 2.4: Overall Grade

Grade	Combined Score
Low Grade (I)	3-5
Intermediate Grade (II)	6-7
High Grade (III)	8-9

2.5 Difficulties in Cancer Grading

Breast cancer grading with NGS is performed by the histopathologist by observing the tumour sample with the help of microscope. Such a manual method has many disadvantages.

- Variability amongst different pathologists
- Time consuming

The calculation of the grade by a histopathologist, keeping in mind all the parameters described, is a complex and time consuming process with high risk of human error. Even the best histopathologists have been shown to exhibit variability in their scoring of the grade. Therefore, there is a need to develop automated methods for this complex problem that minimise the risks of false classification of grade. This false classification can affect the prognosis and will result in incorrect treatment of patients [80, 89, 104]. The automated method will also result in reduction of workload, increased consistency and time saving. In the next section, we investigate the use of FTIR as a potential candidate to be used in complex biomedical applications. In the literature, FTIR has been used in combination with various computational techniques to differentiate between healthy and

cancerous cells of breast cancer [6, 9, 14, 30, 54, 109, 109] but few researchers have focussed on applying this technique for breast cancer grading. Anastassopoulou et al, [4] have used FTIR with an unsupervised learning clustering algorithm (Hierarchical clustering algorithm or HCA) and principal component analysis (PCA) to differentiate between breast cancer grades with NGS criteria. HCA is an unsupervised learning method that groups the data in a nested series of clusters. The output of HCA is called dendrogram which represents the similarity level between patterns of the data. A major disadvantage of HCA is that it is not considered suitable and computationally efficient for large data sets as in case of breast cancer grading [44, 45]. The authors were able to find good classification on their data set. Other than spectral data sets, Petushi et al. [89] developed an automated imaging system to classify breast cancer grades. The imaging system used different chemical features of the images to find distinct features for each grade. The authors concluded that the proposed method was able to help reducing intra-observer variability. They also suggested the need to do more experimentation. There is a need to further explore this area with more advanced computational techniques that can be used for large data sets.

2.6 Fourier Transform Infra-red Spectroscopy

Fourier Transform Infra-red Spectroscopy (FTIR) is based on the principle that when an infrared (IR) beam is passed through a sample, the functional groups within the sample absorb the infrared radiation and the rest of the radiation passes through. The resulting spectrum represents the molecular absorption and transmission as shown in Figure 2.4. FTIR creates a molecular fingerprint of a sample, no two unique molecular structures produce the same infrared spectrum [24]. If the characteristic spectrum of a sample under analysis are known (in a fingerprint library), it may be possible to compare each of the obtained spectra to reference spectra within the fingerprint library and find the correct result. Typical FTIR equipment can be seen in Figure 2.5. FTIR has been used to

differentiate cancer cells from normal cells in a previous study with the Breast Cancer Pathology Group, School of Chemistry and School of Computer Science, University of Nottingham with advanced machine learning methods [111, 112]

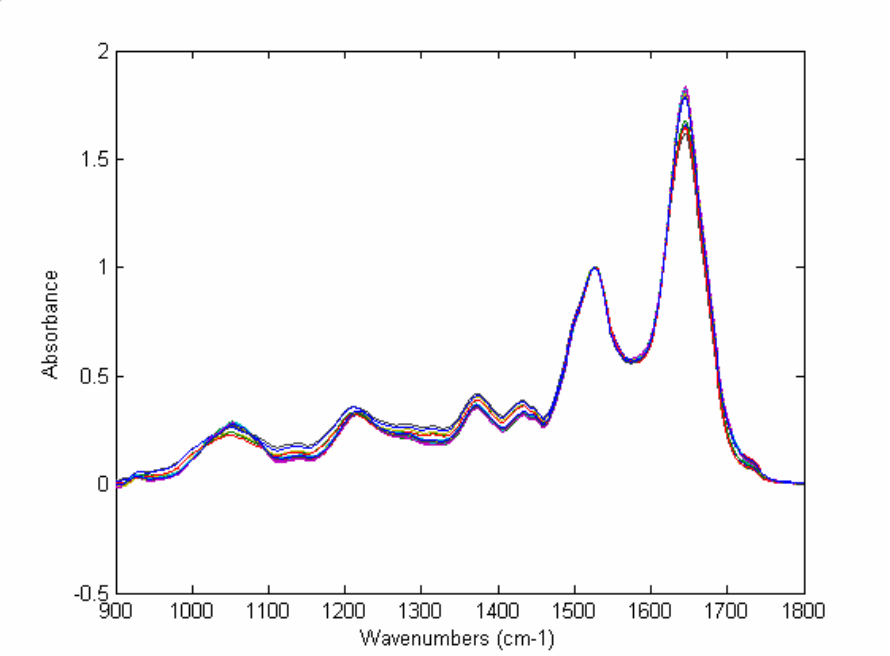


Figure 2.4: An example of FTIR spectra (taken from [109])

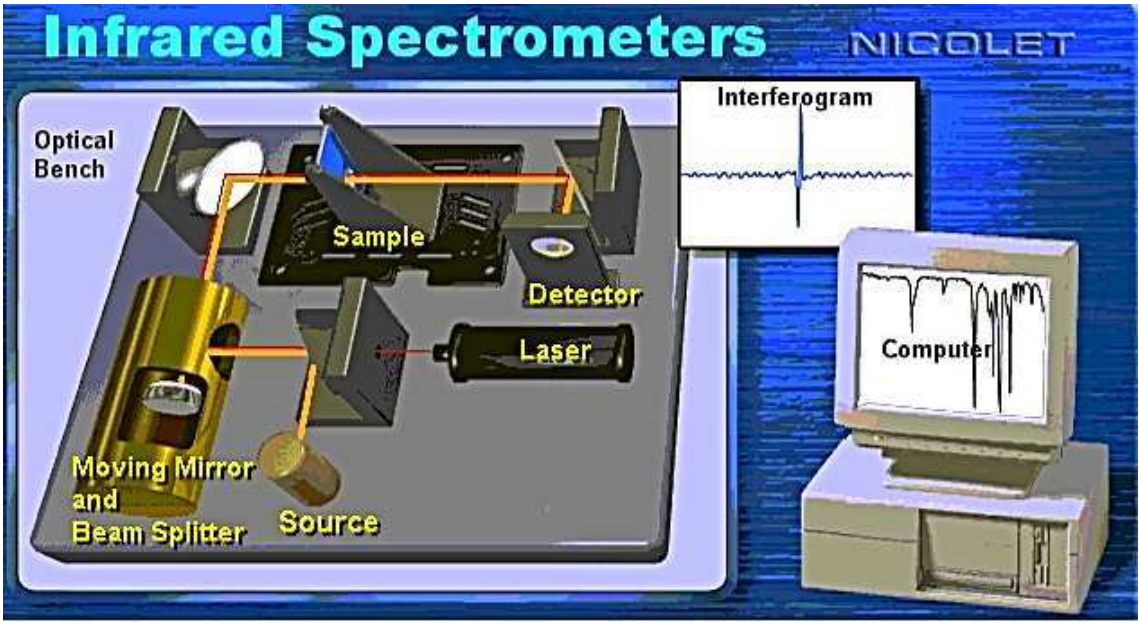


Figure 2.5: A typical FTIR Spectrometer (taken from [2])

Besides FTIR, other spectroscopy techniques also exist. Examples are Raman Spectroscopy and Near Infra-red (NIR) spectroscopy. Raman spectroscopy is used to observe low frequency modes in a system based on interaction of laser light on molecular structures. In NIR, the near infra-red region of electromagnetic spectrum (from 800nm to 2500nm) is used [114]. Each spectroscopic technique has its advantages and disadvantages, and these two have also been used in the literature [17, 18, 57, 96, 121]. We have selected FTIR as it is one of the most common spectroscopic methods found in the literature.

2.7 Spectral Pre-processing

The raw spectra obtained from FTIR are generally not used for analysis as they may contain abnormalities. These abnormalities are removed by standard procedures that include base line correction and normalisation. Base line correction is performed to remove base line abnormalities found in the spectra. These abnormalities are caused by various reasons including carbon dioxide, oxygen, impurities in the air etc . Normalisation is performed to remove the effects produced by varying thickness levels of samples [112]. For our work, we have done basis pre-processing that includes simple base line correction and vector normalisation using standard procedures with the help of School of Chemistry, University of Nottingham. We have not used other methods in order to keep our pre-processing as simple and quick as possible. Advanced pre-processing techniques could potentially be used in future work. Some other methods available for pre-processing not considered for this work are:

1. Savitzky Golay algorithm [6, 30, 85, 116]
2. Self deconvolution and curve fitting [3, 4]
3. Scaling spectra to a particular band (for example, Amide-I or Amide-II) [4, 109]

2.8 Spectral Features Extraction using Specific Regions

Instead of using the whole spectral region, researchers have used various areas within the region as representative features for extraction of key information about the spectra. Researchers are focussing on different feature extraction techniques along with using them for advanced mathematical and computational analysis. A recent critical review of this research can be found in [103]. In Table 2.5 a few examples of the spectral regions used for the analysis of the spectral data have been shown.

Table 2.5: Different spectral regions used for analysis

S.No	Articles	Year	Spectral Features
1	Chiu et al. [20]	2013	950-1350 cm^{-1} 1350-1480 cm^{-1} 1480-1800 cm^{-1}
2	Kumar et al. [58]	2013	1600-1780 cm^{-1}
3	Pallua et al. [85]	2012	3050-3650 cm^{-1} 2800-3000 cm^{-1} 850-1750 cm^{-1}
4	Benard et al. [9]	2010	1000-1800 cm^{-1}
5	Manzano et al [67]	2009	600-1450 cm^{-1} 1500-1750 cm^{-1} 1750-1850 cm^{-1} 2900-3600 cm^{-1}
6	Jusman et al. [50]	2009	950-1800 cm^{-1}
7	Thumanu et al. [102]	2009	900-1800 cm^{-1} 2800-3000 cm^{-1}
8	Anastassopoulou et al. [4]	2009	900-1800 cm^{-1}
9	Wang [109]	2006	900-1800 cm^{-1}

It can be seen from the Table 2.5 that a spectral region considered as a bench mark is around 900-1800 cm^{-1} (S.No 6,7,8,9) as it has been frequently used in the literature. It can also be seen from the table that this region has been further subdivided into smaller regions (S.No 1). All of these smaller regions have different distinct peaks and troughs considered as representative features of the area. We shall be focussing more on the bench mark region throughout this thesis for our experiments instead of using whole spectral region. This will reduce the size of the data set as well as save computation time while keeping majority of the key information intact. We do not assume that this area contains all of the features of the data set and other regions considered by researchers could also be used in a separate investigation.

Now, we discuss various methods used in combination with FTIR data sets for classification various types of spectra.

2.9 Principal Component Analysis

Principal Component Analysis (PCA) is a mathematical procedure that transforms a number of possibly correlated variables into a smaller number of uncorrelated variables called principal components (PCs). It is also called the (discrete) Karhunen-Love transform, or the Hotelling transform depending upon the application area. The first principal component called PC1 contains the maximum variation of data and the subsequent principal components represent reducing variations in values. In this way a set of principal components may contain the maximum features of the data sets and other principal components can be ignored. The PCA is very helpful in the analysis of large data sets as it reduces the data set substantially while keeping most of the important characteristics of the data set intact. [49].

Zwielly et al. [129] implemented FTIR-microscopy to view spectral changes between drug-sensitive and drug-resistant human melanoma cells. PCA was used to reduce the original data of 512 valid measured variables in the spectra to six principal components. These PCs contained 98.4% of the variety in the data. The points represented by the selected PCs were used to distinguish between the two cell types. A t-test was used to find the Significance of the results, which were considered significant with $p < 0.05$. The authors concluded that the best separation was attained using two principal components, PC1 vs PC3 with 91% true identification while addition of PC4 increased the ratio to almost 96%.

Ly et al. [64] reported a study carried out on a selected set of paraffin embedded skin biopsies in order to assess the relevance of FTIR micro-imaging for the differential diagnosis of skin carcinomas. In the suggested model PCA was used and the first eight principal components (describing more than 95% of the total variance) were selected. Different derivative orders and numbers of principal components were tested and the results were presented. The average prediction rate was 74.6% on the spectra from the validation set created by the pathologists. It showed that the data reduction by PCA was not recommended for the precise detection of these types of tumours. Although the first

eight principal components contained the maximum information, they were not precise enough to describe the spectral differences between the eight groups.

Bin et al. [12] have combined PCA with support vector machines (SVM) for the diagnosis of colon cancer on FTIR spectral data. The SVM is a supervised learning algorithm that separates the two classes with a hyperplane in feature space. It is dependant upon parameters like cost factor and the parameters of kernels. 20 PCs were used for the experiments as they contained more than 90% of the data variance. They conducted 500 repeats of the experiments with random initialisations. For the 25 samples used, the combination of PCA and SVM was able to classify 92.6% data correctly. The authors concluded that the results were very good and validated the performance of the proposed method.

Goncalves et al. [35] investigated the use of FTIR Microscopy with PCA to differentiate between different sugarcane bagasse samples. The authors found that using PC1 and PC2, the different types of samples can be detected through visualization and the PCs were able to retain 88% of the variation of the original data.

Lasch et al. [59] employed PCA to generate a coloured image of different types of tissue from the FTIR spectra. Initially, six representative reference spectra from the FTIR maps were used to generate these images. In order to confirm that biochemical information obtained from FTIR analysis was in agreement with conventional light microscopic analysis, PCA was employed again. The Hierarchical clustering algorithm (HCA) (based on Ward's algorithm) was used with six PCs as input. The results showed that IR-based classification agreed with the visual light microscopic investigations.

Kim et al. [52] used PCA as a pre-processing step for cluster analysis using FTIR data from seven different species plants. The first two principal components were plotted and it was shown that the different categories of plants could be identified simply using the visualization. Furthermore, these results based on PCA were used as the input data to the HCA. The results showed that FTIR was successful in reflecting the phylogenetic relationships between the plants.

Kim et al. [51] tried to determine whether multivariate analysis of FTIR spectral data

from leaves and fruit of Fragaria ananassa could be used for rapid discrimination of commercial strawberry cultivators. PCA was used to analyze preprocessed spectral data from each cultivar; Euclidian distances between samples were then calculated. Following this, HCA was used to construct the dendrograms from PCs. This was done by the unweighed pair group method with arithmetic mean analysis using the Euclidean distance as the similarity measure. PCA scores extracted from PCA analysis were used for the calculation of the correlation matrix. Examination of five strawberry cultivars was performed. The first two principal components accounted for 87.58% and 5.8% (93.48% total) of the total variation respectively. A dendrogram based on HCA of the FTIR spectral data was constructed which separated the five cultivars into two major groups. The first PC axis of the score plot did not result in any separation pattern between strawberry cultivars where as the second PC axis of the score plot showed a discrete separation pattern into two groups. The results indicated that PCA was able to display the natural relationship among these samples without prior knowledge along with HCA. The authors concluded that FTIR could be applied as to genetic fingerprinting techniques for investigating the genetic relationship between genotypes of a higher plant species.

Kloss et al. [55] applied FTIR Microscopy to evaluate the biodegradation of polyurethanes of different composition. The infrared spectra profiles obtained for all the samples were very similar. To find minor discrepancies in the spectra the PCA method was applied. Various matrices of different dimensions were created from the collected data within the spectra. Initially a matrix of dimensions of 1739*63 was built with all the 63 spectra. The results indicated that this method was able to identify the differences between these spectra. The samples collected after 12 months of biodegradation in the soil formed a group which was more differentiated from all the others. It was shown that the use of the PCA applied to the infrared spectra was able to determine a better identification of the differences before and after the biodegradation .

Manzano et al. [67] performed a preliminary study on the ageing process of proteinaceous binder materials used in painting under ultra voilet light with FTIR Microscopy and

PCA. Two different methods of covariance data matrices (scaling by mean-centered data) and the correlation data matrices (scaling by unit variance) were used for the PCA. The results of the correlation data matrices were found to be better for this kind of data set. The authors also reported that it was the first study of this kind involving FTIR Microscopy and PCA together for this application.

Pallua et al. [85] applied FTIR for a discrimination study of squamous cell carcinoma. PCA was used for dimension reduction of data set. Scores of first two PCs were plotted against each other. The results showed that these PC plots were able to distinguish cancer cells from stroma cells.

Recently, Kumar et al. [58] applied FTIR on histopathological speimens of breast cancer of different tumour histological grades. The analysis was performed on ECM (extracellular matrix) which separates epithelial cells from the surrounding environment. Normal breast tissue contains layers of epithelial cells which do not perform their functions properly in case of cancer. PCA was used as part of the anlysis and PCA plots were used to show the continuous evolution of the ECM spectra from the tumour towards a position that was far from tumour. Scores of first two PCs were used on the basis that they contained 83.3% and 9.25% of the total data variance respectively. The results indicated there was a significant evolution of distance mainly with PC1 and minor contributions with PC2. Region between 1600-1700 cm^{-1} was found to be containing more information on PC scores. The authors concluded that it was a preliminary study and FTIR could be used for improving the diagnostic of breast cancer in pre-invasive stages.

The litereture review on PCA shows that this is a reliable technique to be used for dimension reduction of large data sets and can be used for breast cancer grading with spectral data sets.

2.10 Clustering

Clustering is defined as a methodology to group a set of unlabelled multidimensional data segments or points such that the members of group called clusters share the most similar attributes. In other words, the members of two different groups will have the most dissimilar attributes. Clustering is a type of unsupervised learning in which no training is performed on data and there are no pre-defined labels reflecting upon the important characteristics of data. In conventional types of clustering algorithm, each member of the data is exclusively assigned to one cluster. In data sets where it is easier to define the cluster boundaries, the results of such method are very good. But in real world applications where it is hard to identify cluster boundaries, this approach does not work well. For this type of complicated data sets, fuzzy clustering may provide more optimal results. In fuzzy clustering each member is a member of every cluster with an associated membership value [44]. For this thesis, we have considered k-means clustering algorithm (k-means) and fuzzy c-means clustering algorithm (FCM). Both k-means and FCM are type of partitional clustering. In partitional clustering a single partition is created by the algorithm. Partitional clustering algorithms have been found to work well with large data sets. We have selected these two algorithms for our study in order to compare the results of a conventional clustering algorithm with a fuzzy based algorithm to find an optimal choice of clustering algorithm. In the next subsections we describe these two algorithms and the key literature associated with them.

2.10.1 K-means Clustering Algorithm

K-means is the most well known centroid algorithm. This means that a cluster in k-means is defined by a cluster centre or a centroid. The algorithm partitions the data set into k-subsets or clusters trying to keep all subsets closest to the same centre. The algorithm first selects a criterion and then executes it with a fixed number of clusters multiple times with different starting values. The result of the algorithm is the best partition found during this

optimisation. A draw back of the algorithm is that it does not find the number of clusters in the data automatically, this number has to provided as an input [44].

The criterion used for k-means clustering is the most commonly used criterion for partitional clustering algorithms and is called the squared error criterion. This criterion has been found to work well with isolated and compact clusters [44, 45]. To understand this, lets take an example of a data set, $X = \{x_1, x_2, ..., x_n\}$ that contains n patterns or elements. Lets assume we need to divide these n patterns to c groups. We consider, $V = \{v_1, v_2, ..., v_n\}$ as the corresponding set of centres and c_j as the number of patterns in cluster j. We also assume that each pattern can only belong to one cluster. Now we define the squared error criterion in Equation 2.3:

$$e^2 = \sum_{j=1}^{c} \sum_{i=1}^{c_j} ||x_i - v_j||^2 \tag{2.3}$$

where e^2 is the squared error, x_{ij} is the i^{th} pattern in the j^{th} cluster, v_j is the j^{th} cluster centre, and $||x_i - v_j||$ is the Euclidean distance between x_i and v_j.

If v_j is the centre of the cluster, then it is calculated as:

$$\begin{aligned} v_j &= \frac{1}{c_i} \sum_{i=1}^{c_j} x_{ij} \\ where \quad & j \in \{1, ..., c\} \end{aligned} \tag{2.4}$$

In Equation 2.4, c_j is the number of patterns in cluster j, x_{ij} is the i^{th} pattern in the j^{th} cluster, and c is the total number of clusters. The working of a typical k-means clustering is shown in Figure 2.6.

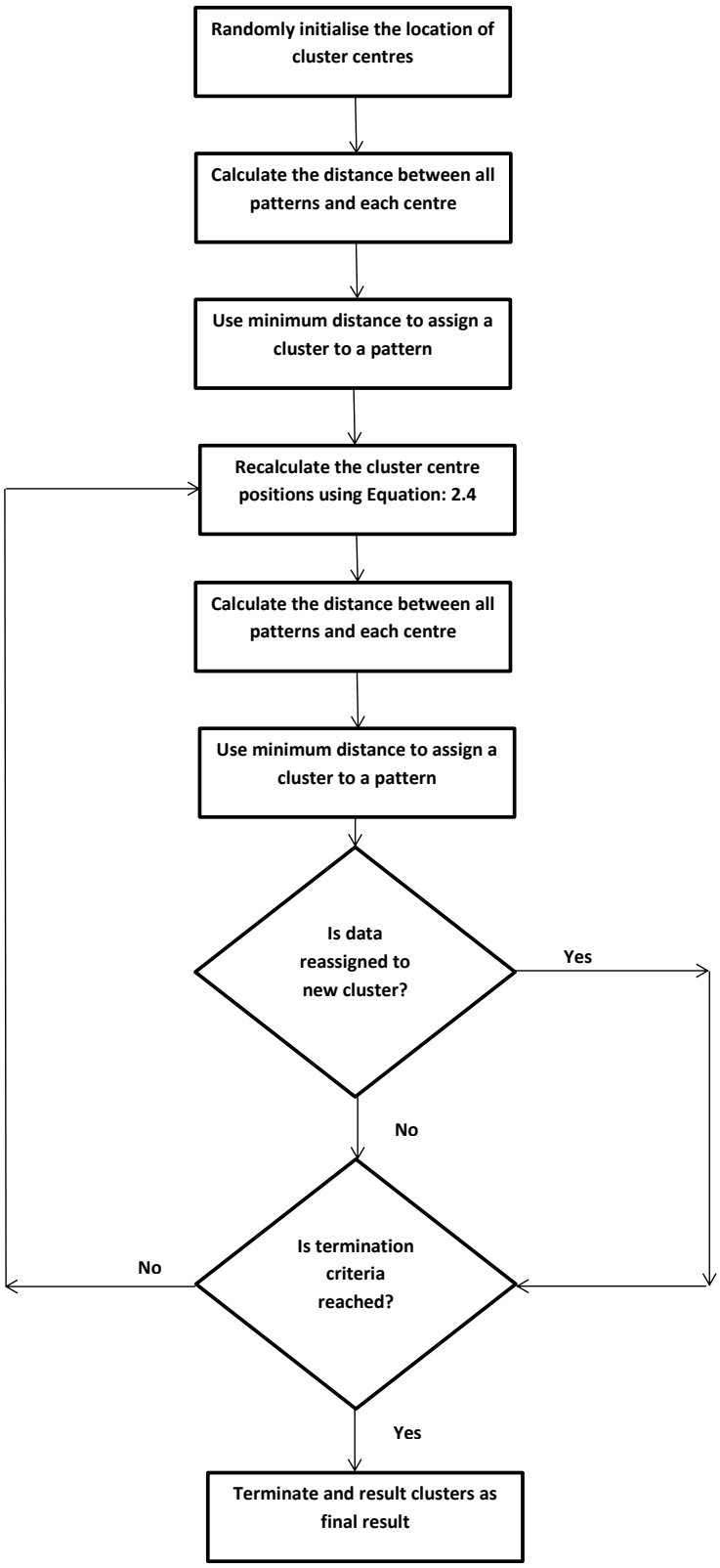


Figure 2.6: K-means clustering algorithm

Wang and Garibaldi [110] applied k-means clustering to axillary lymph node data to distinguish between various types of cells. The IR image created was composed of 7497 spectra that was a significantly higher number than in previous studies. Each spectrum consisted of 821 wave number absorbencies. PCA was used to reduce the dimension of the data set. The first 10 PCs were selected as they contained 99.08% variation of the whole data set. For k-means, the squared Euclidean distance was used as the distance measure; the initial cluster centres positions were randomly selected. The maximum number of iterations was set to 100. The number of clusters were varied from 2 to 10 as different initial positions results in differing performance. The results indicated that k-means was not able to distinguish between normal and cancer tissues with two clusters. However, it was able to split normal and cancerous tissues starting with 6 clusters. In comparison when FCM Clustering algorithm was used on the same data set, it differentiated well between normal and cancerous cells including when the cluster number was set to 2.

Krafft et al. [57] studied Congenital Cystic Adenomatoid Malformations (CCAMs) with FTIR imaging. They claimed it was the first study of this type. Lung tissue specimens obtained from two infant patients were used for the experiments. Data sets were obtained from four tissue sections that allow comparison (i) the biochemical composition of CCAM and normal lung tissue, (ii) the morphology at macroscopic and microscopic resolutions and (iii) the information content of Raman and IR spectroscopy. The pre-processed data sets were subjected to k-means cluster analysis using a Euclidean distance metric. It was used in this study because of its simplicity on large data sets. The result of the cluster analysis was represented by colour coded figures representing the segmentation into four classes or clusters. The result showed that there was overlapping in the clusters and features were not clearly distinguished. It was argued that vibrational spectroscopic imaging could complement the existing diagnostic techniques as data acquisition by FTIR imaging is more rapid and yields spectra with higher signal-to-noise ratios.

Ly et al. [65] applied FTIR spectral imaging on formalin-fixed paraffin-embedded biopsies from colon and skin cancerous lesions. PCA was used on a set of paraffin spectra

to keep the maximum variance in the paraffin dataset while reducing the amount of data modelled. It involved removing of paraffin spectra from the raw images based on outlier detection preceded by a quality test. K-means clustering was performed to highlight tumour tissue within non cancerous tissue. Pseudo-colour images computed by k-means clustering were used to highlight histological structures of interest. K-means maps were calculated several times to make sure a stable solution had found; the percentage of convergence was set to 99.9% in all cases. The retained number of clusters was set to 11, which appeared to match the histology of the epithelial tissues analysed. This methodology was applied on the two samples. The results indicated that tumour areas were successfully demarcated from the rest of the tissue in both colon and skin independently of the embedding material and of the substrate.

Ly et al. [64] carried out a study on a selected set of paraffin embedded skin biopsies in order to assess the relevance of FTIR micro-imaging for the differential diagnosis of skin carcinomas. K-means clustering was applied on pre-processed images to highlight relevant histological structures. For each image, k-means clustering was used to regroup spectra based on similar spectral properties. K-means maps were calculated several times to ensure a stable solution. The percentage of convergence was set to 99.9% and the number of clusters was selected as 11 matching the histology of tissues analysed. A pseudo-colour map was plotted to find the cluster membership information by assigning a colour to each different cluster. Each pseudocolour map was then provided to the collaborating pathologist to correlate the spectral maps and the corresponding H& E stained sections. An almost one-to-one correlation between the IR k-means-clustered images and the histology was found. This showed the potential of the technique for the direct analysis of paraffin-embedded biopsies with FTIR Microscopy.

Pallua et al. [85] used FTIR in combination with k-means as part of study to differentiate between different types of cells in squamous carcinoma. The aim was mainly to distinguish between three types of cells namely cancer, stroma and cornified material. The results were compared with Hematoxylin and eosin stain (HE) images. HE is a method

used to highlight biological structures in biological tissues for better viewing. The results showed that this method was able to distinguish between these three types. Further study was also done by increasing the number of clusters but no other meaningful information was obtained.

2.10.2 Fuzzy c-means Clustering Algorithm

Fuzzy c-means (FCM) clustering algorithm is one of the most popular clustering algorithms found in the literature. Instead of hard clustering algorithms like k-means, the FCM clustering algorithm associates each data point with all the clusters using a membership function. This algorithm was developed by Dunn in 1973 and improved by Bezdek in 1981 introducing a fuzzifier parameter , $1 \leq m < \infty$ [10, 27]. The algorithm is based on the minimisation of the fuzzy objective function as defined in the Equation 2.5:

$$J(U,V) = \sum_{i=1}^{n} \sum_{j=1}^{c} (\mu_{ij})^m ||x_i - v_j||^2 \tag{2.5}$$

As in the case of k-means, we assume that $X = \{x_1, x_2, ..., x_n\}$ is the data that contains n patterns that require dividing into c groups. $V = \{v_1, v_2, ..., v_n\}$ is assumed that the corresponding set of centres are the same as there defined in section 2.10.1. In the Equation 2.5, μ_{ij} is the membership degree of the pattern x_i to the cluster centre v_j. It is compulsory for μ_{ij} to satisfy the following two conditions:

$$\begin{gathered} \mu_{ij} \in \{0,1\} \\ where \quad i \in \{1,...,n\} \quad and \quad j \in \{1,...,c\} \end{gathered} \tag{2.6}$$

$$\sum_{j=1}^{c} = 1 \tag{2.7}$$

The parameter m is used to control the fuzziness of each data point of the set and is called the fuzzifier or fuzziness index. Its higher and lower values make the method more and less fuzzy respectively. There is no theoretical information available to define fuzzifier

but generally value of 2.0 has been widely accepted [10]. $||x_i - v_j||$ is used to represent the Euclidean distance between x_i and v_j. A fuzzy partition defined by $U = (u_{ij})_{n\times c}$ is a matrix that consists of all the membership degrees from every data point to all cluster centres. The fuzzy centres are calculated with the help of the following Equation:

$$v_j = \frac{\sum_{i=1}^{n} (\mu_{ij})^m x_i}{\sum_{i=1}^{n} (\mu_{ij})^m} \qquad \forall j \in \{1,...,c\} \tag{2.8}$$

where v_j represents the fuzzy centres. The fuzzy partition matrix U is updated with the help of the following Equation:

$$\mu_{ij} = \frac{1}{\sum_{k=1}^{c} (\frac{d_y}{d_{ik}})^{\frac{2}{m-1}}} \qquad where \quad d_{ij} = \|x_i - v_j\| \qquad i \in \{1,...,n\} \quad and \quad j \in \{1,....,c\} \tag{2.9}$$

The termination criteria of the algorithm is generally the maximum number of iterations but users can also define any specific criteria according to their requirements. The working of a typical FCM clustering algorithm has been shown in Figure 2.7.

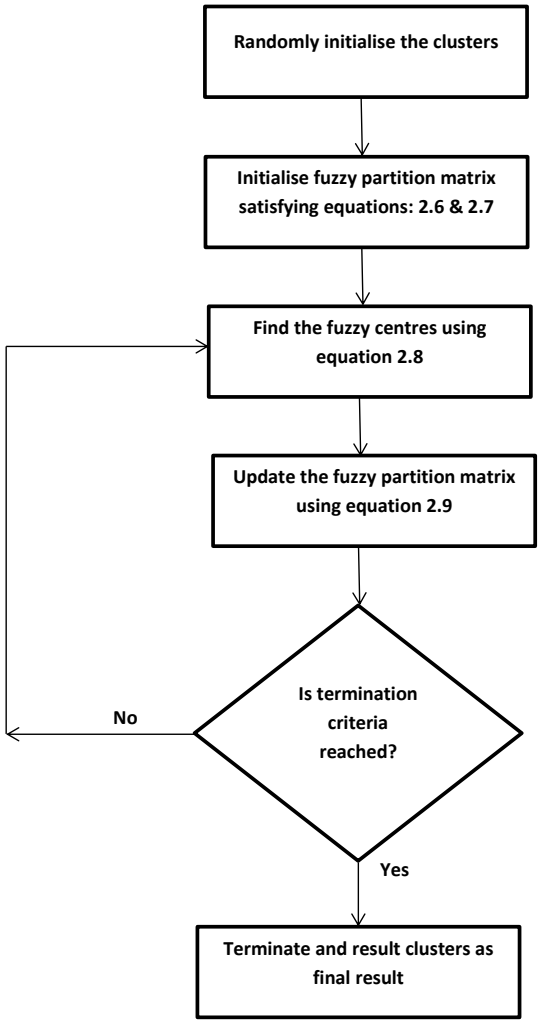


Figure 2.7: FCM clustering algorithm

For all the experiments performed in this thesis with FCM clustering algorithm, after executing the FCM clustering algorithm, each pattern is associated with the cluster to which it has the largest membership degree. This is called hardening of clusters and in the literature it is found to produce better solutions [39, 126].

Lasch et al. [59] implemented a FCM clustering algorithm with FTIR Microscopy on colorectal adenocarcinoma tissue sections. Colour intensities coding was used to find the membership values using PCA. Two dimensional plot was used for the comparison of clustering results with clinical results. The results showed that for upto 6 clusters, FCM images were successfully assigned to a specific tissue. Further increases in number of clusters produced poor results and a weak relationship with the clinical results.

Wang and Garibaldi [110] applied a FCM clustering algorithm to a lymph node tissue section which had been diagnosed with metastatic infiltration (cancer spread from its original location). IR spectra used in experiments were collected from a large area of an axillary lymph node tissue section. The IR image created was composed of 7497 spectra, a significantly higher number than in previous studies. Each spectrum consisted of 821 wave number absorbencies. PCA was used to reduce the dimension of the data set. The first 10 PCs were selected as they contained 99.08% variation of the whole data set. In FCM clustering algorithm, the fuzziness index m was set to a value of 2; the maximum number of iterations was set to 100. The minimal amount of improvement was initially set 10^{-5} (the stopping criterion) of the iteration. Initial results of the FCM clustering algorithm were poor. After setting the minimal amount of improvement of algorithm to 10^{-7}, the results were significantly better. The FCM clustering algorithm produced consistent clustering results including when the number of clusters was set to 2 and it was able to differentiate between normal and cancer cells. The results indicate that the FCM algorithm can separate the major different tissue types using a small number of clusters. As the number of clusters increased upto 9, more information about the tissues was available that could not be recognized by the histopathologists. The performance of the algorithm was also compared with the same experiments using the k-means algorithm which was found to be slightly better.

Steller et al. [101] used FTIR Microscopy with FCM clustering algorithm and HCA to investigate thin sections of cervix uteri encompassing normal tissue, precancerous structures, and squamous cell carcinoma. One hundred and twenty two images of cervical tis-

sue recorded by a FTIR spectrometer were used for the experiments. A two step approach was used combining the advantages of both clustering methods to distinctly enlarge the tissue area to be characterized at once by IR spectroscopic imaging. Initially, FCM clustering was performed so that every tissue type in the tissue sections is represented by at least one FCM cluster. The respective cluster centroids were used as the starting point for subsequent HCA. This procedure reduced the number of potential HCA clusters to less than 100 (a manageable size) for computation. This approach also improved the signal-to-noise ratio, because each FCM cluster centroid was the average over a certain set of IR spectra. The dendrogram resulting from HCA was used to distinguish between different tissue types. Clinical results were used for the verification of the results In the first step, cervical stroma, epithelium, inflammation, blood vessels, and mucus could be distinguished. The authors claimed that it was the first successful attempt in distinguishing basal cells, dysplasia, and tumour in a single sample by IR spectroscopy.

Steiner et al. [100] investigated the influence of hydrophobicity of the substrate surface on structural changes during protein adsorption. FCM clustering algorithm was used for the clustering of FTIR images spectra. The proposed method was demonstrated on an example with two preselected clusters. It was found that when the preselected number of clusters was increased, spectra could be assigned more homogeneously. It was also found that a number of clusters over 5 did not result in significantly lower variance or improved information retrieval. The experiments showed that significant differences were found between hydrophobic and hydrophilic surfaces. The authors concluded that FTIR Microscopy along with a FCM clustering algorithm reliably characterized the thin layers of adsorbed fibrinogen and also importantly spots of structural changes within larger sample areas.

A main draw back of both k-means and the FCM clustering algorithm is that because of their limitations like advance knowledge of number of clusters and convergence to local optima, they can result in sub-optimal solutions that are locally optimal.

2.11 Hill Climbing

Hill Climbing (HC) is a technique which is used for optimisation problems. In HC, a random solution to a problem is found and then the solution is improved (either by climbing down the hill or climbing up the hill) until an optimal solution is reached based on a certain condition or a termination criteria is reached [94]. A draw back of the algorithm is that if a better solution is not available with immediate solutions, the algorithm gets stuck at that point and never goes beyond that point which results in a locally optimal solution rather than globally optimal solution. A typical HC algorithm can be seen in Figure 2.8.

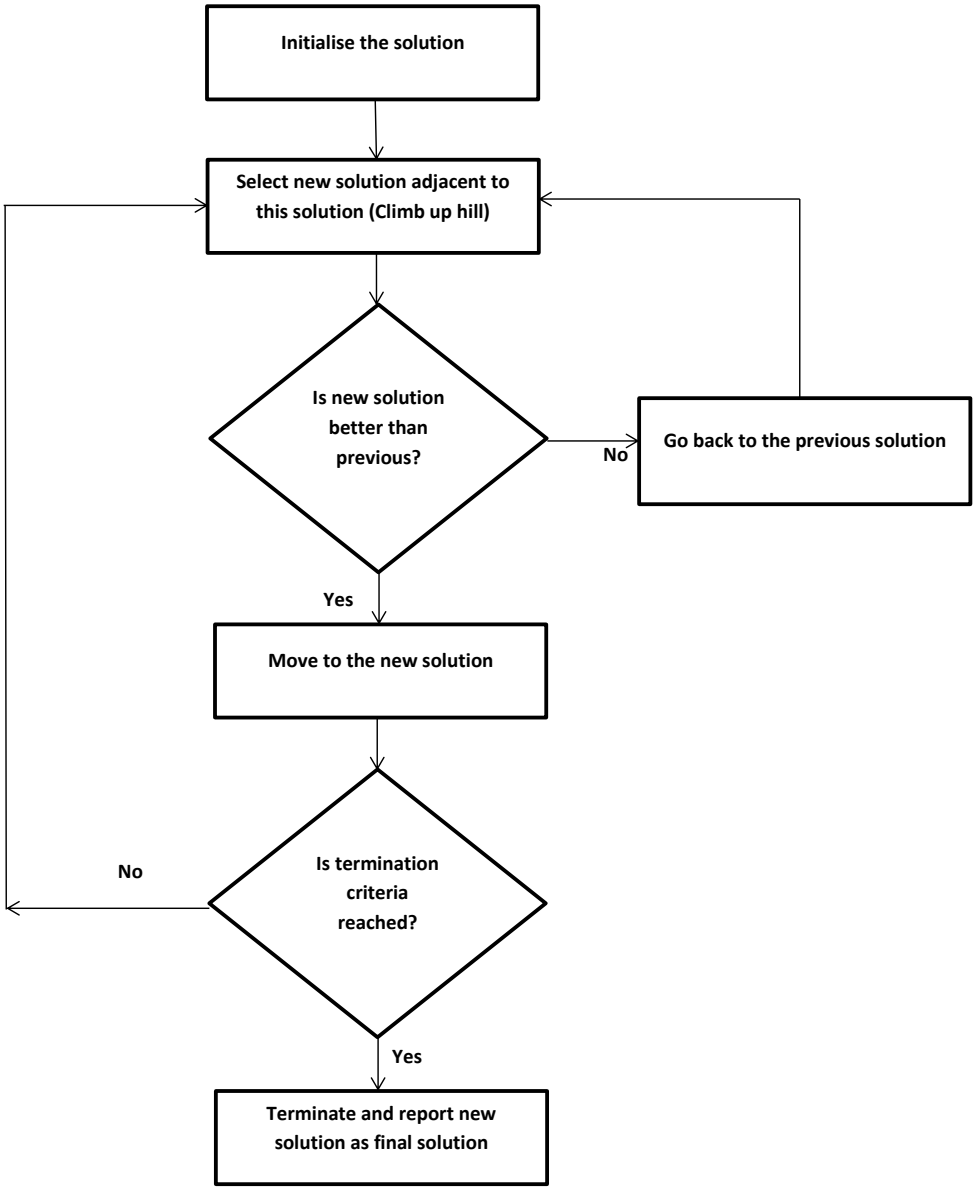


Figure 2.8: Hill Climbing algorithm

The HC algorithm has been used in the literature for problems that involve spectral data sets.

Luts et al. [63] used a HC algorithm for extracting features from Magnetic Resonance Spectral (MRS) data set. Magnetic resonance spectroscopy (MRS), also known as nuclear magnetic resonance (NMR) spectroscopy, is an analytical technique which is used to

study various metabolic changes in brain tumour and other diseases like strokes, seizure disorders etc. It uses the magnetic properties of certain atomic nuclei to determine the physical and chemical properties of atoms or molecules they are part of. The aim of this study was to find the best features suitable for brain tumour classification. HC was used because the high dimensionality of data causes a high computational cost. A subset of features with the best performance was obtained with the help of cross validation. Cross validation is a statistical technique that is used to predict the performance of a model. In this technique, data is divided into various subsets and on a certain criteria, these sets are used for training and testing of the data. A measure for quantifying separation was obtained by stepwise addition or deletion of features over whole spectral region. 13 other methods for the same purpose were also used. The results indicated that this HC method did not perform well and it was excluded.

Schief et al. [95] have used a HC algorithm along with a genetic algorithm (GA) for correction of the shift uncertainty on NMR based spectral data set. GA belongs to a class of evolutionary algorithms which is used to find solutions for optimisation and search algorithms. GA is inspired by the techniques of natural evolution such as inheritance, mutation, selection and crossover [8]. NMR spectra were described by means of a set of overlapping peaks. In order to find the correct target peak, a method of peak picking was required. The authors selected the HC method because of its simplicity over other methods. The results indicated that the combined method was able to find highly effective error estimates for the shift uncertainties in NMR measurements.

HC has also been used in other medical applications. For example, Sasic [97] has used a HC based algorithm in an application to determine the coating thickness of tablets by chiseling and image analysis. The main aim of the HC algorithm was to find the boundaries of the thickness coating on both sides of the tablet. The results indicated that the proposed approach was able to make a quick and inexpensive assessment of the coating thickness of tablets.

As both selected clustering algorithms (k-means and FCM) and the HC algorithm

suffer from locally optimised solutions, we may consider a stochastic search based algorithm like simulated annealing. Stochastic search algorithms may produce a near-optimal solutions quickly. They also avoid converging towards locally optimal solutions [53].

2.12 Simulated Annealing

Simulated Annealing (SA) is a stochastic search technique which has been used for many years for finding global optimum solutions of the problems [53]. It is inspired by the physical process of annealing solids. In this process a solid is heated upto a certain high temperature. After that it is cooled down at a slower pace. The aim is to keep the system in thermodynamic equilibrium at any instant. At any state in equilibrium, the system has a certain energy level. To move from one state to another, there may be several solutions available depending upon the energy levels. The decision to move from one state to another is dependent upon the difference between the energy levels. To understand the scenario within artificial intelligence frame work, we consider E_p as energy of the present state and E_n as energy of the new state. We always move to the new state if $E_n < E_p$. If $E_n \geq E_p$ then we accept the solution and move to the new state with a probability $e^{-(E_n - E_p)/CT}$, where CT is the current temperature. In this way, we might accept a worse solution and avoid the search getting stuck at a local minima. CT is decreased slowly and the process is repeated until the solution does not improve any further or a termination criteria (for example any low temperature) is reached [44]. The workings of a typical SA algorithm are shown in Figure 2.9.

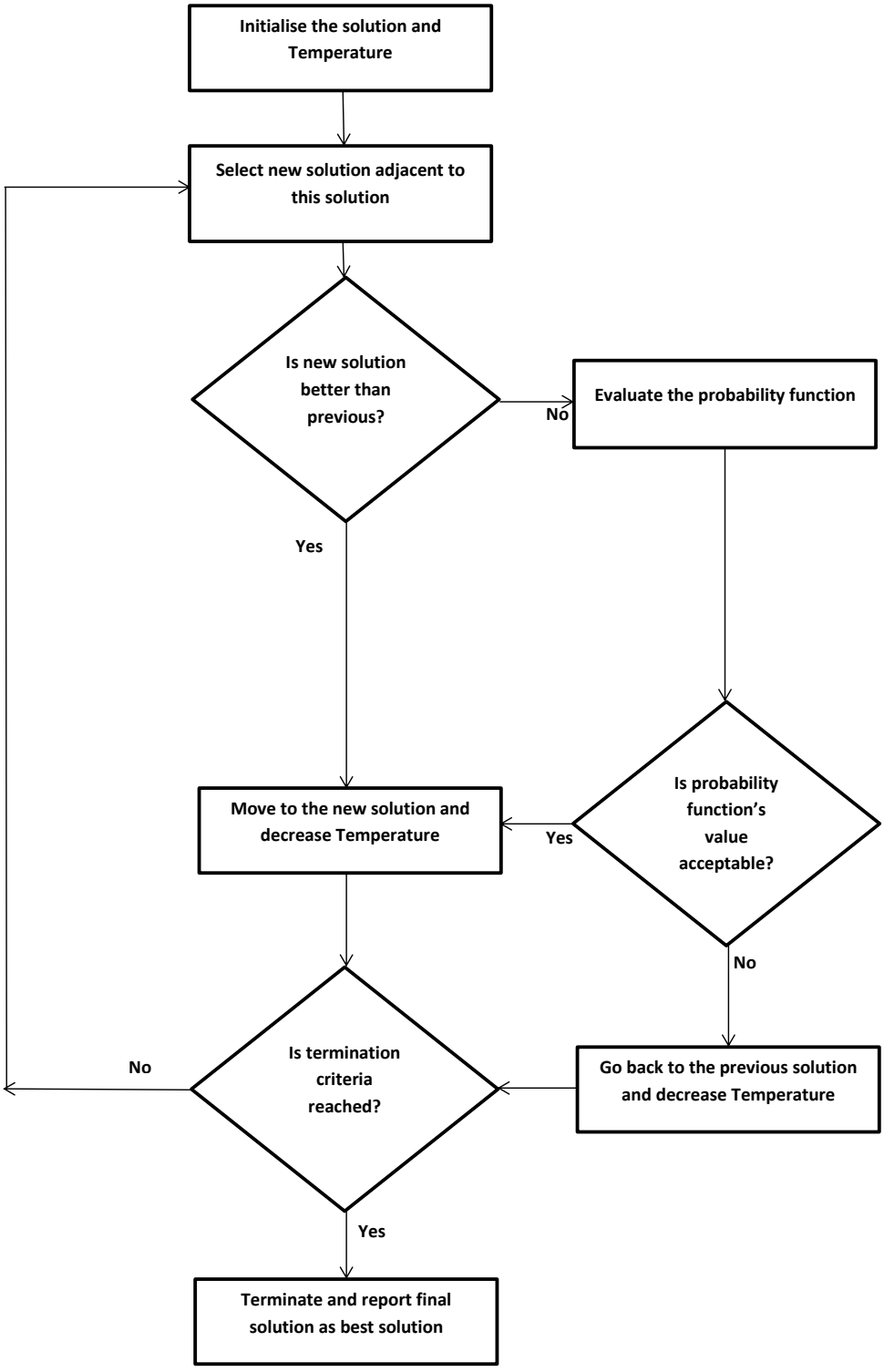


Figure 2.9: Simulated Annealing algorithm

SA has been used in the literature with various spectral data sets. Chen et al. [19] have

used SA to search for the optimal cut off threshold for detecting the quality of glycerol monolaurate (GML) with a pre-processed FTIR spectral data set and with wavelet transform (WT). GML is a popular emulsifier that is found mainly in the food and cosmetics industry. GML contains many impurities and it is essential to identify these impurities. WT is a technology that is used to transform original spectra into a wavelet domain that is represented by wavelet coefficients. SA was used to eliminate the wavelet coefficients whose value is less than the obtained cut off threshold. The starting temperature of the algorithm was set to 100 while the algorithm terminated when the temperature reached 0. Student's t distribution was used to generate new solutions in the SA algorithm. The criteria for accepting a worse solution was Boltzman's probability distribution (Metropolis criterion) which is a function of temperature T as given by the equations:

$$\rho(\Delta F) = exp(\frac{-\Delta F}{T}) \tag{2.10}$$

Where

$$\Delta F = F(\hat{x}) - F(x_i) \tag{2.11}$$

Where F is the objective function, ΔF is used to increment the objective function, x_i is the current values, and $\hat{x}$ is a new solution close to x_i. The performance of the SA was found using a fitness function that was used to move the algorithm towards global optimum solution. The results indicated that the SA part of the model worked well and it was able to successfully eliminate the irrelevant coefficients from the WT.

SA has also been used in other spectral data analysis. Jha et al. [47] have used a SA algorithm to decide the descriptors that decide nearness between molecules on mass spectrometry data. These descriptors help to relate the dependencies of the antimicrobial activity of new compounds on the nature of substitution in oxadiazoles compounds. The authors conclude that the results were able to provide valuable information relating to structure of the compounds.

Schumacher et al. [96] have used SA with Raman Spectroscopy in order to develop a

tree like classifier that differentiates inorganic and organic particulate matter. The method compared classifiers with different algorithms including artificial neural networks (ANN), linear discriminant analysis (LDA) and support vector machines (SVM). SA was used as a classifier for the support vector machine (SVM). SA was used for the optimisation of these parameters. Metropolis criteria with Boltzman's function was used for SA. Cross validation was used for the estimation of the accuracy of the classifier. The results indicate that the SVM classifier with SA in general performed better than LDA but not better than ANN.

SA has also been used on other types of data set in the medical domain. For example, Filippone et al. [31] have used SA along with SVM for gene selection in the classification of gene expression data. The aim of the algorithm was to select the input for aggregating an ideally minimal subset of inputs with strong discriminative power. They called it a simulated annealing input selection (SAIS) algorithm. The authors compared their results with other variants of SAIS and found the results comparable with other versions.

2.13 Type-I Fuzzy Sets

Classical fuzzy sets or Type-I (T-I) fuzzy sets were introduced by Zadeh in 1965 [122]. The aim of fuzzy logic is to represent and analyse data where uncertainties are involved. In fuzzy logic, each element has a degree of membership to a fuzzy set which is described by often a real value in (0,1). For example, classifying a person as young or old is a transition process rather than a quick switch so a crisp set is not able to represent it correctly as shown in Figure 2.10. This example has been taken from [77]. In the figure, x-axis represents age of a person in years and y-axis shows the membership grade as μ. For age 59, the crisp variable considers it as the Young as boundary for Old age starts from 60. Where as in a fuzzy logic system, a transition is made with the help of membership functions as shown in Figure 2.11. Age 59 has membership value for Young as 0.4 and for Old as 0.6. This indicates that age 59 is more towards old age than young. This

information can not be shown with the help of a crisp set. We can say that fuzzy logic is different from classical set theory where an element can either be in a set or not where as in fuzzy theory an elements fuzziness indicates its inclination towards a set. An element has fuzzy values for all the sets in the system.

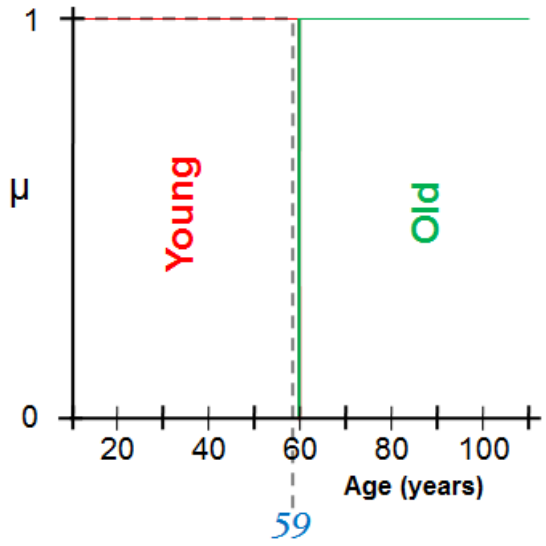


Figure 2.10: Example of a crisp set (taken from [75])

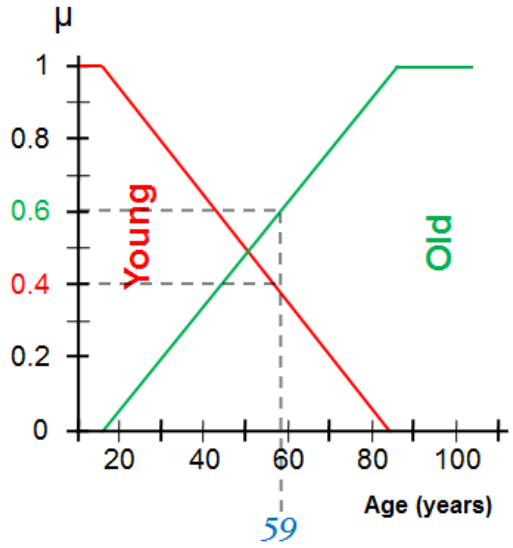


Figure 2.11: Example of a fuzzy set (T-I) (taken from [75])

Fuzzy sets are associated with linguistic terms and they better reflect human reasoning

and decision making. For example if we say that *water* is *very cold* then *very* is the term that reflects upon the magnitude of the fuzzy variable *cold*. In 1975, Professor Zadeh introduced the concept of linguistic variables instead of numerical values in [123].

The fuzzy sets describing linguistic variables are connected through fuzzy rules that take the following expression

IF antecedents(s) **THEN** consequent(s)

The antecedent implies fuzzifying the inputs where as consequent is the result after applying inferencing with those fuzzified input values. Fuzzy rules can have multiple antecedents. These antecedents are connected through fuzzy operators as defined by Zadeh [122]. They are generally referred as fuzzy conjunction (AND), fuzzy union (OR) and fuzzy complement (NOT). A brief description of these operators is as follows.
The intersection between two fuzzy sets x and y is defined as:

$$\mu_{x\cap y} = Min[x,y] \tag{2.12}$$

The union between fuzzy sets A and B is defined as:

$$\mu_{x\cup y}(x) = Max[x,y] \tag{2.13}$$

The compliment of fuzzy set x represented as $\overline{A}$ is defined as:

$$\mu_{\overline{x}} = 1 - x \tag{2.14}$$

where μ is a membership function.

There are other types of operators found in the literature besides Zadeh's conventional operators of *Min* and *Max*. They are generally referred to as Triangular operators (T-operators) where Triangular Norm (T-Norm), Triangular Conorm (C-Norm) and Negation are used for finding Union, Intersection and Compliment of fuzzy sets [36]. Now we briefly describe some of these operators.

Probabilistic operators are defined by Goguen [115] and Bandler et al. [7]. A brief description of operators is as follows.

The intersection between fuzzy sets x and y is defined as:

$$\mu_{x\cap y} = x.y \tag{2.15}$$

The Union between fuzzy sets x and y is defined as:

$$\mu_{x\cup y} = x + y - x.y \tag{2.16}$$

The compliment of fuzzy set x represented as $\overline{A}$ is defined as:

$$\mu_{\overline{x}} = 1 - x \tag{2.17}$$

In Lukasiewicz logic [34], T-operators are described as.

The intersection between fuzzy sets x and y is defined as:

$$\mu_{x\cap y} = Max[x + y - 1, 0] \tag{2.18}$$

The Union between fuzzy sets x and y is defined as:

$$\mu_{x\cup y} = Min[x + y, 1] \tag{2.19}$$

The compliment of fuzzy set x represented as $\overline{x}$ is defined as:

$$\mu_{\overline{x}} = 1 - x \tag{2.20}$$

There are other types of T-operators found in the literature. A few examples of such operators are described by Weber [115] and Yager [119]. Multiple antecedents connected

with fuzzy operators evaluate to give a single number. Consequents can also have multiple parts that can be aggregated to define a single output [81]. The obtained output needs to be defuzzified in order to get a precise solution as a crisp value. Various defuzzification methods have been proposed in the literature [23]. Some of them are

- Centre of Area (COA)
- Mean of Maximum (MOM)
- Bisector
- Largest of Maximum (LOM)
- Smallest of Maximum (SOM)

2.13.1 Fuzzy Inference System

A fuzzy inferencing system (FIS) is a rule-based system that uses fuzzy logic, rather than boolean logic to reason about data. A FIS can also be called a fuzzy expert system (FES) or a fuzzy logic controller (FLC) dependent upon the area of its application. The four main components of a FIS are:

- A fuzzifier, which translates real-valued inputs into fuzzy inputs
- An inference engine, which applies a fuzzy reasoning mechanism to obtain a fuzzy output using a knowledge base
- A defuzzifier, which translates the fuzzy output into a crisp value
- A rule base, which contains an ensemble of fuzzy rules

The fundamental elements of a FIS can be seen in Figure 2.12. The two most common inferencing methods are Mamdani's fuzzy inference method and the Takagi-Sugeno-Kang (TSK) method. In Mamdani's method, the consequents of the rules are fuzzy sets which

are aggregated to produce an output fuzzy set and the final output is obtained after defuzzification over all fuzzy outputs. Where as in TSK which is also called singleton the output is the weighted average of each rule's output [81]. Each input of the FIS is associated with one or more membership functions. There are several types of membership function available, for example, gaussian, triangular, trapezoidal etc. These membership functions are evaluated with fuzzy rules to find the output. A good classification is dependent upon adequate rules, linguistic variables and membership functions. One of the main advantages of a medical based FIS is that it is possible to include the knowledge of specialists, even if statistics data is not available. In the case of medical data, there is a high level of uncertainty and a FIS may help in providing an optimal solution in such cases [43, 87].

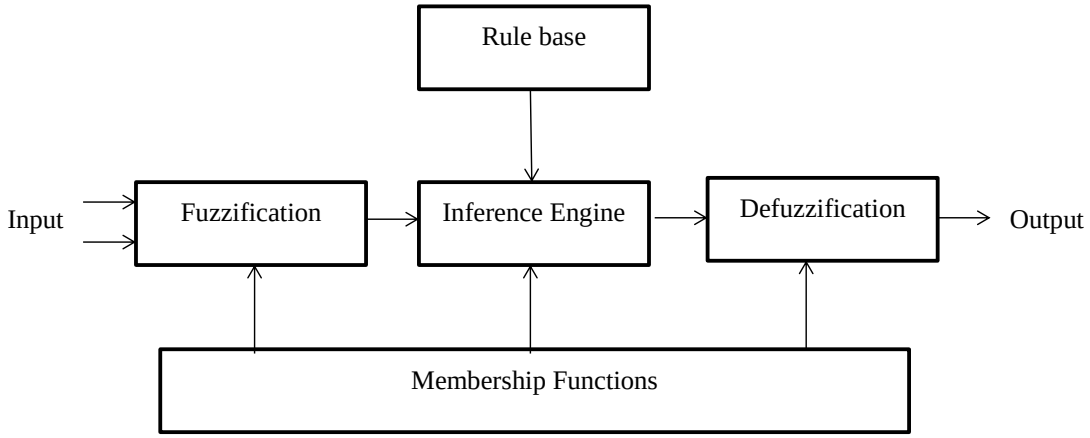


Figure 2.12: Structure of a Fuzzy Inferencing System (taken from [86])

A brief literature review of FIS used with spectral data in medical science follows.

Castanys et al. [17] described a three phase case-based reasoning system (CBR) to identify unknown materials by means of the automatic recognition of their Raman spectra. The first phase consists of dimensionality reduction by means of PCA. The second phase consists of defining similarity measures to objectively quantify the spectral similarity with a final value obtained by the fuzzy logic system. The final phase consists of revision and

validation of the results. The total number of rules developed for the fuzzy logic system was 4. The results indicated that the proposed system worked well and has potential to be used for other problems like identifying pigment mixtures etc.

Evsukoff et al. [29] presented a frame work for intelligent data analysis of spectral data in classification and regression problems. In this frame work, the number of interpolation functions was computed using spectral analysis. Each function was then associated with a symbol to generate a fuzzy rule. Each symbol was related with a prototype that can be computed using a clustering algorithm. A rule induction algorithm was used to determine the minimum number of rules for the frame work. The proposed frame work was tested on various data sets including iris, cancer, wine etc. The results indicate that the frame work performed well in classification of these data sets. The authors concluded that the frame work had the potential to be used in complex problems although there are areas like finding optimal number of rules that require more in depth research.

Cernuda et al. [18] proposed a specific fuzzy system called the TSK fuzzy system for calibrating the chemometric models based on NIR spectra. This fuzzy system was used to model the non-linearity contained in the production process of polytheracrylat (PEA). The TSK based fuzzy system was used to express non-linearities contained in the mapping between NIR spectra samples and measured concentrations or target values. The calibration results obtained by the proposed method were also compared with other state of the art methods. The results showed that the proposed system outperformed other methods in terms of properties associated with calibration and was also computationally comparable.

Mahmoodabadi et al. [66] presented a fully automated system in order to analyse and classify magnetic resonance spectroscopy (MRS) signals of patients with metabolic brain diseases. The selected features from MRS constituted the universe of discourse (input). Every input feature was fuzzified using low (L), normal (N) or high (H) group with a membership value in range [-1,1]. Only normal (N) had a membership value starting from 0. Trapezoidal membership functions were used. The proposed fuzzy membership

functions were used in the classifier to categorize the metabolic brain diseases. The authors stated that the use of specific membership functions was able to increase accuracy and interpretability of the system.

Zhengmao Ye [121] used livers, lungs, kidneys and glands Raman spectra and created and artificial intelligence approach along with fuzzy logic filtering to categorise them. For the fuzzy logic part of the method, from Raman spectra, consecutive intensity differences between any point and its adjacent points are normalsised and linear combination of difference terms was considered as crisp inputs to the fuzzy logic filter. Positive or negative signs of intensity differences were considered for rule making. A Mamdani type fuzzy system was created. Linguistic variables were expressed as fuzzy sets of Negative big (NB), Negative small (NS), Zero (ZE), Positive small (PS), Positive big (PB)]. Centroid defuzzification was used. The authors concluded that their method was able to perform well on these data sets and argued that the method had potential to be used for various cancer cell classifications as well.

Similarly, Pueyo et al. [88] and Kong et al. [56] have also developed fuzzy systems for spectral data sets.

FIS has also been used for various breast cancer data sets in general other than spectral data sets. A brief literature review of such examples follows.

Reyes [87] used a FIS with evolutionary algorithm for development of an automated method, and later on created Fuzzy CoCo (a Fuzzy modelling technique with evolutionary algorithms) and applied it to the Wisconsin Breast Cancer Database (WBCD). However, no experiments were performed on real data sets, therefore, the authors themselves stressed the need for a more practical approach to understand the actual performance of the model.

Uriarte and Castillo [33] compared the results of FCM Clustering algorithm and a FIS based on a combination of FCM and a Genetic Algorithm (GA) on WBCD. This data base consists of 569 cases, 357 benign and 212 malignant. For each case, there are 10 variables defined. The membership values for the fuzzy system were obtained by the re-

sultant grouping after the FCM clustering algorithm was applied. In each group, average, minimum and maximum values obtained from the FCM were used as membership function values. A genetic algorithm (GA) was used to find the rules. For the comparison of the methods, the final number of grouping was considered as a measure of accuracy. After training the FIS, four rules were selected. The results of both the methods were good although the overall FCM clustering algorithm was more accurate at 99.315% as compared to the FIS which achieved 80.136%.

Jain and Abraham [46] created four fuzzy rule generation methods and compared their performance on WBCD data set for breast cancer diagnosis. In the first method, a single fuzzy if then rule was generated for each class using the mean and the standard deviation of attribute values. For each attribute of the data set, 20 membership functions were created and a fuzzy partition matrix was used to create the histogram. In the second method, histogram attribute values were normalised to 1 and used for rule generation and a single fuzzy if then rule was used as in the first method. In the third method, rules were created by homogeneously partitioning each attribute creating a simple fuzzy grid was created. Each attribute had multiple rules instead of a single rule. The last method was a modified version of the fuzzy grid approach. In this method the shape of the membership functions is modified by partitioning only areas which are overlapping. The results showed that modified fuzzy grid approach provided a high classification rate of 99.73% where as modified grade achieved the lowest accuracy of all methods at 62.57%.

Auephanwiriyakul et al. [5] also used FIS to detect abnormalities in mammograms. One abnormality was microcalcification which is a small deposit of calcium and the other was mass which is a lump of fat detected in mammograms by expert radiologists, sometimes a small presence of these abnormalities can be ignored. Real mammograms were used for the experiments. Two FIS were proposed by the authors based on Mamdani's system and their performances were compared. The first system was called the microcalcification detection system and it consisted of four features extracted which were parameters extracted from the mammograms. The second system was called the mass classification

system and it consisted of 3 features. The results indicated that both systems provide good classification performance. The microcalcification detection system accuracy was 78.07% where as mass classification's accuracy was 98.33%.

FIS have also been developed for the diagnosis of other diseases as well for example, Zhtogullari et al. [127] developed a FIS for diagnosis of urinary system illnesses that cause obstruction in the urethra and the bladder, Castanho et al. [25] developed a FIS for the diagnosis of prostate cancer and Nagata et al. [79] used FIS to predict cervical lymph node metastasis in carcinoma of the tongue.

2.14 General Type II Fuzzy Sets

General Type-II fuzzy sets (GT-II) were introduced by Zadeh as a generalisation of the T-I fuzzy sets [74, 123]. A GT-II set denoted by $\tilde{A}$, is characterized by a T-II membership function $\mu_{\tilde{A}}(x,u)$ and can be expressed as [74]:

$$\tilde{A} = \{((x,u),\mu_{\tilde{A}}(x,u))|x \in X, u \in J_x, J_x \subseteq [0,1]\} \tag{2.21}$$

where X is the primary domain, J_x is secondary domain, $\mu_{\tilde{A}}(x)$ is the secondary membership function at x, and all secondary grades $\mu_{\tilde{A}}(x,u) \in [0,1]$. The GT-II set can also be expressed as [73]:

$$\tilde{A} = \int_{x \in X} \int_{u \in J_x} \mu_{\tilde{A}}(x,u)/(x,u) \qquad J_x \subseteq [0,1] \tag{2.22}$$

where $\int\int$ denotes union over all admissible x and u. For discrete universes of disclosure, $\int$ is replaced by Σ. Figure 2.13 shows a GT-II fuzzy set for our example regarding height of a person. It can be seen that secondary membership function itself is a T-I fuzzy set (in blue colour).

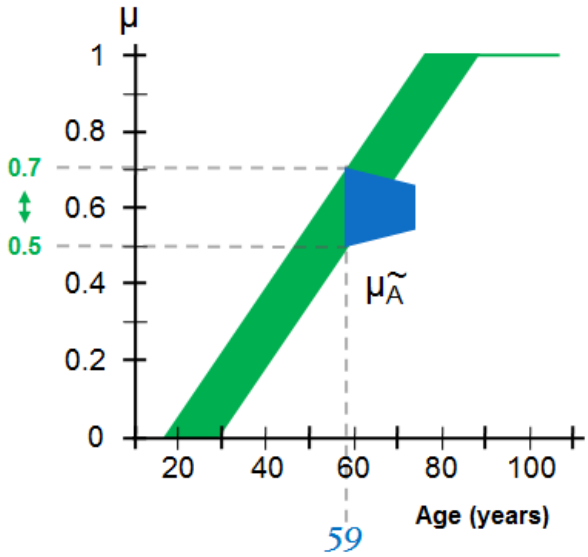


Figure 2.13: Example of a General T-II fuzzy set (taken from [75])

2.14.1 Interval Type-II Fuzzy Sets

An interval Type-II fuzzy set (IT-II) is considered as a special case of GT-II fuzzy set in which the secondary grade is equal to 1 for all $x \in X$ and for all $u \in J_x \subseteq [0,1]$. Thus the IT-II set $\tilde{A}$ can be expressed as [73]:

$$\tilde{A} = \{((x,u),1) | x \in X, u \in J_x, J_x \subseteq [0,1]\} \tag{2.23}$$

The IT-II set $\tilde{A}$ can also be expressed as [73]:

$$\tilde{A} = \int_{x \in X} \int_{u \in J_x} 1/(x,u) \qquad J_x \subseteq [0,1] \tag{2.24}$$

Figure 2.14 shows the height example in the form of IT-II set. It can be observed that now the secondary grade has been fixed to 1 (in blue colour as a line) and the primary membership grade takes the form of an interval between [0.5, 0.7].

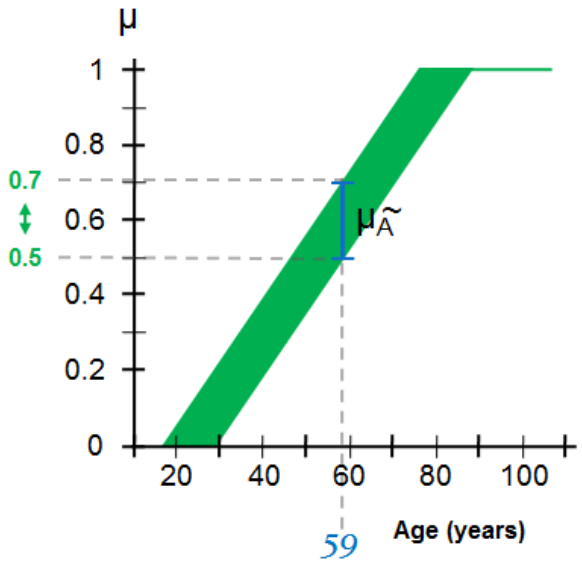


Figure 2.14: Example of an IT-II fuzzy set (taken from [75])

A main advantage of IT-II fuzzy sets is its simplicity and reduced computational cost and that is why they are the most commonly used T-II fuzzy sets [117].

2.14.2 zSlices Representing Type-II Fuzzy Sets

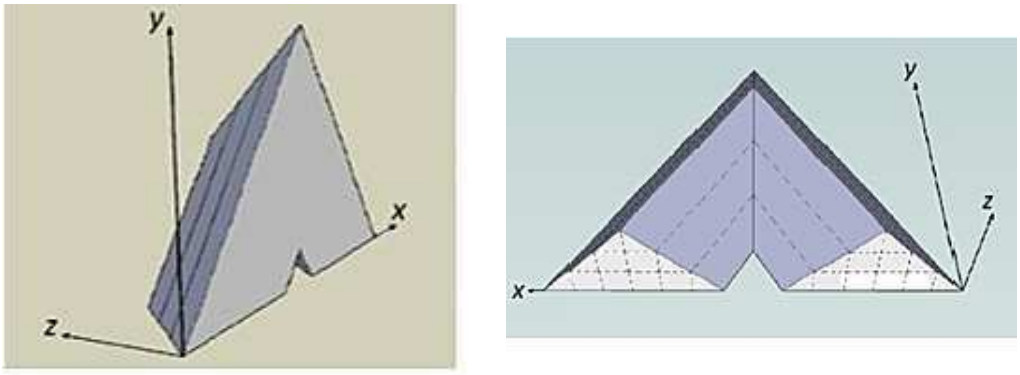
Researchers have focussed on reducing the complexity of the GT-II sets. We shall be using Wagner and Hagras approach for the research work carried out throughout this thesis which uses zSlices approach [37]. In this approach, a GT-II fuzzy set can be represented by slicing in the third dimension (z) at a level z_i to create a zSlices based type-II fuzzy set (zGT-II). The result of this process is a set of zSlices which are IT-II fuzzy sets with a secondary membership grade of z_i which is in contrast to the regular IT-II fuzzy sets whose secondary membership grade is always 1. Thus the zSlice can be written as:

$$\tilde{Z}_i = \int_{x \in X} \int_{u \in J_{i_x}} z_i / (x, u_i) \tag{2.25}$$

Then fuzzy set $\tilde{A}$ is represented as a collection of zSlices.

$$\tilde{A} = \sum_{i=1}^{I} \tilde{Z}_i \tag{2.26}$$

Where I represents the number of zSlices. It is important to note that zSlice z_0 is disregarded because its secondary grade is zero, therefore, z_0 does not contribute to the fuzzy set [37]. Increasing the number of zSlices to represent a type-II fuzzy set increases the accuracy of the set. Figure 2.15 shows an example of a zGT-II fuzzy set with 3 zSlices, the z-axis shows the three zSlices.



(a) Side view with three zSlices (b) Tilted view with three zSlices

Figure 2.15: Example of a General T-II fuzzy set with three zSlices (taken from [107])

Miller et al. [76] have showed that using interval data, both inter and intra-observer variabilities can be incorporated into zSlices based GT-II sets (zGT-II) on survey data. At first T-I fuzzy sets were created with the help of interval data obtained from experts over multiple surveys. These sets represent the intra-expert variability i.e, variability between values of one expert for different surveys. The T-I fuzzy sets are created with the help of

the following Equation:

$$\begin{aligned}\mu(A) = {} & y_1 / \bigcup_{i^1=1}^{N} \bar{A}_{i^1} \\ & + y_2 / \left(\bigcup_{i^1=1}^{N-1} \bigcup_{i^2=i^1+1}^{N-1} \left(\bar{A}_{i^1} \cap \bar{A}_{i^2} \right) \right) \\ & + y_3 / \left(\bigcup_{i^1=1}^{N-2} \bigcup_{i^2=i^1+1}^{N-1} \bigcup_{i^3=i^2+1}^{N} \left(\bar{A}_{i^1} \cap \bar{A}_{i^2} \cap \bar{A}_{i^3} \right) \right) \\ & + \ldots \\ & + y_n / \left(\bigcup_{i^1=1}^{1} \ldots \bigcup_{i^N=N}^{N} \left(\bar{A}_{i^1} \cap \ldots \cap \bar{A}_{i^N} \right) \right), \\ & \textit{where } y_i = \frac{i}{N}\end{aligned} \tag{2.27}$$

Where y is the degree of membership over the survey domain x. It represents the number of intervals overlapping at a certain point. $\tilde{A_n}$ is a series of intervals where $i \in \{1....N\}$ and N is the number of the intervals. The T-I fuzzy set A is defined by the membership function $\mu(A)$. In the Equation 2.27, the '/' sign refers to degree of membership and is not a division sign and the addition symbol represents the union and it is not the arithmetic addition. The T-I fuzzy set is created by taking the union of all the intervals which are associated with a membership of y_1, the union of all possible two tuple intersections of intervals associated with y_2 and so on.

To add multiple experts opinion as inter-expert variability (the variability between values of different experts for different surveys), a zGT-II fuzzy set is created with the help of the following equation:

$$\begin{aligned}\mu(\tilde{A}) = z_1/\bigcup_{i^1=1}^{N} A_{i^1} + \left(z_2/\bigcup_{i^1=1}^{N-1}\bigcup_{i^2=i^1+1}^{N}(A_{i^1}\cap A_{i^2})\right)\\ + z_3/\left(\bigcup_{i^1=1}^{N-2}\bigcup_{i^2=i^1+1}^{N-1}\bigcup_{i^3=i^2+1}^{N}(A_{i^1}\cap A_{i^2}\cap A_{i^3})\right)\\ + \ldots\\ + z_N/\left(\bigcup_{i^1=1}^{1}\ldots\bigcup_{i^N=N}^{N}(A_{i^1}\cap\ldots\cap A_{i^N})\right),\\ \textit{where } z_i = \frac{i}{N}\end{aligned} \tag{2.28}$$

In Equation 2.28, each zSlice is calculated separately and the total number of zSlices is the number of experts involved. A_n is a series of T-I fuzzy sets where $i \in \{1....N\}$ and N is the number of sets. To combine these T-I fuzzy sets into a zGT-II fuzzy set, the agreement principle is applied and a higher secondary membership level or zLevel is associated with areas where more T-I sets overlap. In this way zGT-II fuzzy set created incorporates both the inter and intra-expert variability. It is worth mentioning that no information is added and no assumptions are made. Outliers are not removed but are modelled at a lower level of agreement.

zSlices based logic has also been used in other areas. For example Mbede et al. [68] have used zSlices based T-II fuzzy logic to develop a navigation system for the autonomous navigation of a robot in crowded and dynamic indoor environments. For the system, zSlice based primary membership functions were based on sigmoid membership functions where as secondary membership functions were of triangular. The results showed that proposed system was robust and intelligent. The authors concluded that zGT-II based logic was able to perform better and also was less complex when compared to GT-II sets.

Medical decision making is an area that has high level of uncertainty especially linguistic uncertainty and T-II fuzzy logic is expected to provide better solutions [48]. We provide a brief literature review on T-II fuzzy sets applied in the medical domain.

Chumklin and Auephanwiriyakul [21] developed a system based on GT-II for the detection of microcalcification in Mammograms for breast cancer. Microcalcification is a very small deposit of calcium which is a breast abnormality that can cause cancer. The input to the system was the original image of a mammogram. Four features were selected from the image. Two algorithms, the FCM clustering algorithm and Possibilistic c-means clustering algorithms (PCM) were used. The membership functions were designed using IT-II fuzzy logic. The upper and lower membership functions were defined with reference to the position of data points on the left and right of the centroid with predefined membership values. The results were compared with membership function generation with the FCM clustering algorithm and a Mamdani fuzzy inference system. The results with interval T-II fuzzy logic system with PCM were found to be the best with 89.47% accuracy.

Wang and Yu [113] proposed a T-II fuzzy membership test (T-II FM test) for disease associated gene identification with microarrays of diabetes and lung cancer. They emphasised the need to use T-II fuzzy logic as the data obtained from microarray was noisy with a lot of uncertainties involved and traditional fuzzy logic is not able to handle this complex situation. IT-II fuzzy sets were used because of their simplicity and reduced computational cost. They used a heuristic method for the generation of T-II membership functions. The heuristic method has three main steps.

1. Selection of a heuristic T-I membership function suitable for the data set
2. Setting the parameters for the membership function (These can be provided by the experts)
3. Designing the upper and lower membership functions using a suitable method

The authors used Gaussian membership functions. The primary and secondary membership values for T-II were combined into a traditional membership. Then the genes were ranked based on these values. A comparison was made between patients of a normal gene and patients of infectious gene. For both diabetic and lung cancer genes, a total of 10

genes were used for each disease. 7 out of 10 genes were correctly classified with the help of the proposed T-II based testing method.

Hosseini et al. [42] have used T-II fuzzy logic in combination with a GA for rule extraction for classification. The WBCD was used for to test the proposed method. IT-II Gaussian membership functions were created for the system. The lower and upper bound parameters of the IT-II Gaussian membership function with mean m and standard deviation s were described by the following equations:

$$\overline{m} = m + k_m s, \underline{m} = m - k_m s, k_m \in [0,1] \tag{2.29}$$

$$\overline{s} = s * k_v, \underline{s} = s/k_v, k_v \in [0.3,1] \tag{2.30}$$

where k_m and k_v are parameters for tuning the footprint of uncertainty (FOU), $\overline{m}$ and $\underline{m}$ are the lower and upper bound of the mean and $\overline{s}$ and $\underline{s}$ are the lower and upper bound of the standard deviation of the IT-II lower and upper membership functions respectively. The structure of the rules was also T-II based. For a given pattern $X_p = (X_{p1}, X_{p2}, ..., X_{pn})$, the rule is defined as:

$Rule\,R_i$: If x_{p1} is $\tilde{A}_{i1}$ and ,.....,and x_{pn} is $\tilde{A}_{in}$ then Class C_i with GC_i

where $\tilde{A}_{i1}, ..., \tilde{A}_{in}$ are IT-II fuzzy sets, $i = 1, ..., M$ is the number of rules. A GA was used for rule selection and elimination. A 10 fold cross validation with 100 runs was used to give more accurate results. The average accuracy of the classifier using three rules with only one variable per rule was found to be 95.96% which is better than previously used methods. The authors concluded that the proposed method was able to handle more uncertainties in the rules with the help of T-II fuzzy sets.

Phong and Thien [90] have used IT-II fuzzy sets to create a TSK based fuzzy system for the classification of electrocardogram (ECG) arrhythmic classification. They used a FCM clustering algorithm and the back propagation technique to determine parameters of the T-II fuzzy classifier. Gaussian membership functions were used for the experiments. 70 ECG samples were used for the training and testing of the classifier. The performance

of the classifier was also compared with a T-I fuzzy classifier and with a T-II Mamdani classifier. The results indicated that the performance of T-II base classifier with both TSK and the Mamdani system was better than T-I. TSK was slightly better than the Mamdani T-II as it was found to have better training. The authors concluded that because of the complex nature of the data, T-II fuzzy logic produced better results.

Zarandi et al. [124] have proposed an expert system based on T-II fuzzy logic for processing of brain tumour images. The authors used PCM clustering to create T-II membership functions. Eight rules were developed for the system. The aim of the system was to correctly identify the tumour grade among four grades. The results were also compared with a T-I system. The results indicated that the T-II based expert system performed better than T-I system.

Ozen and Garibaldi [84] have used IT-II fuzzy logic for the development of a fuzzy expert system for the problem of Umbilical Acid-Base (UAB) assessment. UAB assessment is a procedure which is based on analysis of blood taken from umbilical cord. This information is used to obtain information regarding infant's health. Blood samples may contain errors and these errors lead to uncertainty and T-I based system is not good enough to deal with such complex issue. Results indicated that as uncertainty in IT-II fuzzy sets increased, it also resulted in increase of variation in 50 UAB assessments.

A recent review by Melin and Castillo [71] has also shown that applications of T-II in classification and pattern recognition in different areas including medical sciences are increasing and more researchers are inclining towards the use of T-II fuzzy logic in complex scenarios and problems where more and more uncertainty is involved.

There are other types of T-II fuzzy approaches found in the literature. For example, Coupland and John [22] have used an approach where fuzzy sets and fuzzy logic operators are considered as geometric objects and are manipulated only with the help of geometry, Liu [61] has defined α plane for the representation of fuzzy sets and Hamarwi et al. [40] have described an α cut representation for T-II fuzzy sets.

The use of zGT-II fuzzy sets for classification of cancer cell types especially with

a spectral data set is an under explored area of research. In the current work, we have investigated this area with real spectral data sets of breast cancer for the classification of breast cancer grade.

2.14.3 Similarity Measures for Fuzzy Sets

A similarity measure between fuzzy set is the representation of the degree to which fuzzy sets are similar. Similarity measures for T-I fuzzy sets are common and have been used by many researchers [60, 62, 128]. One of the most common methods used for both crisp and fuzzy sets is Jaccard's similarity measure [108]. For crisp sets, it is the division of the intersection and union of two sets as described by the following equation:

$$S_j^{CS}(A,B) = \frac{A \cap B}{A \cup B} \tag{2.31}$$

Equation 2.31 can be modified for a T-I fuzzy set as:

$$S_J^{FS}(A,B) = \frac{\sum_{i=1}^{N} min(\mu_A(x_i), (\mu_B(x_i))}{\sum_{i=1}^{N} max(\mu_A(x_i), (\mu_B(x_i))} \tag{2.32}$$

Equation 2.32 results in a value between $[0,1]$ where 0 shows completely disjoint sets and 1 means that sets A and B are identical.

Similarity measures for T-II fuzzy sets are less common. For IT-II fuzzy sets, some of the common measures have been described by Zeng and Li [125], Jaccard [118], Gorzalczany and Bustince [16]. For GT-II sets, similarity measures have been described by Mitchell [78] and Yang and Lin [120]. Four properties are commonly used to describe any similarity measure. They are:

1. Reflexivity : $s(\tilde{A},\tilde{B}) = 1 \Longleftrightarrow \tilde{A} = \tilde{B}$
2. Symmetry : $s(\tilde{A},\tilde{B}) = s(\tilde{B},\tilde{A})$
3. Transitivity : If $\tilde{A} \leq \tilde{B} \leq \tilde{C}$, then $s(\tilde{A},\tilde{B}) \geq s(\tilde{A},\tilde{C})$

4. Overlapping If $\tilde{A} \cap \tilde{B} \neq \phi$, then $s(\tilde{A}, \tilde{B}) > 0$, otherwise, $s(\tilde{A}, \tilde{B}) = 0$

It is not compulsory for a similarity measure to have all of the properties because applications of a measure may not depend on all of them [69]. As part of current work is to find similarity between zGT-II fuzzy sets, we shall be using a new similarity measure introduced by McCulloch et al [69] which finds the similarity between zGT-II sets. This similarity measure has all four properties previously defined for a similarity measure. We describe this similarity measure now, which is a modification of Jaccard's similarity measure.

For zGT-II fuzzy sets, a measure of similarity for IT-II fuzzy sets can be applied to each zSlice, and the results of each zSlice are combined with the following Equation:

$$S_{ZS}(\tilde{A}, \tilde{B}) = \frac{\sum\limits_{i \varepsilon L} z_i S_\lambda(\tilde{A}_{z_i}, \tilde{B}_{z_i})}{\sum\limits_{i \varepsilon L} z_i} \tag{2.33}$$

where $S_\lambda(\tilde{A}_{z_i}, \tilde{B}_{z_i})$ is any similarity measure for IT-II fuzzy sets. $\tilde{A}_{z_i}$ and $\tilde{B}_{z_i}$ are zSlices from two sets $\tilde{A}$ and $\tilde{B}$ at zLevel z_i. L is the set of zlevels used by $\tilde{A}$ and $\tilde{B}$. A high number of zSlices provides more accurate information from the similarity measure. In this method, each set will use an equal number of zlevels throughout the T-II system based on zSlices.

In McCulloch's method, each zSlice is weighted and the weighted average of Jaccard's similarity for IT-II fuzzy sets is computed for each zSlice. The method can be summarized by the following equation:

$$S_J^{\text{zGT-II}}(\tilde{A}, \tilde{B}) = \frac{z_i \sum\limits_{i=1}^{I} S_J^{\text{IT-II}}(\tilde{A}_i, \tilde{B}_i)}{\sum\limits_{i=1}^{I} S_J^{\text{IT-II}}(\tilde{A}_i, \tilde{B}_i)} \tag{2.34}$$

In Equation 2.34, $\tilde{A}$ and $\tilde{B}$ are zGT-II sets, I is the number of zSlices, i shows a particular zSlice and z_i represents the degree of membership for each zLevel. A value of 0 indicates that zGT-II sets are disjoint where as a value of 1 shows identical sets. An important aspect

of this approach is that because of the weighting of zSlices, the uncertainty represented by each set is properly presented.

Table 2.6 shows a summary of the major computational techniques reviewed in this Chapter. This table provides a categorisation of the type of the method used with relevant paper and the application area which has been investigated in that manuscript. It can be seen from the table that all the methods have been used for a variety of problems from cancer to categorisation of fruits, navigation of robots etc.

Table 2.6: Summary of literature review

Type	Authors	Application area
PCA	Zwielly et al. [129]	Drug-sensitive and drug-resistant human melanoma cells
	Ly et al. [64]	Diagnosis of skin carcinoma
	Bin et al. [12]	Diagnosis of colon cancer
	Goncalves et al. [35]	Differentiate between different sugarcane bagasse samples
	Lasch et al. [59]	To generate a coloured image of different types of tissue
	Kim et al. [52]	Different species plants
	Kim et al. [51]	Different categories of fruits
	Kloss et al. [55]	Diodegradation of polyurethanes
	Manzano et al. [67]	Ageing process of proteinaceous binder materials
	Pallua et al. [85]	Discrimination study of squamous cell carcinoma
	Kumar et al. [58]	Grading of breast cancer
K-means	Wang and Garibaldi [110]	Study of axillary lymph node
	Krafft et al. [57]	Study of Congenital Cystic Adenomatoid Malformations
	Ly et al. [65]	Biopsies from colon and skin cancers
	Ly et al. [64]	Study of Skin cancer
	Steller et al. [101]	Study of squamous cancer
FCM	Lasch et al. [59]	study of colorectal adenocarcinoma tissue sections
	Wang and Garibaldi [110]	Study of lymph node sections
	Steller et al. [101]	Study of cervix uteri
	Steiner et al. [100]	Study of hydrophobicity of protein adsorption
HC	Luts et al. [63]	Extracting features from MRS
	Schief et al. [95]	Shift uncertainty of NMR data set
	Sasic [97]	Coating thickness of tablet
SA	Chen et al. [19]	Study of quality of GML
	Jha et al. [47]	Study of molecules of mass spectrometry data
	Schumacher et al. [96]	Classification of Raman Spectroscopy data
	Filippone et al. [31]	Classification of gene expression data
Type-I	Castanys et al. [17]	Identification of Raman Spectra
	Evsukoff et al. [29]	Classification and regression problem
	Cernuda et al. [18]	Classification of NIR spectra
	Mahmoodabadi et al. [66]	Classification of MRS signals
	Zhengmao Ye [121]	Classification of kidney,lungs etc spectra
Type-II	Miller et al. [76]	Modelling inter & intra expert variabilities
	Mbede et al. [68]	Mobile navigation of Robots
	Chumklin [21]	Mammograms for breast cancer
	Wang and Yu [113]	Gene identification of diabetes and lung
	Hosseini et al. [42]	Rule extraction for WBCD
	Phong and Thien [90]	Classification of ECG
	Zarandi et al. [124]	Classification of brain tumour images
	Ozen and Garibaldi [84]	For assessment of UAB

2.15 Summary

This chapter presents a general overview of breast cancer and the problems related to classification, FTIR, a literature review of spectral data sets in particular with the clustering algorithms k-means and FCM, T-I fuzzy logic and fuzzy inferencing, T-II fuzzy logic and similarity measures for T-II fuzzy sets specifically with zGT-II fuzzy sets.

Initially a description of breast cancer, its diagnosis and prognosis especially in line with widely accepted NPI has been described. Grade is a critical parameter of NPI and problems with manual methods have been described. FTIR is a technique that has been frequently used in the literature in biomedical applications because of its ability to extract key information from molecular cells in the form of spectra. Instead of using the whole spectral regions, specific areas in literature have been described as key features and examples have been given from the literature.

A brief literature review of the commonly used clustering algorithms k-means and FCM with FTIR has also been included as we have used these two algorithms for our data sets.

Fuzzy logic and fuzzy inferencing has been used to handle variability and uncertainty in data found in real world applications. A review of some important work in the medical domain with traditional or T-I fuzzy logic has also been presented.

For more complex and highly uncertain data sets, T-II fuzzy logic is a better choice. An introduction of T-II fuzzy sets has also be given in this chapter. As we have used zGT-II fuzzy sets for our work with interval data, we have described this in detail. Finally, similarity measures used for T-II fuzzy sets with emphasis on zGT-II fuzzy sets has also been given.

In the next chapter, we give an introduction to the data sets. We have used three different data sets for this study each with a different level of complexity. Initial experiments with clustering algorithms of k-means and FCM and their results will also be discussed.

Chapter 3

Data Sets Description and Initial Experiments with Clustering Algorithms

This chapter includes a detailed description of the data sets used during this work for the initial experiments. The preliminary experiments done with standard FCM and k-means clustering algorithms are also part of this chapter.

3.1 First Data Set Description

For our initial experiments, we created a data set which was derived from a real oral cancer spectral data set. This data set was created by combining two data sets previously used in work on oral cancer tissues provided by Derby General Hospital with full consent of patients [109]. The dataset consists of 33 records obtained from three oral cancer patients. This data set is different from the original work because in the original study both data sets, with different histopathological background, were used separately where as for our experiments, we consider them as single data set regardless of histopathological differences. The classifications made by the histopathologists between tumour and stroma cells were recognized as the original results for each set. Stroma is made up of the non-

malignant host cells and plays an important role in cancer pathology. The FTIR spectra for 33 different locations for two patients were obtained. These locations can be seen in Figure 3.1. The data points 1-15 belong to one data set and 16-33 from the second data set. The dotted white line separates the regions between tumour and stroma. Data points 1-5, 11-15 and 16-25 belong to the tumour cells whereas points 6-10 and 26-33 belong to the stroma cells. The spectral range was limited to 900-1800 cm^{-1}. Each spectrum consisted of 900 absorbance values. The aim of the experiments was to separate tumour cells from stroma cells.

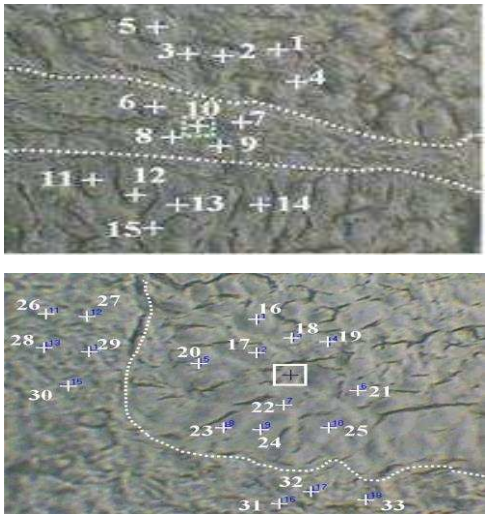


Figure 3.1: Location of data points in tissue samples of two data sets used for FTIR analysis (taken from [109])

3.1.1 Methods

A two step approach was used. First, PCA was performed on the data set and then the FCM clustering algorithm was run 10 times on the reduce data set obtained from PCA. PCA was used to reduce the dimensionality of data while keeping most of the information contained intact. We have used first 10 PCs for this study. The original data set was of size 33*900 wave numbers cm^{-1}. After selecting first 10 PCs, the data set was of size 33*10 PCs that indicates that data set was reduced substantially. For the FCM clustering

algorithm, the fuzziness index (m) was set as 2.0, previous experiments had examined various values of m and found that this was the most effective value, the maximum number of iterations was set to 100. For distance measurements, the squared Euclidean distance was used. The number of clusters was set to 2. The variation of the spread of the data was calculated by the method described in [109]. In this method the percentage variance is calculated as described in Equation 3.1.

$$PV = 100 * \frac{sum[firstN]}{sum[All]} \tag{3.1}$$

Where PV is percentage variance. $firstN$ has variance of the number of PCs included (in our case 10) where as All includes variance from all PCs (in our case 900). The results were compared with the original classifications made by histopathologists to determine the accuracy. Both PCA and the FCM clustering algorithms were implemented using a script written in MATLAB 6.5 (Mathworks, Natick, MA, USA).

3.1.2 Results

The results show that the proposed method created two clusters of tumour and stroma cells having 23 and 10 members respectively. A comparison of results obtained by this experiment with the original results is shown in Table 3.1.

Table 3.1: Comparison of results with PCA+FCM

Tissue Type	Original Result	PCA+FCM	Correctly Classified	Incorrectly Classified
Tumour	20	23	19	1
Stroma	13	10	9	4

These results indicate that there were five data points of the original data that were assigned to the incorrect cluster. In other words, 15.15% of the data was misclassified. This may be due to the adjustments made in the original data sets. The first 10 principal components, along with their percentage variances, are shown in Table 3.2.

Table 3.2: First 10 PCs with the associated variance in data

Number of PCs	Variance (Percentage)
1	50.77%
2	92.16%
3	97.93%
4	98.65%
5	99.09 %
6	99.42 %
7	99.57 %
8	99.78%
9	99.75 %
10	99.81 %

This shows that as the number of components increases, the percentage variance also increases (as expected). It also shows that after a certain point this change becomes smaller (for example from PC5 to PC10). The first two PCs contain 92.16% of the variance of the data and their plot gives a clear visual appreciation of the spread of the members of the clusters as shown in Figure 3.2. It shows cluster members of tumour and stroma cells after the execution of PCA and FCM clustering algorithm. It also indicates that more stroma cells (4) were misclassified by the procedure and most of the tumour cells were correctly classified with only one misclassified.

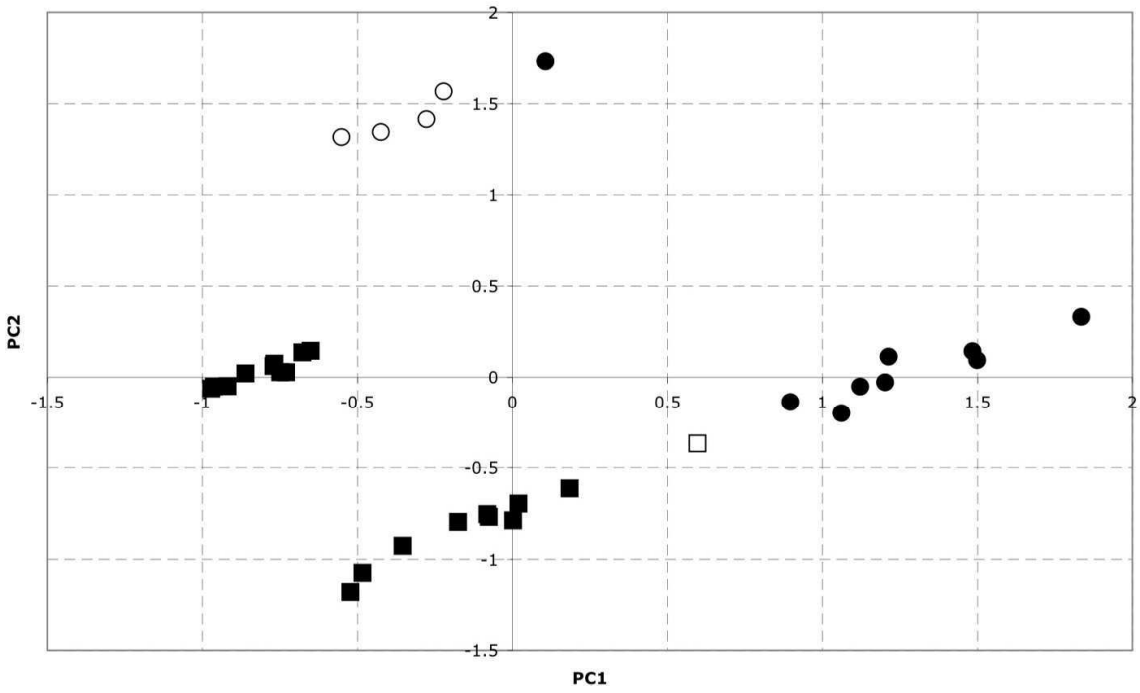


Figure 3.2: Plot of PC1 and PC2. The squares are actual tumour cells; the circles actual stroma cells. The filled (black) symbols are correctly classified; the open symbols incorrectly classified

These results indicate that the combination of FTIR spectra data sets with PCA and clustering algorithms is helpful in classification of the data. It also shows that with a few PCs, the majority of information about the data spread can be obtained helping decrease computational cost for larger data sets that are anticipated to be used for future experiments.

3.2 Second Data Set Description

For this data set, a total of 25 breast cancer tissue samples from 5 different breast cancer biological subtypes were identified from the archives of the Nottingham Tenvous Primary Breast cancer Series. These samples were collected by a team at the breast cancer research group in Queens Medical Centre Hospital in Nottingham. NGS was used to define the

grade for each sample. The samples belong to different breast cancer categories classified as either G-I, G-II or G-III. The aim of these experiments is to classify the grade of a sample with the help of standard commonly used clustering algorithms i.e k-means and FCM.

3.2.1 Methods

Figure 3.3 shows the flow of the work for our experiments in the form of a step wise pipe line. We will now describe each stage of the pipe line in detail pointing out the difficulties involved.

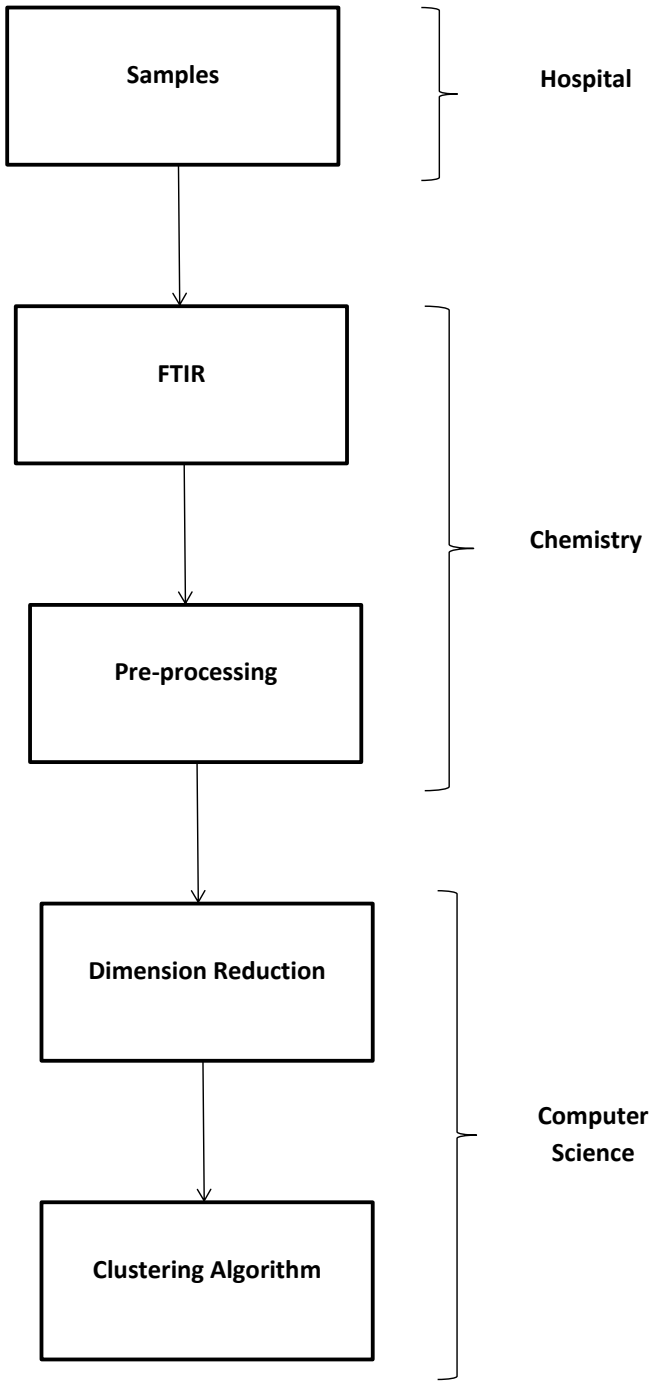


Figure 3.3: FTIR pipeline

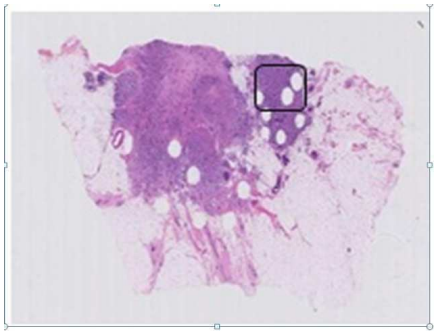
3.2.2 Samples

After detailed discussions with clinical experts at Nottingham City Hospital, we selected one sample of each case for our experiments which were recommended by the experts as these samples were clinically better and reliable.

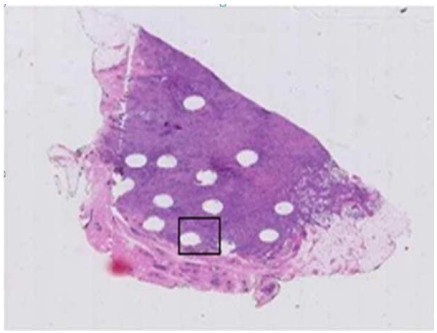
3.2.3 FTIR of Samples

FTIR was carried out on a Nicolet Continum FTIR Microscope in the school of Chemistry, University of Nottingham. For this purpose, selected samples were mounted on a slide and then placed within a slide holder on the microscope and the spectrum was recorded within the region of 800-4000 cm^{-1}.

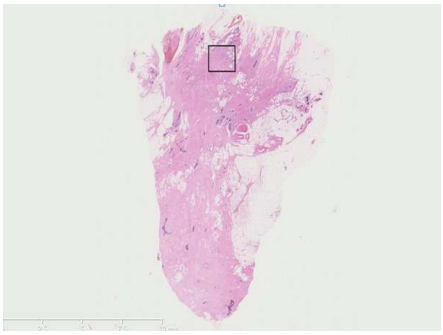
Each spectrum obtained by this method contained 3319 absorbance values and the total size of data for each sample was very large. For the G-I sample, the data size was 25944*3319 wave numbers, for G-II sample, it was 18400*3319 wave numbers and for G-III sample, it was 9393*3319 wave numbers. The data size for each sample varied because of different size of the samples. It is computationally expensive to process such a large amount of data, therefore, a section of the cancerous region from each sample was identified with the help of pathologists and data was extracted for those cancerous regions using a script written in Matlab version 7.02 (Mathworks, Natick, MA, USA). These regions have been represented as boxes in Figure 3.4 for G-I, G-II and G-III selected cases. The number of spectra in each region varies as it is not possible to get same size of section for each sample because of the different shapes of samples. In order to get the best spectral data available, only 100 spectra from each section were used for our experiments. The selection of these spectra was made possible by visual inspection of each spectrum with the help of the clinical experts. In total, 300 spectra (100 of each grade) were used as data set for the current work. It is also important to note that the sections identified were not the only cancerous regions in the samples. The samples also contained non-cancerous regions as well as fatty tissues not included in our experiments.



(a) G-I sample with selected area in box



(b) G-II sample with selected area in box



(c) G-III sample with selected area in box

Figure 3.4: Samples of second data set with selected areas in box

3.2.4 Pre-processing

Baseline correction was performed to correct the sloping baseline that is present with cell spectra. Subsequently all data underwent vector normalisation, to remove effects arising from the thickness of the sample. It was achieved by scaling all spectra such that the squared deviation over the indicated wave number range equals unity. An example of normalised spectra has been shown in Figure 3.6. The two spectra belonging to G-II were clearly different from each other before pre-processing as shown in Figure 3.5. After pre-processing, the blue colour spectrum and red colour spectrum have overlapped and it seems that there is only one spectrum visible in Figure 3.6. Normalisation has synchronized the raw spectra which can now be used for analysis to distinguish it from spectra of other grades. All of the corrections mentioned were achieved using a script written using Matlab version 7.02 (Mathworks, Natick, MA, USA). Spectra were also cut

to the region between 900-1800 cm^{-1} as a finger print region. Each spectrum consist of 934 absorbance values over this region which is significantly less than 3319 found in the original data set.

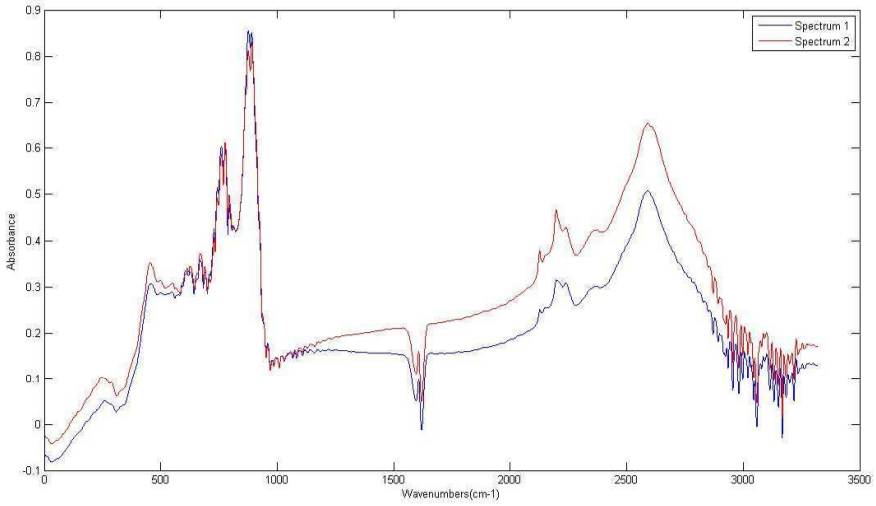


Figure 3.5: Example of non-processed spectra

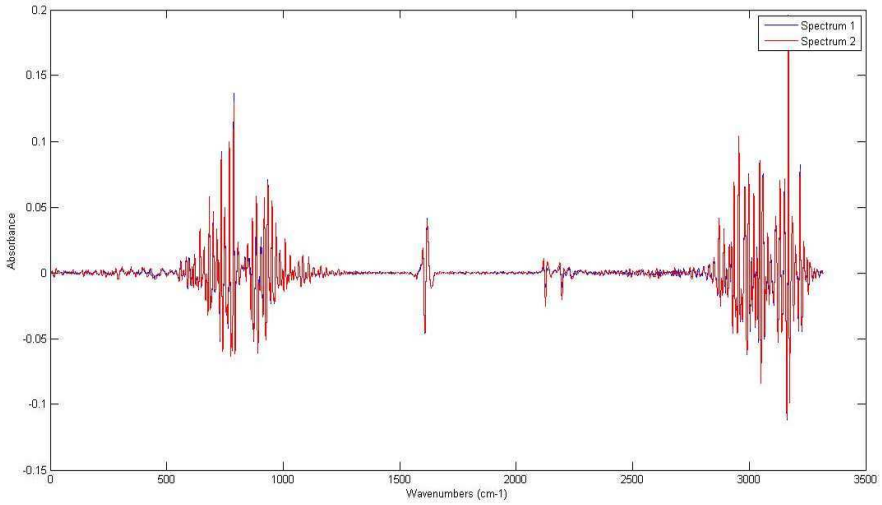


Figure 3.6: Example of processed spectra

A data set was created by combining the selected 100 spectra from the cancerous region of each case. The final data set was of the size 300*934 wave numbers.

3.2.5 Dimension Reduction

For this data set, we have selected PCA as the standard method for dimension reduction. Although our data set size is not very large (300*934), but we have used it as a standard procedure. We have used the first 10 PCs for our experiments. The selection was made on the basis that first 10 PCs contained more than 95% variation of the overall data set. After dimension reduction, the final data set was of size 300 spectra * 10 PCs.

3.2.6 Clustering Algorithms

We have used k-means and FCM clustering algorithms for our experiments. The selection has been made to compare a hard clustering algorithm (k-means) and a fuzzy based

algorithm (FCM).

For the FCM Clustering algorithm squared Euclidean distance was used. Fuzziness index was set a value of 2 and minimal amount of improvement was set as 10^{-5}. For the k-means clustering algorithm, again squared Euclidean distance was used for the measurement and maximum number of iterations was set as 100. The number of clusters for both FCM and k-means clustering algorithms was set as 3. The results obtained from clustering were compared with the classification made by the expert histopathologists.

3.2.7 Results

Table 3.3 shows the results with FCM and k-means clustering algorithm with 3 clusters. Table 3.3a shows results with FCM clustering algorithm. It indicates that cluster one mainly contains members of G-I. Twenty four members of G-I became part of cluster 3 which mainly contains G-II members. Cluster 2 was able to successfully differentiate the G-III members from the data set and only one member was misclassified and became part of cluster 1. Cluster 3 represents 87 members of G-II where as the remaining 13 were part of cluster 1. Table 3.3b describes the results obtained by the k-means clustering algorithm. Cluster 1 has majority of grade 2 members and only one G-III member where as cluster 2 consists of G-II members. G-III is clearly separable by cluster 3. Both FCM and k-means clustering algorithms results indicate that spectral data of G-I and II had less variation, therefore, cluster members became part of each other. In case of G-III data, both FCM and k-means clustering algorithms were able to clearly distinguish it from rest of the grades.

Table 3.3: Results with FCM and k-means clustering algorithm with data set 2

(a) FCM Clustering

Cluster with members	G-I	G-II	G-III
1(87)	73	13	1
2(102)	3	0	99
3 (111)	24	87	0

(b) K-means Clustering

Cluster with members	G-I	G-II	G-III
1(113)	40	73	1
2(87)	60	27	0
3 (100)	0	0	100

Our experiments indicate that both the FCM and k-means clustering algorithms can be used to differentiate between different breast cancer grades with spectral data sets. But there are certain things which need to be considered while dealing with medical spectral data sets. In our experiments G-III was clearly separable form the rest of the grades with both FCM and k-means clustering algorithms, but it may be because G-III spectra was more distinct when compared to G-I and G-II.

The second data set only had one case of each grade. To further investigate a complex data set we have selected the following data set which has more cases of each of grades of breast cancer. The data set also has different number of cases for each grade.

3.3 Third Data Set Description

The third data set is called BR804 obtained from University of Illinois at Urbana Champaign, USA [13]. It consists of 80 cores of 40 cases of paired breast invasive cancer and matched normal adjacent tissue with a single core per class. Figure 3.7 shows the Tissue Microarray (TMA) slide of the data set. TMAs are paraffin blocks in which a very large number of tissue cores (up to 1000) can be assembled together. A major advantage TMA is that it allows experts to do multiplex histological analysis [91]. The TMA of this data set consists of 10*8 Matrix where each pair of circles indicate a cancer case and a relevant normal case. It can also be observed that cancer cases are darker than the normal cases. A detailed microarray panel display is also shown in Figure 3.8 where a normal case is indicated by NAT (not a tumour). The cancer grade break down of the data set is described in Table 3.4. These grades were calculated by expert histopathologists with NGS criteria. Table also shows that there are 6 cases whose grade could not be determined by the normal histopathological procedure and are undefined. The three grades and their relation with cases and spectra is shown in Figure 3.9. It is also important to note that each case has different n_i number of spectra as size of each case is different from other. The FTIR data set for this TMA slide is of size 85.4 GB. It has been calculated between (722-4000)

cm^{-1} for every alternative wave number. It consists of three parameters.

X: Samples = 4062 (along x-axes)

Y: Bands = 3420 (along y-axes)

For each (X, Y) pair there is an absorbance Z that consists of 1641 wave numbers. Each point is of type float (4 bytes). Therefore, the total data size is calculated as:

Total data set size = 4(X*Y*Z) = 85.4GB.

The data set is of an extremely large size. FTIR spectra have been calculated for the whole TMA slide resulting in a large data set. Also, the TMA slide includes normal spectra cases as well as cancerous samples and their spectral results also contribute to the size of the data set. It is worth mentioning here that we have not used the normal sample cores or the undefined cores as the aim is to differentiate cancer grades from one another rather than differentiating them from normal samples.

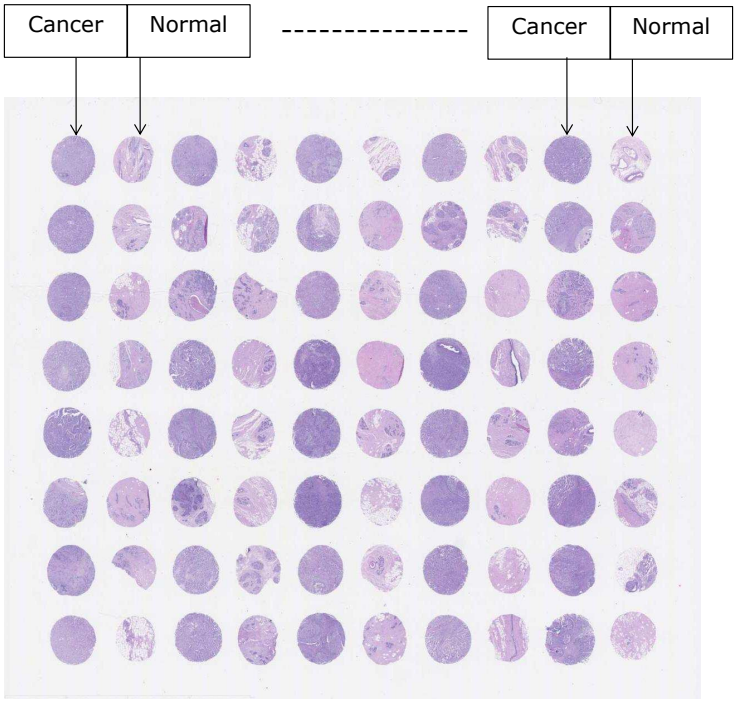


Figure 3.7: TMA Slide of Data set 3

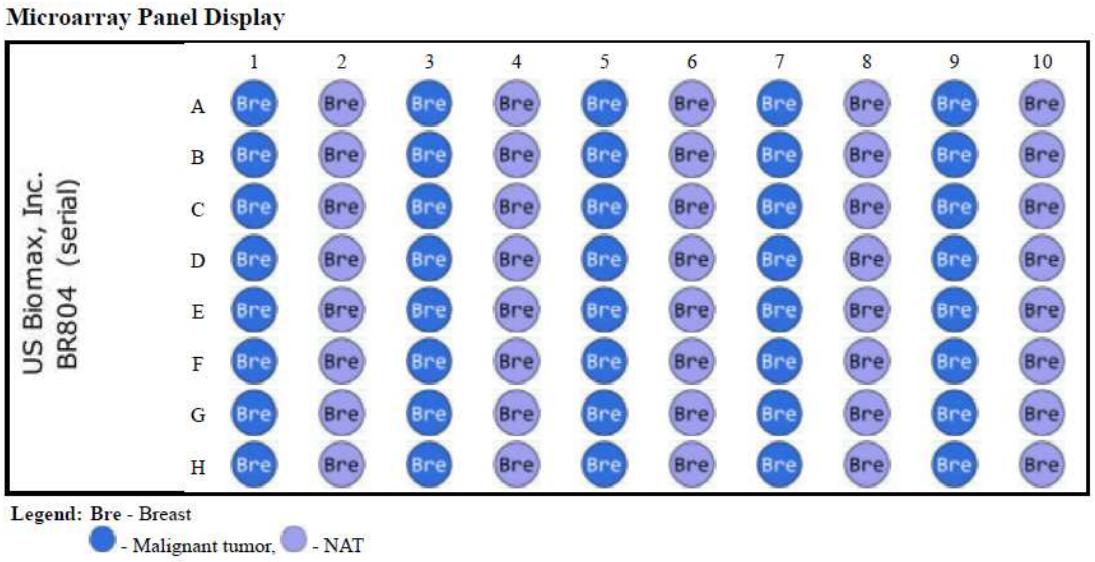


Figure 3.8: Microarray panel display for Data set 3 (taken from [13])

Table 3.4: Categorisation of grades (cases)

Grade-I	Grade-II	Grade-III	Undefined
2	26	6	6

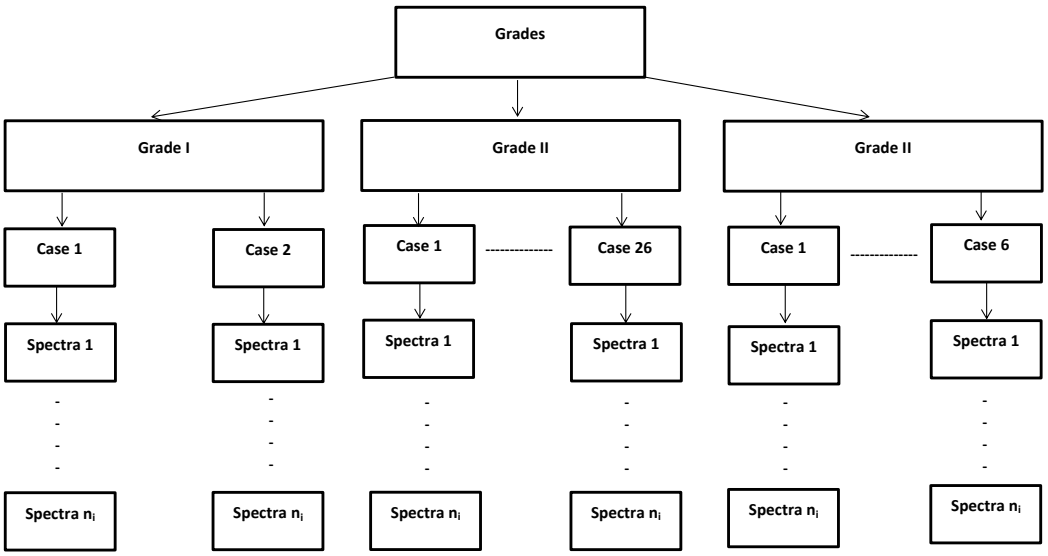


Figure 3.9: Cancer grades and their relation with cases and spectra

Figure 3.10: Example of a sample used for data extraction

3.3.1 Data Extraction

As the data set is very large, extraction of all spectra at the same time is computationally expensive. To resolve this issue, smaller areas were identified with the help of clinical experts. The spectra were extracted from 10*10 square areas from all cases (100 spectra for each square). An example of such a square has been shown in Figure 3.10.

3.3.2 Data Pre-processing

Instead of using whole spectral range for the experiments, we have selected the spectral region between 1000-1800 cm^{-1} for our experiments as in literature the spectral region around this region (sometimes starting with 900 cm^{-1}) has frequently been used as it is estimated to include spectra that can provide valuable information about the data [109]. Thus reduced the wave numbers values from 1641 to 401 which was significantly less. Data from the selected spectral region was pre-processed with standard base line correction and normalisation process using a script written in Matlab provided by the School of Chemistry, University of Nottingham. The processed data was discussed with experts in

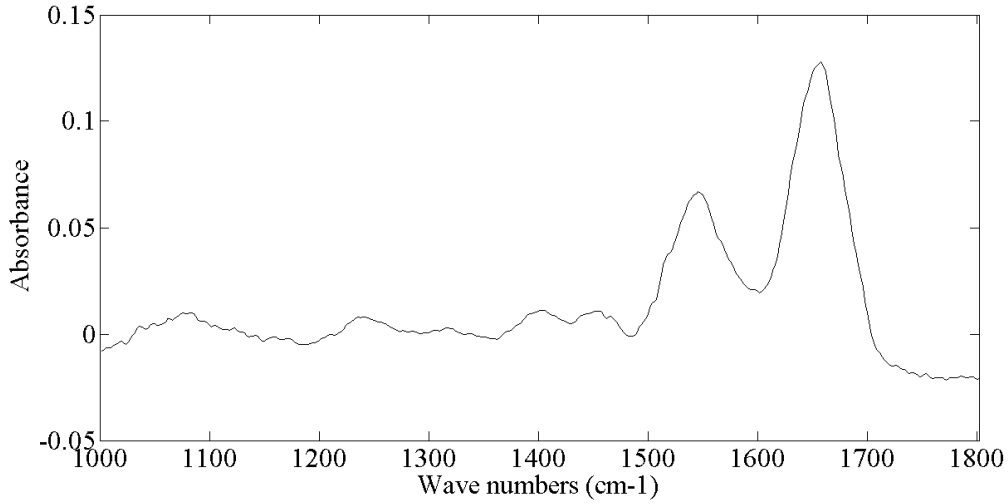


Figure 3.11: An Example of a pre-processed spectra in region 1000-1800 cm^{-1}

the School of Chemistry and then used for the experiments performed here.

3.3.3 Dimension Reduction

PCA was used to reduce the dimensions of the data. Initially we have selected 10 PCs as they cover more than 90% of the variance of the data. These PCs were used in combination with clustering algorithms for our experiments.

3.3.4 Clustering Algorithms

We have created a data set of 300 spectra * 10 PCs for experiments with clustering algorithms. As we have different number of cases for each grade, we have used the following approach to create a balanced data set.

- G-I: 100 spectra * 10 PCs (2 cases, 50 spectra from each case)
- G-II:104 spectra * 10 PCs (26 cases, 4 spectra from each case)
- G-III: 96 spectra * 10 PCs (6 cases, 16 spectra from each case)

For this complex data set, we have again used the two clustering algorithms, one a hard clustering algorithm, k-means and other a fuzzy logic base algorithm, FCM. Both algo-

rithms were used with 3 clusters as an input in order to classify the three grades. The parameters of the algorithms were the same as described in subsection 3.2.6. The aim of the research is to create a decision support system with the help of advanced fuzzy logic for breast cancer grading. Therefore, We have used both k-means and FCM clustering algorithms in order to explore only the complexities of the data set.

3.3.5 Results

Table 3.5 shows results with the k-means clustering algorithm with 3 clusters. It can be seen that cluster 1 mainly consists of values of G-II and G-III but with no clear distinction. Cluster 2 has G-I and G-III members and cluster-3 mainly consists of members of G-I and G-II. The results indicate that because of the complexity of the data no cluster was able to differentiate between the three grades.

Table 3.5: Results with k-means clustering algorithm with data set 3

Cluster with members	G-I	G-II	G-III
1(112)	13	51	48
2(91)	40	17	34
3 (97)	47	36	14

Table 3.6 shows results with FCM clustering algorithm with three clusters. It can be seen that cluster 1 include a small number of members from G-I and majority classified as G-II and G-III. Cluster 2 include members from all grades. Cluster 3 was able to differentiate 50 members of G-I out of 100 correctly.

Table 3.6: Results with FCM clustering algorithm with data set 3

Cluster with members	G-I	G-II	G-III
1(83)	5	41	37
2(135)	45	37	53
3 (82)	50	26	6

The results with clustering algorithms indicate that neither of the two algorithms are able to differentiate between the three grades with 10 PCs. The results indicate that be-

cause of high level of variabilities involved in the data set, unsupervised learning with standard clustering algorithms is not able to find a clear distinction between cancer grades. These uncertainties may exist between between spectra of same case of grade (intra-case) as well as between multiple cases same grade (inter-case). The other type of uncertainties may exist between cases of same grade (intra-grade) and between cases of different grades (inter-grade).

The results indicate that both k-means and FCM clustering algorithms are able to show the complicated nature of the data as both algorithms performed poorly on the data set. In the next stage of the research, we take this complicated data set and move towards a decision support system with fuzzy logic.

3.4 Summary

In this chapter, we have used three different data sets with standard k-means and FCM clustering algorithms to differentiate between different classes with the help of PCA. Data set 1 was used to distinguish between tumour and stroma cells with PCA and FCM. The results indicated that the method was able to make good classification. Data set 2 was a real breast cancer spectral data set used to differentiate between three cancer grades with PCA and k-means and FCM clustering algorithms. Results indicate that both methods are able to differentiate between three grades. Data set 3 was a real complex spectral data set involving a variety of cases from all grades. The same methods of PCA with k-means and FCM clustering algorithms were applied to differentiate between breast cancer grades. Results indicate that because of variabilities between cases of same grade and between grades, the clustering algorithms perform poorly and are not able to distinguish between grades emphasising high level of uncertainties involved in the data set. In the next Chapter, we take this complex data set and move towards a decision support system supervised learning approach by developing a fuzzy inferencing system that can classify the correct grade for such a complicated data set.

Chapter 4

Experiments with Fuzzy Inferencing System

In this chapter, we have used data set 3, and a Mamdani type fuzzy inferencing system (FIS) has been developed with 300 spectra and using three PCs taken from different cases of each grade for classification. The system uses HC and SA algorithms to train membership functions and rules. The developed system has also been tested with unseen data. The results are compared with the standard k-means clustering algorithm and the performance of the system is discussed.

4.1 System Structure

Figure 4.1 shows a block diagram of the main structure of the FIS used. A data set has been created either by selecting data taken from a single case per grade or from all cases of all grades. The created data set goes through PCA and first three PCs are selected as an input to the system. Each input has three membership functions. These membership functions are trained with the help of three training methods namely, Hill Climbing with Membership Function Tuning (HCMT), Simulated Annealing with Membership Function Tuning (SAMT) and Simulated Annealing with Membership function and Rule Tuning (SAMRT). The HC algorithm is selected because it has been previously used in a spectral

problem to find correct target spectral peak and was able to perform well [95]. We have selected SA in order to avoid limitations of HC as in complex scenarios, HC tends to tilt towards local minima. In case of SA, the chances of getting a better solution increase. In the literature, SA has been found to perform well with complex FTIR spectral data in order to find the optimal cut off threshold for detecting the quality of glycerol monolaurate (GML) [19]. It shows that SA can be useful in problems where complexity of spectral data is high. The best trained FIS is found by comparing the results on training data. The selected FIS is tested on unseen data. The next sections describe the processes involved in each of these steps.

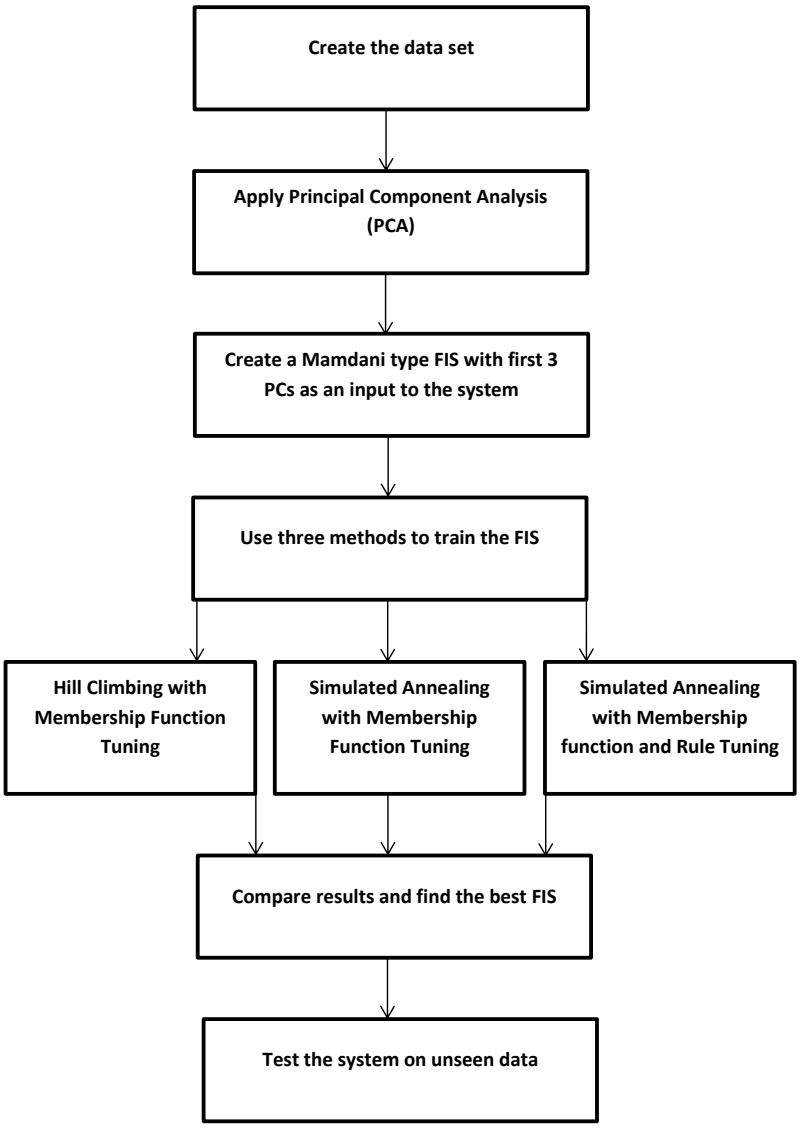


Figure 4.1: Main structure of FIS

4.2 Spectral Regions

We have used all of the spectral region available (722-4000 cm^{-1})and the spectral region between 1000-1800 cm^{-1} wave numbers in two different sets of experiments using spectra

from all the cases available (multiple case study). The first set uses spectral data from all the spectral regions available where as the second set uses spectral data from 1000-1800 cm^{-1} spectral region. The reason is to show that instead of using the complete spectral region, a smaller region is also able to achieve similar results. It has been shown to produce good results in previous studies done with breast cancer spectral data sets [109]. The advantage of using a smaller region is that it reduces the number of spectra, making it more computationally efficient. All spectra have been pre-processed by the standard methods as described in Chapter 3.

4.3 Case Studies

We have considered two types of case studies, namely, a single case study and a multiple case study for our experiments. We now describe each of them.

4.3.1 Single Case Study

100 spectra were extracted from each of the cases F1 (G-I), A9 (G-II) and D7 (G-III)(from Figure 3.8). These cases were arbitrarily chosen as a test case. The data set consists of 300 spectra*1641 wave numbers for the whole spectral region, and 300*401 for the 1000-1800 cm^{-1} spectral region. PCA was used to reduce the dimensions of the data, and the first three PCs were selected. The selection of the number of PCs was made in order to keep the initial number of fuzzy rules manageable. The final training set after PCA was 300 spectra * 3 PCs for both regions. It is worth mentioning that the single case study was done as a proof of concept for the new system. A large number of possible combinations of single cases are available, and different combinations may produce different results depending upon how close the PCs of cases are.

4.3.2 Multiple Case Study

With all available cases involved, the data set was 300 spectra * 1641 wave numbers. The same pre-processing and PCA was performed as with a single case, and the first 3 PCs were used as input to the system. As the number of cases was not the same for all grades, the following number of spectra were extracted to make a balanced data set.

- G-I: 100 spectra * 3 PCs (2 cases, 50 spectra from each case)
- G-II: 104 spectra * 3 PCs (26 cases, 4 spectra from each case)
- G-III: 96 spectra * 3 PCs (6 cases, 16 spectra from each case)

4.4 Development of Fuzzy Inferencing System

All experiments were carried out with Matlab using the Fuzzy Logic Tool Box. We have used a Mamdani type fuzzy system for our experiments with three input variables, three membership functions (MFs) for each input variable and one output variable with three membership functions. The three input variables are the first 3 PCs. The total number of membership functions is 12 (9 for input variables and 3 for output). We have used Triangular type membership functions for our experiments as they are commonly used and computationally efficient. The output of the system is a grade classification matching the possible cancer grades I, II or III. Initially, 27 rules were created for the system comprising of all possible combinations of the input membership functions. Table 4.1 shows these rules. The rules were joined by fuzzy *AND* operator. The consequent of the rules are decided by a majority vote out of the three membership function values. Rules 6, 8, 12, 16, 20 and 22 were not included for experiments with Hill Climbing and Simulated Annealing with membership functions. They were excluded from these experiments because they result in a tie when majority vote is used. For Hill Climbing and Simulated Annealing with Membership Function Tuning, the rule set was comprised of 21 rules. For Simulated Annealing with both Membership Function and Rule Tuning,

they were included for experiments as consequents of the rules were changed during the experiments and majority vote was not used for the final assignment. The Defuzzification method selected was Largest of Maximum (LOM). This method finds the maximum value out of three output membership functions and assigns MF with the largest value as grade for the input data.

Table 4.1: Fuzzy rule set for FIS

Rule	MF1	MF2	MF3	Consequent
1	1	1	1	1
2	1	1	2	1
3	1	1	3	1
4	1	2	1	1
5	1	2	2	2
6	1	2	3	1/2/3
7	1	3	1	1
8	1	3	2	1/2/3
9	1	3	3	3
10	2	1	1	1
11	2	1	2	2
12	2	1	3	1/2/3
13	2	2	1	2
14	2	2	2	2
15	2	2	3	2
16	2	3	1	1/2/3
17	2	3	2	2
18	2	3	3	3
19	3	1	1	1
20	3	1	2	1/2/3
21	3	1	3	3
22	3	2	1	1/2/3
23	3	2	2	2
24	3	2	3	3
25	3	3	1	3
26	3	3	2	3
27	3	3	3	3

4.4.1 Single Control Point and Multiple Control Points (SCP and MCP)

Two types of control points (CPs) were defined for the membership functions of each input variable. When using a single control point (SCP), three membership functions parameters are set to zero and the other six are set to either the maximum or minimum of the range of each input variable. Figure 4.2 shows an example of the membership function parameters for an input variable. The vertical red line represents the control point passing through the parameters that are controlled by the control point i.e, right parameter of MF1, the centre parameter of MF2 and the left parameter of MF3. Out of the remaining 6 parameters (2 for each membership function), for MF1, the left parameter is set to the minimum value of the input range and the centre parameter is set to 1, for MF2, right and left parameters are set to minimum values within range for the input variable, for MF3, the centre parameter is set to 1 and right parameter is set to the maximum value within input range. In the case of SCP when the control point is incremented by a random value within the range, all three parameters that it controls take the same value and move in the same direction as shown in Figure 4.3. They are controlled by single value which is why we call it SCP. In case of MCP, all three parameters are controlled independently of each other meaning that each of them is assigned a different random value generated within the range and they move in that direction regardless of other parameters as shown in Figure 4.4. This provides three independent control points instead of a single value for all parameters. The rest of the parameters are kept constant to keep the working of control points simple and efficient. If we use all 9 control points then moving all of them will be a bottleneck as we shall have to restrict movement to avoid values that are not valid for left, right or centre parameters of a triangular membership function.

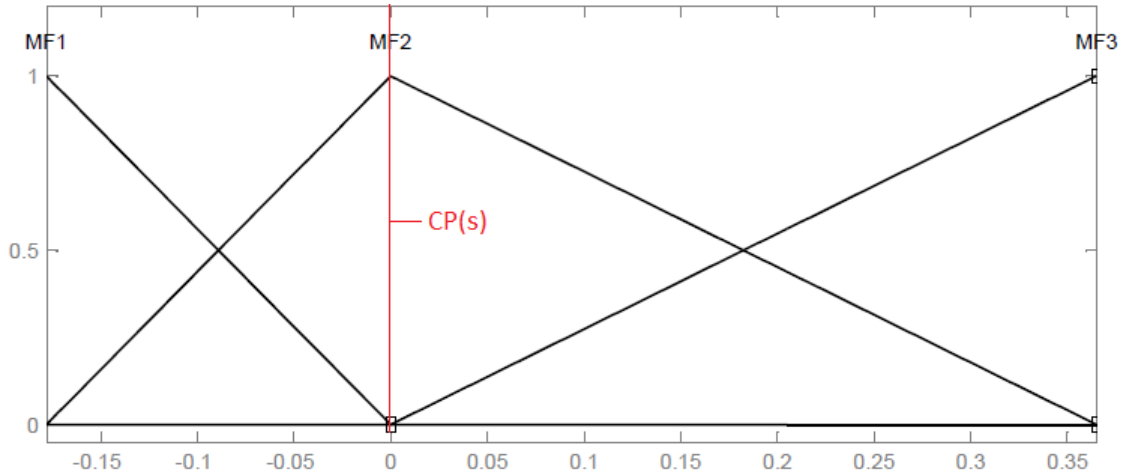


Figure 4.2: Control Points (CP=0)

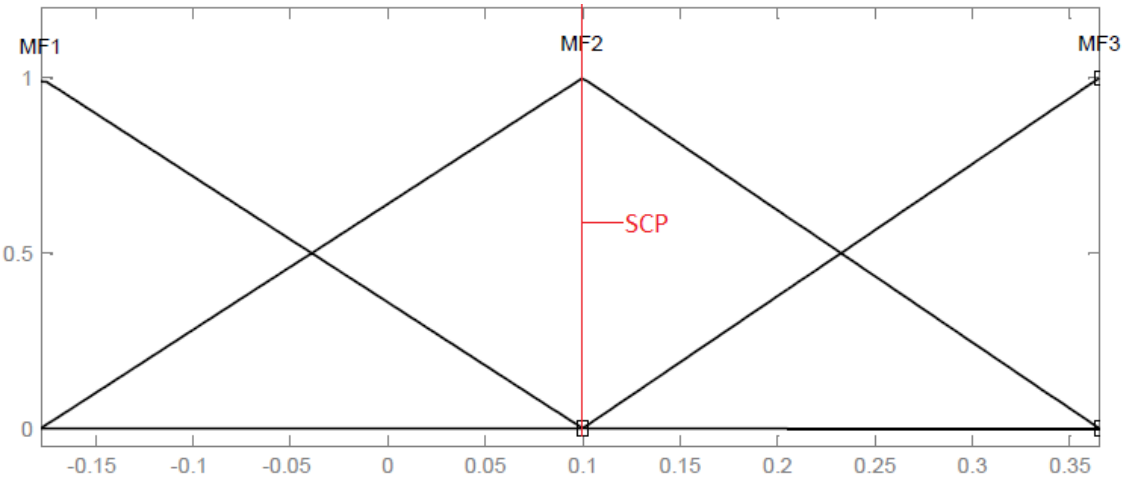


Figure 4.3: Single Control Point (CP=0.1)

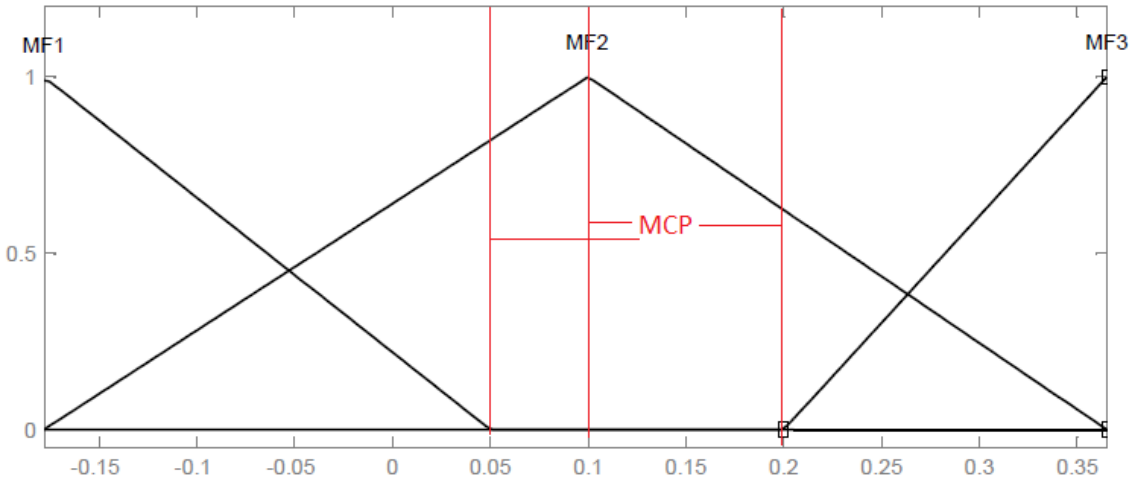


Figure 4.4: Multiple Control Points (CP1=0.05, CP2=0.1, CP3=0.2)

4.5 Fuzzy Inferencing System Training Methods

The following methods were used to train the FIS. Training was performed on CPs for the MFs for each input variable with Hill Climbing and Simulated Annealing with Membership Function and Rule Tuning by changing the consequents of the fuzzy rules. The aim of the training was to find the membership function parameters and rules that provide the best possible classification for breast cancer grading.

4.5.1 Hill Climbing with Membership Function Tuning

Figure 4.5 shows a flow chart of the Hill Climbing with Membership Function Tuning (HCMT) method used for our experiments. In the HCMT method, CP(s) are initialised at zero and then CP(s) are incremented or decremented by a small random value within the input range. The input range is defined as the values between the minimum and maximum values of the input variables. The rule set consists of 22 rules as described in Table 4.1. After that, the FIS is evaluated and the output is compared with the known output for each spectra. The squared error is calculated using Equation 4.1.

$$Er = (GI - c1)^2 + (GII - c2)^2 + (GIII - c3)^2 \tag{4.1}$$

where Er is the squared error, GI, GII and $GIII$ are the number of spectra available for each grade and $c1$, $c2$, $c3$ are the number of correct classifications of each grade. If it is lower, it is accepted and then, with the CP(s) and $c1$, $c2$, and $c3$, the process is repeated and the FIS re-evaluated. A lower value of Er shows improvement in the result towards a better solution. The process is continued for 85 iterations. This number is selected to keep it in line with the other two methods as they also have the same number of iterations. The aim of the method demonstrated here is to find the optimal values of CP(s) by reducing Er to as low value as possible and accepting parameter values only when Er is reduced and rejecting all other solutions. The squared error is used to avoid tilting of accepted values towards correctly classifying one grade. It keeps a balance between correct grade values of all grades by using squared error. If summation of correct classification is used as the criteria then if one grade is completely classified correctly, it will indicate a reasonable solution where in reality it will be a poor solution. To avoid such scenarios, we have used squared error criterion. In the case of SCP, only one CP per input is used for HCMT where as in case of MCP, 3 CPs per input are used for HCMT, and the rest of the procedure remains the same as described before. The entire process is repeated, initialising the system with 10 different random numbers within the input range. The exit criteria is the maximum number of iterations which is 85. At the end, the system with the least squared error is considered to be the best system and is used for testing of the system. Testing is based on unseen spectra taken from all cases.

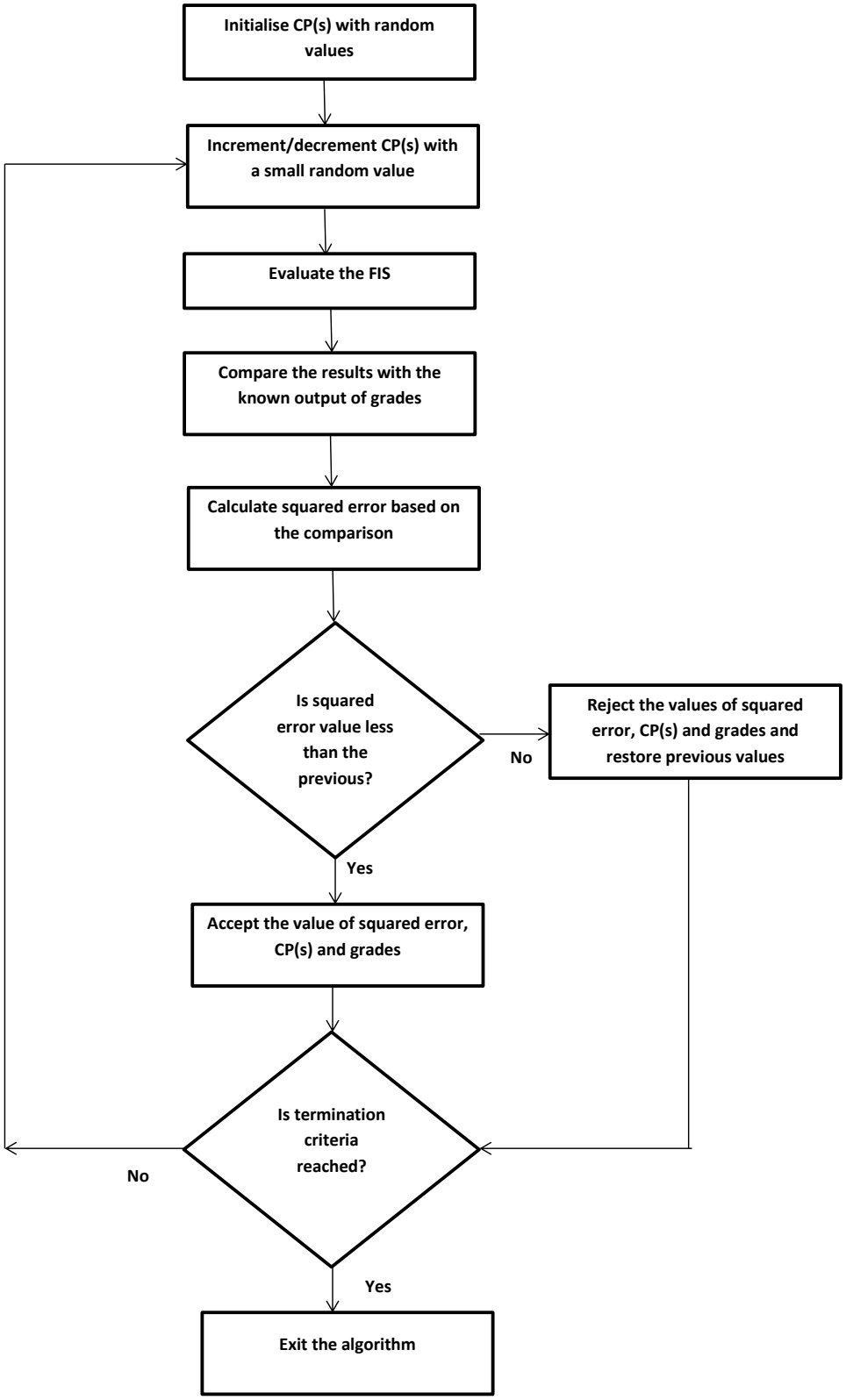


Figure 4.5: Flow chart for Hill Climbing method

4.5.2 Simulated Annealing with Membership Function Tuning

Figure 4.6 shows the flow chart of the Simulated Annealing with Membership Function Tuning (SAMT) method used for our experiments. In this method, CP(s) are initialised with a random value(s) within the input range. A starting temperature of 170 has been selected. This selection was made after a number of experiments with different starting temperatures. The selected temperature was found to provide the best result. The parameters of the Simulated Annealing algorithm are as follows.

- Starting Temperature (T)=170
- Cooling Schedule after each iteration: T=T-2
- Stopping criteria: T=0
- Total number of iterations:85

As before the rule set consists of the 22 rules defined in Table 4.1. The initial FIS is evaluated, then perturbed, then evaluated again. If the squared error is greater than the previous value, then it may still be accepted based on the probability function as defined in equation 4.2.

$$accept = rand() < e^{-(er2-er1)/T} \tag{4.2}$$

where $rand()$ generates a random number from 0 to 1, $er2$ and $er1$ are new and previous squared error values respectively, T is the temperature value at that time. The solution is accepted if the random value is less otherwise it is rejected. Initially, simulated annealing accepts every solution, as the temperature decreases, the cooling down starts to reject solutions where squared error is high. By the end of the search, eventually, simulated annealing reduces to hill climbing. The algorithm stops when the temperature reaches zero which takes 85 iterations. The best CP(s) are saved and used for testing on unseen data. We have used 10 different random initialisations to cover the range available for membership function parameters. Like in HCMT, for SCP, only one value per input PC

is used and in case of MCP, 3 different values for membership functions parameters per input PC are used.

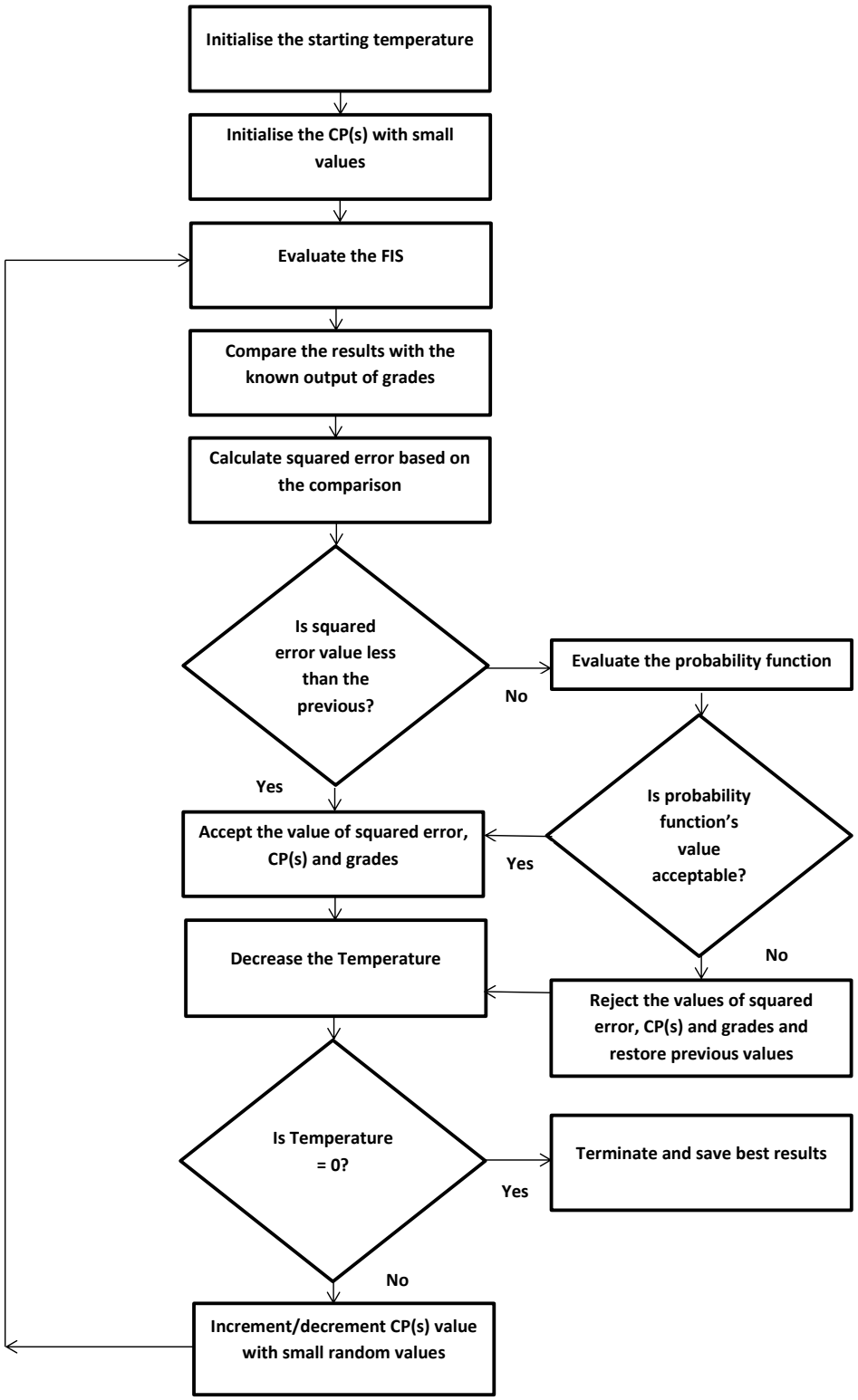


Figure 4.6: Flow chart for Simulated Annealing with Membership Function Tuning

4.5.3 Simulated Annealing with Membership Function and Rule Tuning

Figure 4.7 shows a flow chart for the Simulated Annealing with both Membership functions and Rule Tuning (SAMRT) algorithm used for our experiments. The FIS CP(s) are initialised in the same way as in SAMT. In this method, the consequents of the fuzzy rules shown in Figure 4.1 are also changed before the FIS is evaluated. The rule set for this method includes all 27 rules as defined by Table 4.1. We have selected to change 3 out of 27 rules consequents in each iteration. The value of the consequents of 3 rules selected at a random are changed to a random value in (1,3). The parameters for the simulated annealing algorithm remain the same as for SAMT. The aim is to find whether any rule consequent changes the results. The rule changes also allows membership functions to move along within input parameters and behave differently because of changes of rules. After the FIS has been evaluated and the squared error and probability functions have been executed, we change the values of the CP(s) and the consequents of the 3 rules. The algorithm is repeated for 10 random initialisations and the best values of the FIS parameters are saved and used for testing on unseen data. The process is repeated for both SCP and MCP in the same way as for HCMT and SAMT. After evaluating all of the methods, a comparison between the results is made and the best method is identified.

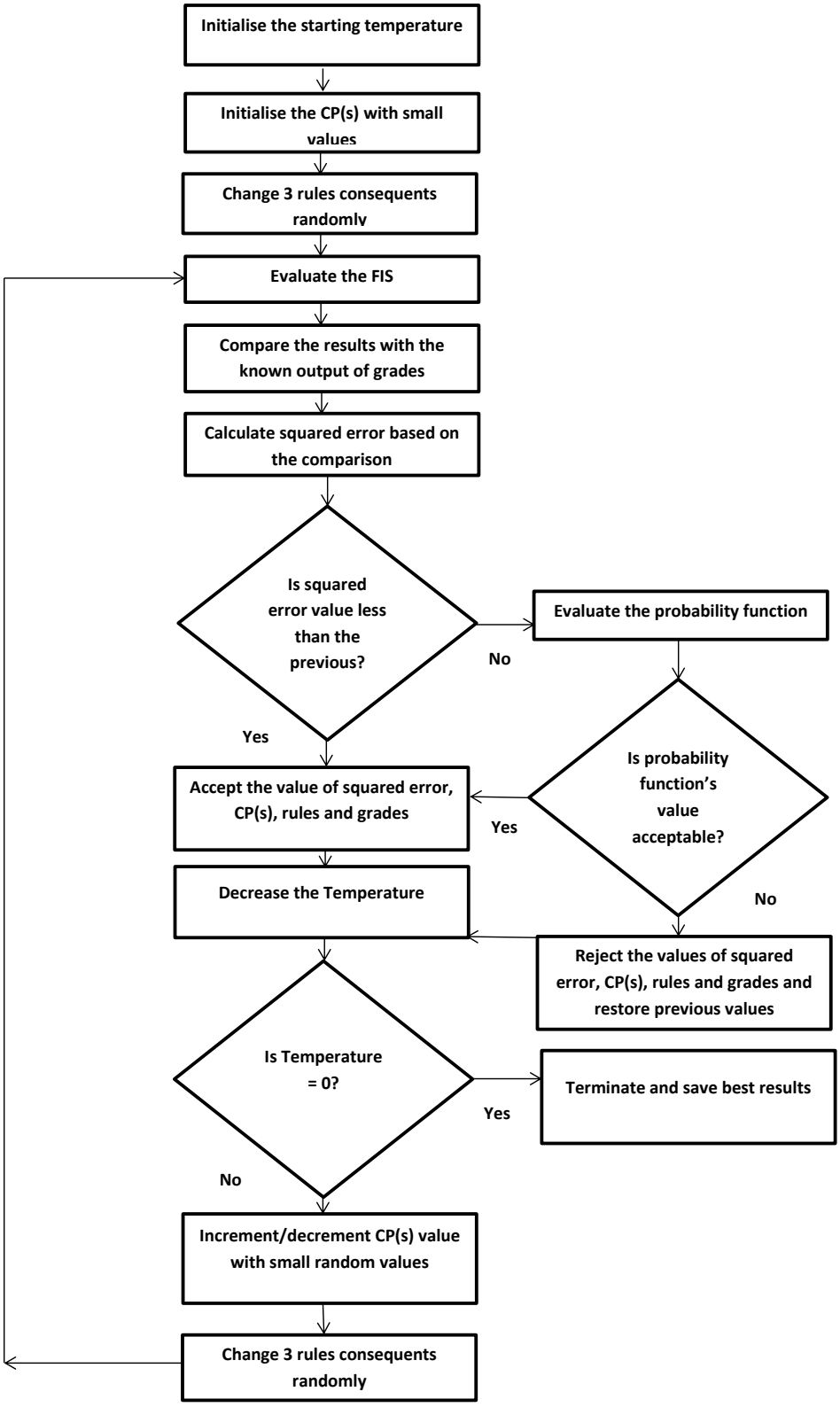


Figure 4.7: Flow chart for Simulated Annealing with Membership Function and Rule Tuning

4.6 Results with Single Case Experiments

For all experiments with a single case, 10 random initialisations were used for each method. All initialisations used a fixed random seed to ensure that numbers are generated in same sequence for all the experiments and can be repeated. The best training result out of each 10 was selected for use with the testing data. The results are shown as FIS percentage accuracy out of 300 test spectra (100 from each grade) which has been calculated by the formula shown in Equation 4.3.

$$PA = \frac{SumCP}{Tspectra} * 100 \tag{4.3}$$

where *PA* is Percentage Accuracy, *SumCP* is sum of correct spectra and $Tspectra$ is total number of spectra used for an experiment.

Table 4.2 shows the results of all the six methods. It can be seen that in the case of HCMT-SCP, HCMT-MCP, SAMT-SCP and SAMT-MCP, the results are less than 50% accurate which is quite poor for both the whole region and the region 1000-1800 cm^{-1}. In the case of SAMRT-SCP the results improved considerably and in case of region 1000-1800 cm^{-1}, over 80% correct. SAMRT-SCP also performed better than HCMT and SAMT. It can also be observed that in the case of all regions and region between 1000-1800 cm^{-1}, the results were slightly better in the region 1000-1800 cm^{-1}. It shows that instead of using the whole spectral region , region between 1000-1800 cm^{-1} can be used as bench mark region. This observation is also supported by the fact that this region has been frequently used in the literature as benchmark region [4, 9, 20, 50, 102, 109].

Table 4.2: Classification accuracy (%) for single case experiments

Regions	HCMT-SCP	HCMT-MCP	SAMT-SCP	SAMT-MCP	SAMRT-SCP	SAMRT-MCP
All	42.3	36.6	42	34	**66.3**	59.6
1000-1800 cm^{-1}	46.3	33.3	45.3	33.3	**80.6**	63.3

Now, we look at the results for a single case with each method grade-wise to investigate which grade spectra are correctly classified and which were mis-classified as another

grade.

Table 4.3 shows grade-wise categorisation with the HCMT-SCP method. The correct classifications have been highlighted in bold. It can be observed that the method performed poorly. G-I was mostly classified as G-III, G-II was mostly predicted correctly but 24 spectra were incorrectly classified as G-III. In case of G-III, 63 spectra were classified correctly but the rest were classified as G-II. In summary, no spectra were classified as G-I. It indicated that G-II and G-III spectra were more close to each other where as G-I spectra were quite different from them.

Table 4.3: Grade wise categorisation with Single Case using HCMT-SCP with region 1000-1800 cm^{-1}

Actual Grade	Predicted Grade			
	G-I	G-II	G-III	Sum
G-I	**0**	4	96	100
G-II	0	**76**	24	100
G-III	0	37	**63**	100

Table 4.4 shows the results with the HCMT-MCP method. It can be observed that having multiple control points does not make any difference as grade categorisation is still poor. All the spectra are classified as G-III. Results showed that multiple control points were locally optimized thus indicating a major drawback of HC.

Table 4.4: Grade wise categorisation with Single Case with HCMT-MCP with region 1000-1800 cm^{-1}

Actual Grade	Predicted Grade			
	G-I	G-II	G-III	Sum
G-I	**0**	0	100	100
G-II	0	**0**	100	100
G-III	0	0	**100**	100

Table 4.5 shows results with the SAMT-SCP method. It can be observed that the majority of G-III were correctly classified and that there is a split between G-II and G-III for most of the spectra and no spectra are classified as G-I. These results are slightly better

in comparison with the HCMT-MCP method as G-II and G-III are classified correctly to some degree. As in HCMT-MCP, no spectra were classified as G-I.

Table 4.5: Grade wise categorisation with Single Case with SAMT-SCP with region 1000-1800 cm^{-1}

Actual Grade	Predicted Grade			
	G-I	G-II	G-III	Sum
G-I	**0**	4	96	100
G-II	0	**83**	17	100
G-III	0	47	**53**	100

Table 4.6 shows results with SAMT-MCP method. It can be seen that method did not perform well and in this case also, all spectra were classified as G-III.

Table 4.6: Grade wise categorisation with Single Case with SAMT-MCP with region 1000-1800 cm^{-1}

Actual Grade	Predicted Grade			
	G-I	G-II	G-III	Sum
G-I	**0**	0	100	100
G-II	0	**0**	100	100
G-III	0	0	**100**	100

Table 4.7 shows results with SAMRT-SCP method. It can be noted that in this method the rules are also changed along with membership functions. It can be observed that results are better than the HCMT and SAMT methods. In the case of G-I, 67 spectra were correctly classified, in the case of G-II 99 spectra were correctly classified where as in case of G-III, mostly spectra were correctly classified (76) and remaining were mis-classified as G-I (24). Overall this method produced the best result among all methods. It is much better than we would expect at random for all 3 grades.

Table 4.7: Grade wise categorisation with Single Case with SAMRT-SCP with region 1000-1800 cm^{-1}

Actual Grade	Predicted Grade			
	G-I	G-II	G-III	Sum
G-I	**67**	0	33	100
G-II	1	**99**	0	100
G-III	24	0	**76**	100

Table 4.8 shows results with SAMRT-MCP method. It can be observed that results were good for G-II, G-I and G-III are confused with one another.

Table 4.8: Grade wise categorisation with Single Case with SAMRT-MCP with region 1000-1800 cm^{-1}

Actual Grade	Predicted Grade			
	G-I	G-II	G-III	Sum
G-I	**57**	0	43	100
G-II	0	**96**	4	100
G-III	63	0	**37**	100

In summary, SAMRT-SCP was the best method among all the methods used in terms of classifying grades. It also shows that the rules initially set up for the system are not ideal, therefore, both the HCMT and SAMT methods performed poorly. When the rules were altered in SAMRT, results improved with both SCP and MCP. The change of rules also reduced the squared error indicating an improvement on overall grade prediction.

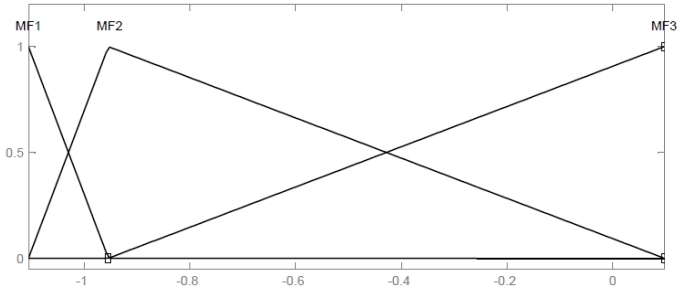
Table 4.9 shows the final rule set obtained for the trained FIS for the SAMRT-SCP method with region 1000-1800 cm^{-1}. It can be seen that these rules are quite different from the initial rules. For example, in rule-1 the consequent is 3 when all input MFs are 1 and it is contrary to original majority vote scheme where consequent was 1. Similar observations can be made for other rules as well. The consequents different from the initial consequent (IC) previously described in Table 4.1 have been highlighted in bold. Six rules (6, 8, 12, 16, 20, 22) were never assigned any specific consequent and they varied with rule changes. It shows that the majority vote method that has been used for HC and SA previously is not suited to this data set and rule changes method is the one

better suited to the data set.

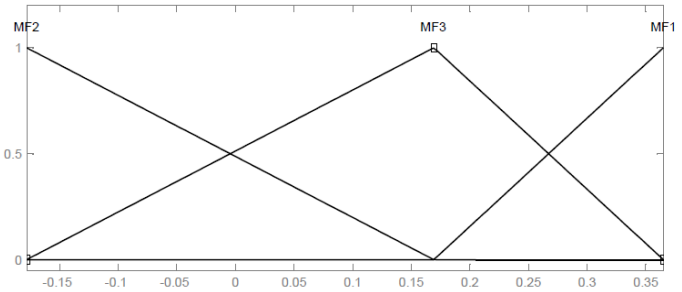
Table 4.9: Fuzzy rule set for FIS with SAMRT-SCP method for Single Case

Rule	MF1	MF2	MF3	IC	SAMRT-SCP Consequent
1	1	1	1	**1**	**3**
2	1	1	2	**1**	**3**
3	1	1	3	**1**	**2**
4	1	2	1	1	1
5	1	2	2	**2**	**3**
6	1	2	3	1/2/3	3
7	1	3	1	1	1
8	1	3	2	1/2/3	2
9	1	3	3	3	3
10	2	1	1	**1**	**2**
11	2	1	2	**2**	**3**
12	2	1	3	1/2/3	1
13	2	2	1	2	2
14	2	2	2	2	2
15	2	2	3	2	2
16	2	3	1	1/2/3	1
17	2	3	2	**2**	**3**
18	2	3	3	**3**	**1**
19	3	1	1	1	1
20	3	1	2	1/2/3	2
21	3	1	3	**3**	**2**
22	3	2	1	1/2/3	3
23	3	2	2	2	2
24	3	2	3	**3**	**2**
25	3	3	1	**3**	**1**
26	3	3	2	3	3
27	3	3	3	**3**	**2**

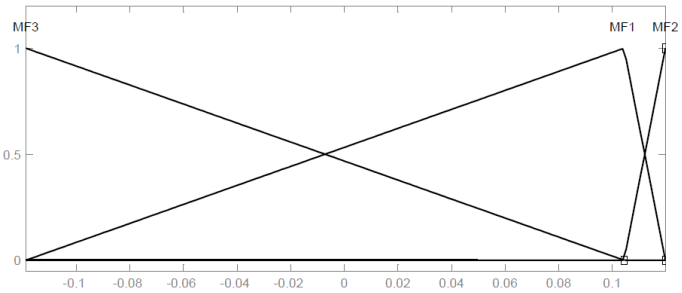
Figure 4.8 shows the final membership functions after training for all input variables. It can be seen that all membership functions have moved for all input variables in order to find the best possible membership functions parameters.



(a) Membership functions for input variable 1



(b) Membership functions for input variable 2



(c) Membership functions for input variable 3

Figure 4.8: Final membership functions for SAMRT-SCP for single case

4.7 Results with Multiple Case Experiments

In multiple case experiments where spectral data was obtained from all the known cases, 10 random initialisations were done for all the experiments as in the case single case experiments. The results obtained by all methods over the whole spectral region and 1000-1800 cm^{-1} spectral region are shown in the Table 4.10. The results are in terms of percentage accuracy calculated using Equation 4.3. It can be observed from the table that HCMT-SCP, HCMT-MCP, SAMT-SCP and SAMT-MCP performed very poorly and were not able to find a FIS with high accuracy, and the percentage accuracy remained very low, equivalent to a random guess. In case of SAMRT-SCP, the accuracy improved

to 51.6% which was not very good, but better than other methods. Similarly, SAMT-MCP resulted in percentage accuracy of 48.6%. In terms of regions, 1000-1800 cm^{-1} region results were found to be slightly better than the whole region. It indicates that this region, which is much smaller than the whole spectral region, can be used as a bench mark region for the experiments. The results with SAMT also showed that changing the rules and membership function parameters at the same time provides a better approach as this method tends to result in optimal membership functions as well as rules for the data set. It means that results are improved by this method which shows a convergence towards an optimal solution.

Table 4.10: Classification accuracy (%) for Multi case experiments

Regions	HCMT-SCP	HCMT-MCP	SAMT-SCP	SAMT-MCP	SAMRT-SCP	SAMRT-MCP
All	35	35.33	35	38.3	41.3	**45.3**
1000-1800 cm^{-1}	34.6	34.3	28.3	29.66	**51**	49.66

It can also be noted that in the case of multiple case experiments results were poorer than experiments with a single case. Although the single case was randomly chosen and was not representative of all the variations and uncertainties involved, it still indicates that adding cases adds considerable complexity to the data set.

Now, we look at the results for multiple cases with each method grade-wise as we did in the single case experiments.

Tables 4.11 and 4.12 show grade wise categorisation with the HCMT-SCP and HCMT-MCP methods. It can be observed that both methods performed poorly as with a single case.

Table 4.11: Grade wise categorisation with Multiple Case with HCMT-SCP with region 1000-1800 cm^{-1}

Actual Grade	Predicted Grade			
	G-I	G-II	G-III	Sum
G-I	**0**	39	61	100
G-II	0	**58**	46	104
G-III	0	50	**46**	96

Table 4.12: Grade wise categorisation with Multiple Case with HCMT-MCP with region 1000-1800 cm^{-1}

Actual Grade	Predicted Grade			
	G-I	G-II	G-III	Sum
G-I	**0**	33	67	100
G-II	0	**52**	52	104
G-III	0	45	**51**	96

Tables 4.13 and 4.14 show results with the SAMT-SCP and SAMT-MCP methods respectively. The results follow the same pattern as in the single case and no spectra are classified as G-I.

Table 4.13: Grade wise categorisation with Multiple Case with SAMT-SCP with region 1000-1800 cm^{-1}

Actual Grade	Predicted Grade			
	G-I	G-II	G-III	Sum
G-I	**0**	74	26	100
G-II	0	**37**	67	104
G-III	0	48	**48**	96

Table 4.14: Grade wise categorisation with Multiple Case with SAMT-MCP with region 1000-1800 cm^{-1}

Actual Grade	Predicted Grade			
	G-I	G-II	G-III	Sum
G-I	**0**	62	38	100
G-II	0	**28**	76	104
G-III	0	35	**61**	96

Tables 4.15 and 4.16 show results with the SAMRT-SCP and SAMRT-MCP methods. It can be seen that changing the rules produce better results. In both methods, a high percentage of G-I spectra are classified correctly. For G-II, most of spectra are classified as G-II or G-III. G-III spectra are the worst of all and are often confused with G-I. In general, SAMRT-SCP performed slightly better than SAMRT-MCP. Results also show that the initial rules set up for the system are not suitable and the altered rules provide better results when compared to the static rules.

Table 4.15: Grade wise categorisation with Multiple Case with SAMRT-SCP with region 1000-1800 cm^{-1}

Actual Grade	Predicted Grade			
	G-I	G-II	G-III	Sum
G-I	**71**	20	9	100
G-II	28	**73**	3	104
G-III	40	47	**9**	96

Table 4.16: Grade wise categorisation with Multiple Case with SAMRT-MCP with region 1000-1800 cm^{-1}

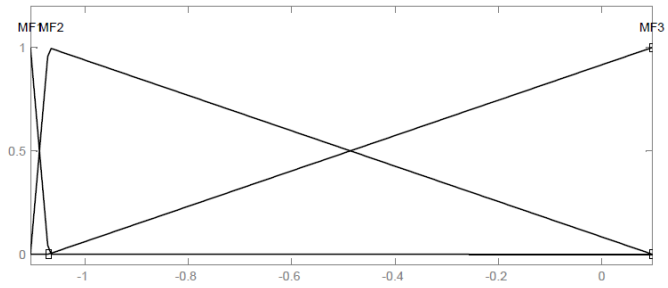
Actual Grade	Predicted Grade			
	G-I	G-II	G-III	Sum
G-I	**64**	1	35	100
G-II	38	**48**	18	104
G-III	43	16	**37**	96

Table 4.17 shows the final rule set obtained after training of FIS with SAMRT-SCP and its comparison with the initial consequents (IC) as described in Table 4.1. It can be observed that rules consequents have changed from the initial rule set. For example, in rule 2, the consequent has changed from 1 to 3, in rule 9 the consequent is 2 where as the input memberships are 1 and 3. Rules whose consequents have changed have been highlighted in bold. Six rules (6, 8, 12, 16, 20, 22) were always flexible and their best consequent has been found with the method as seen in the table. It indicates that changing rules works better as it improves the results and the majority vote criteria was not a good choice for the system.

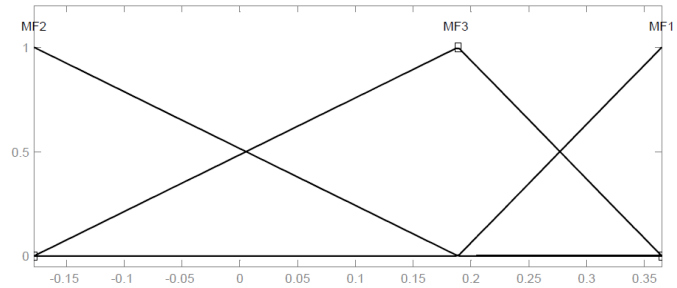
Table 4.17: Fuzzy rule set for FIS with SAMRT-SCP method for Multiple Case

Rule	MF1	MF2	MF3	IC	SAMRT-SCP Consequent
1	1	1	1	1	1
2	1	1	2	**1**	**3**
3	1	1	3	1	1
4	1	2	1	**1**	**2**
5	1	2	2	**2**	**1**
6	1	2	3	1/2/3	1
7	1	3	1	**1**	**3**
8	1	3	2	1/2/3	2
9	1	3	3	**3**	**2**
10	2	1	1	**1**	**2**
11	2	1	2	2	2
12	2	1	3	1/2/3	1
13	2	2	1	2	2
14	2	2	2	**2**	**3**
15	2	2	3	2	2
16	2	3	1	1/2/3	2
17	2	3	2	**2**	**3**
18	2	3	3	**3**	**2**
19	3	1	1	3	3
20	3	1	2	1/2/3	2
21	3	1	3	**3**	**1**
22	3	2	1	1/2/3	2
23	3	2	2	**2**	**1**
24	3	2	3	**3**	**1**
25	3	3	1	**3**	**2**
26	3	3	2	3	3
27	3	3	3	**3**	**1**

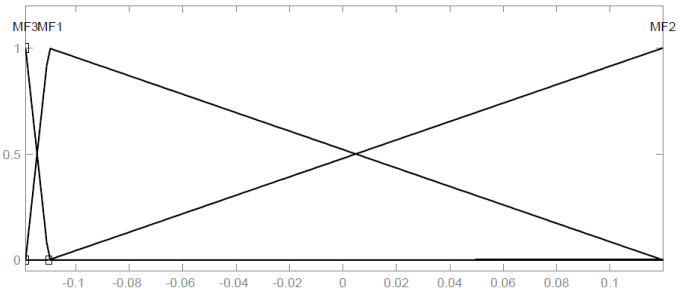
Figure 4.9 shows the final membership functions after training with SAMRT-SCP. It can be observed that all three membership functions for the three input variables have moved around the input range.



(a) Membership functions for input variable 1



(b) Membership functions for input variable 2



(c) Membership functions for input variable 3

Figure 4.9: Final membership functions for SAMRT-SCP for Multi case

The results also indicate that results with HC algorithm were always poor. The reason might be that it stuck to a local minima and because of it, the solution never improved which means that squared error did not reduce. Therefore, we did not use rule tuning with HC as well as it was expected to do the same with rules as well.

We conducted another series of experiments starting with taking spectra from a single case and then by adding cases step by step to see the effect of adding cases and the complexity involved in it. These experiments were conducted with SAMRT-SCP method as it was found to provide better results both in case of single case and multi case experiments. The region used was 1000-1800 cm^{-1} which we consider as a bench mark region.

Table 4.18 shows results with the step wise approach with the SAMRT-SCP method.

Table 4.18: Classification accuracy percentage for step wise cases with SAMRT-SCP over 1000-1800 cm^{-1} region

Exp No	G-I cases	G-II cases	G-III cases	Spectra for each case	Total spectra	Best Summation	Percentage accuracy
1	1	1	1	G-I:100,G-II:100,G-III:100	300	242	80.6
2	2	2	2	G-I:50,G-II:50,G-III:50	300	158	52.6
3	2	4	4	G-I:50,G-II:25,G-III:25	300	181	60.3
4	2	8	6	G-I:50,G-II:12,G-III:16	292	165	56.5
5	2	12	6	G-I:50,G-II:8,G-III:16	292	150	51.3
6	2	16	6	G-I:50,G-II:6,G-III:16	292	159	53
7	2	20	6	G-I:50,G-II:5,G-III:16	300	146	48.6
8	2	26	6	G-I:50,G-II:4,G-III:16	300	155	51.6

The column *Spectra for each case* indicates the number of spectra selected from each case as we have two cases of G-I, 26 cases of G-II and six cases of G-III. The *Total Spectra* column is the total sum of all spectra from all cases for a set of experiments. The number varies for experiments as we wanted to take an equal number of spectra from each case of a particular grade to keep a balanced data set. The final column is the *Percentage accuracy* over 300 spectra calculated as defined in Equation 4.3. It can be observed from the table that as the number of cases of G-II starts to increase, the uncertainty of the results also start to increase. In experiment 3 where 4 cases of G-II are involved, the percentage accuracy is 60.3% where as from experiments 4-8 it does not exceed 51.6%. It is also worth mentioning that from experiments 4-8, the cases of G-I and G-III are fixed to 2 and 6 respectively as this is the maximum number of cases available for these grades. The results indicate that there is a lot of difference between cases of the same grade and grades are not easily separable. Especially in case of G-II when more cases are added, it seems that they are quite different from each other and this confuses the FIS during training and does not provide a good solution. Figure 4.10 shows a 3d scatter plot of three PCs for three grades. It can be observed the three PCs are not able to provide a clear distinction between grades and all grades are overlapping.

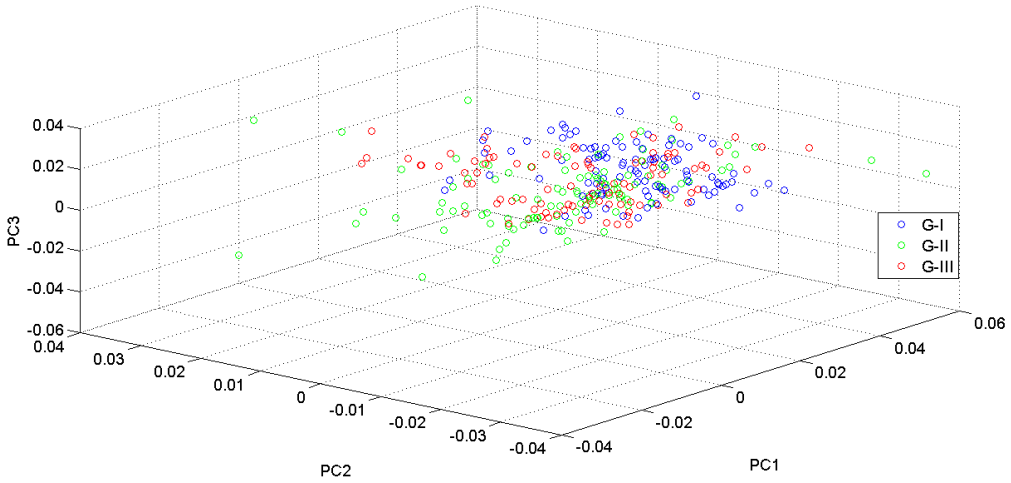


Figure 4.10: 3D Scatter plot of three PCs for all grades

From our experiments, the following observations can be made.

- A particular FIS setting performed well with a single case of each grade
- With data from all cases, clear distinction between grades was not achieved
- There is the possibility of more variance in the data within cases of same grade than between grades

To have more confidence in our results, we also used the k-means clustering algorithm as an example with three PCs to make a comparison. We have shown results with both k-means and FCM with 10 PCs in Chapter 3.

4.8 Results with k-means Clustering

In these experiments, the same number of spectra for each grade were used as in the multiple case experiments i.e, for G-I (100 spectra), for G-II (104 spectra) and for G-III (96 spectra) using three PCs. The number of clusters was set to three, one for each grade. The algorithm was repeated with 10 different start positions. Table 4.19 shows the results of the k-means experiments. It shows three clusters with the number of spectra from each grade in each cluster. It is evident from the Table 4.19 that k-means is not able to create 3 distinct clusters that match the grades. The clusters contain a mix of spectra from all grades.

Table 4.19: k-means clustering results with three clusters

Clusters with members	G-I	G-II	G-III
1(83)	4	47	32
2(111)	46	24	41
3(106)	50	33	23

We also conducted experiments with k-means by setting number of clusters to two to further investigate the matter. In these experiments, the same data was used as in case of

previous k-means experiments i.e, for G-I (100 spectra), for G-II (104 spectra) and for G-III (96 spectra) using three PCs. The results are shown in Table 4.20 . It can be seen that cluster 1 only includes eight members of G-I and is mainly comprised of members from G-II and G-III where as cluster 2 has nearly all members of G-I and about half of G-II and G-III. It also suggests that there is more uncertainty found within cases of the same grades especially G-II and G-III which are adding complexity as both with two and three clusters, G-II and G-III are never predicted clearly where as G-I prediction remained better. The k-means clustering algorithm is not able to find an optimal classification of the grades.

Table 4.20: k-means clustering results with 2 clusters

Clusters with members	G-I	G-II	G-III
1(109)	8	62	39
2(191)	92	42	57

4.9 Summary

In this chapter, we developed a Mamdani type fuzzy inferencing system with triangular type-I fuzzy membership functions. We trained membership functions using HCMT, SAMT and both rules and membership functions using SAMRT for the complex data set 3. Two types of control points SCP and MCP were defined for membership functions. Experiments were performed on the whole region and region between 1000-1800 cm^{-1}. The latter region was selected and SAMRT-SCP was found to provide the best results of all methods, though none of the methods used was able to find a clear distinction between cancer grades. K-means clustering was also used with the same number of PCs as used for the FIS to make a comparison of the results in terms of correctly classified grades. It too performed poorly on our data set. It has been concluded that these methods are not able to find a useful classification of breast cancer grading. This may be because more variability and uncertainty found within spectra from the case of the same grade (intra-case vari-

ability) than spectra from different cases of the same grade (inter-case variability). The evidence is provided by the experiment in which cases are added gradually that ultimately results in poor classification. There is a need to find advanced computational methods that can deal with these types of uncertainties. The next chapter introduces a model based on T-II fuzzy logic in order to classify the cancer grades as T-II fuzzy logic in general has been found to work well when there is more uncertainty in a data set.

Chapter 5

Experiments with Type-II Fuzzy Model

This chapter introduces a step wise approach to create zSlices based General Type-II fuzzy sets (zGT-II) from spectral data by selecting a number of features from a data set. The features extracted from the data will be used as interval data for the model. The aim of the model is to classify breast cancer grade while accounting for the complex uncertainties found in the data. We also test the model with unseen data with different configurations and discuss the results obtained by the model.

5.1 Model Structure

The aim of the proposed methodology is to investigate the use of zGT-II fuzzy sets created using interval data representing features extracted from spectral data. T-II fuzzy sets have been shown to do well when there is more uncertainty involved [38, 72, 74]. In the case of our data set, we have two types of uncertainty, one within the spectra of a case of a grade (intra-case) and other when comparing it with other cases of same grade (inter-case). The proposed model aims to cover these uncertainties by creating a zGT-II set involving both spectra from same case of a grade and spectra from different cases of a grade with interval data covering both types of variabilities. Figure 5.1 shows a broad view of the model from creation to testing. We describe each of these stages of the model in detail in the following sections.

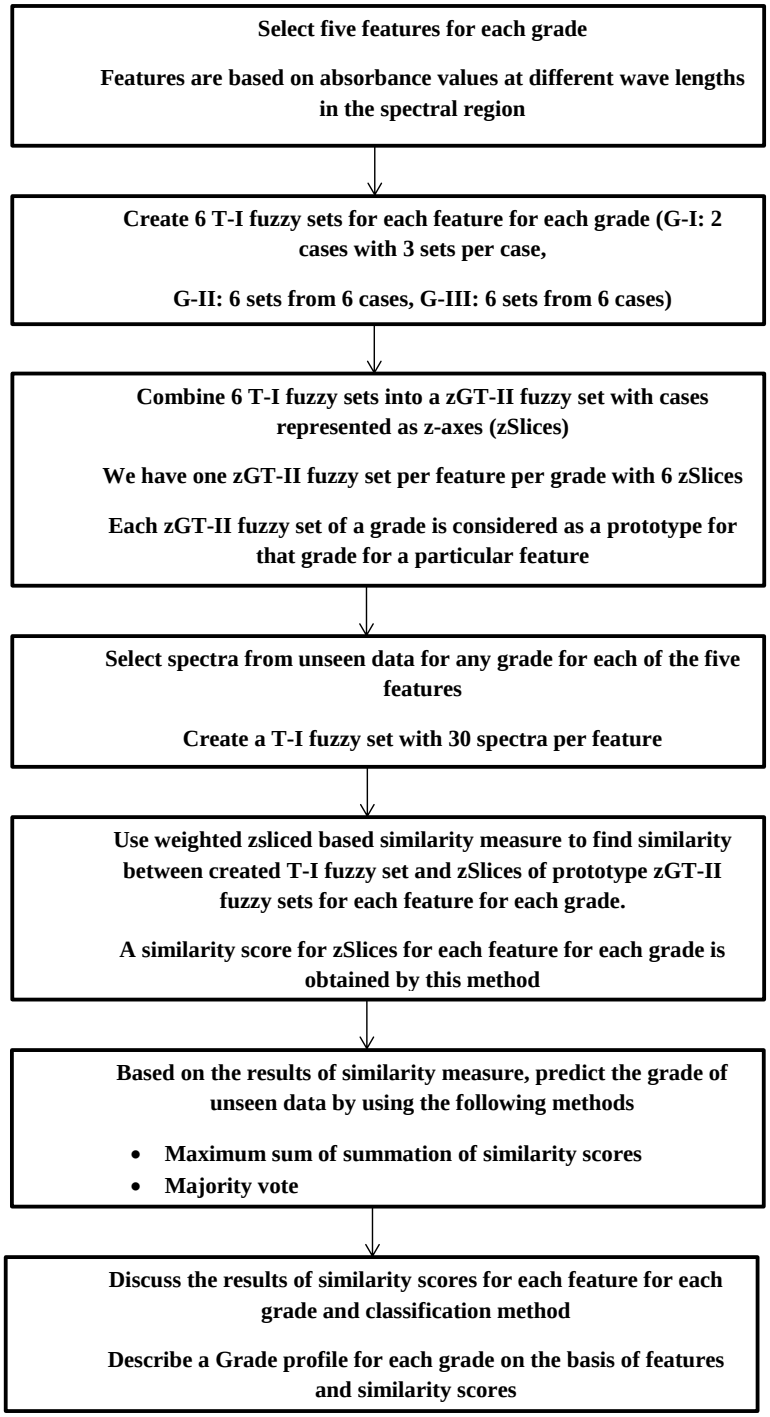


Figure 5.1: Block diagram of the model structure

5.2 Features Extraction from Spectral Regions

Initially, five features have been selected for the experiments. The region 1000-1800 cm^{-1} was divided into three sections in line with Chiu et al's [20] division of the region. A slight modification from Chiu et al is that our spectral range starts from 1000 cm^{-1} instead of 950 cm^{-1} and ends at 1800 cm^{-1} instead of 1780 cm^{-1} because of the availability of the processed data. The regions are described in Table 5.1.

Table 5.1: Regions of Features with Spectral Range

Regions	Spectral Range
A	1000-1350 cm^{-1}
B	1350-1480 cm^{-1}
C	1480-1800 cm^{-1}

For region A, three features have been selected in order to define interval data. The regions and approximate location of the areas covered by the features are shown in Figure 5.2. The bar in the figure indicates the approximate area covered by each feature. It is worth mentioning that these features have been selected based on absorbance values of spectra at certain peak heights and troughs at different wave lengths and no other specific criteria has been used. Intervals representing features have been extracted from each spectrum. For example, for feature 1, minimum absorbance value A and maximum absorbance value B are combined to create an interval (A, B) as shown in Figure 5.2. For feature 5, two distinct peak heights have been used to create an interval. A number of other features comprising of various other combinations of peak heights and troughs is possible, though not considered for this initial study. The description of the selected features is as follows.

Feature 1: This feature consists of minimum absorbance values from the region 1000-1020 cm^{-1} and maximum peak spectral absorbance in the region 1080-1100 cm^{-1}. We have selected this feature to cover the highest distinct left peak available in the region with the left most negative peak or trough in the region.

Feature 2: This feature consists of the maximum peak height absorbance values in the

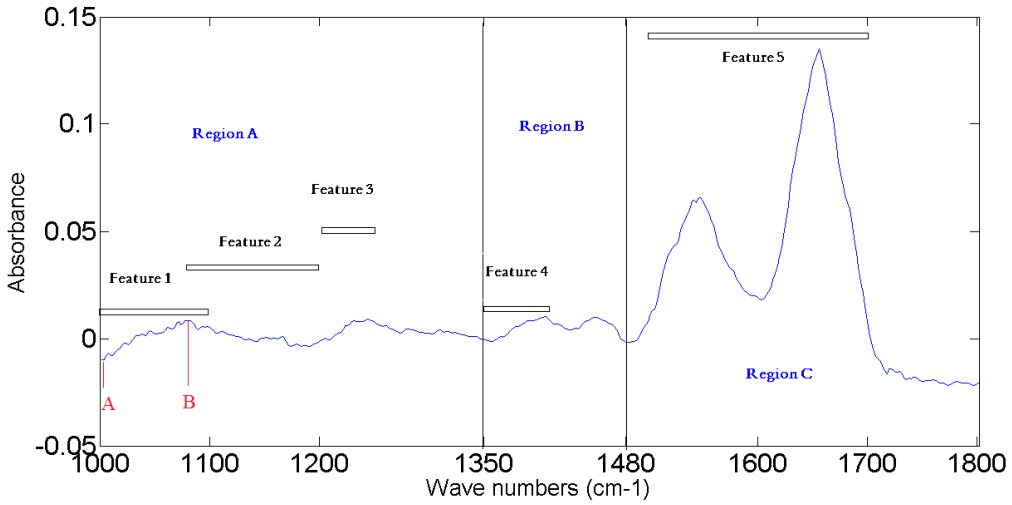


Figure 5.2: Regions and approximate locations of selected features

region 1080-1100 cm^{-1} and minimum absorbance values in the region 1180-1200 cm^{-1}. The aim is to cover maximum left side peak in the region and right side lowest trough value in the region.

For features 1 and 2, the peak height value is the same. The aim is to see how these two features whose one value is common respond in the model.

Feature 3: This feature consists of minimum absorbance values in the region 1200-1220 cm^{-1} and maximum peak absorbance values of the region 1220-1250 cm^{-1}. The feature was selected to cover second main peak in the region 1200-1300 cm^{-1}.

For region B, one feature has been selected.

Feature 4: This feature consists of minimum absorbance values in the region 1350-1400 cm^{-1} and the maximum peak absorbance values in the region 1400-1410 cm^{-1}.

For region C, one feature has been selected.

Feature 5: This feature consists of peak heights of Amide-I and Amide-II region as interval data. This feature has been selected to cover the two most distinct peaks in the region. Amide-II peak height is the maximum peak absorbance values in the region 1500-1600 cm^{-1} and Amide-I is the maximum peak height absorbance values in the region 1600-1700 cm^{-1}.

5.3 Construction of Type-I Fuzzy Sets from Features

The next stage of the model is to construct T-I fuzzy sets from the interval data. A combination of the minimum and maximum absorbance values of the selected features are used to create an interval. We have initially selected 30 spectra to create a T-I fuzzy set. As there are 30 values for each set, the primary membership domain is divided into 30 sections ranging from 1/30 to 30/30. As we have two cases from G-I, 26 cases from G-II and six cases of G-III, we have decided to create six T-I fuzzy sets for each grade per feature from these cases. For G-I, three regions from two cases have been selected making it six sets in accordance with other grades. The hierarchy used for construction of the fuzzy sets for G-I is shown in Figure 5.3. It can be observed from the figure that three sets for each of the two cases have been used to construct six fuzzy sets for each feature. The steps used to create the sets for G-II and G-III can be seen in Figure 5.4. For G-II, a random selection of six cases out of 26 has been made and for G-III, spectra from all six cases have been included. In this way we have six sets of 30 spectra (in terms of interval data) from each grade per feature. All five features use the same set of spectra. Each grade has 30 sets for five features. In total we have 90 T-I sets for all grades.

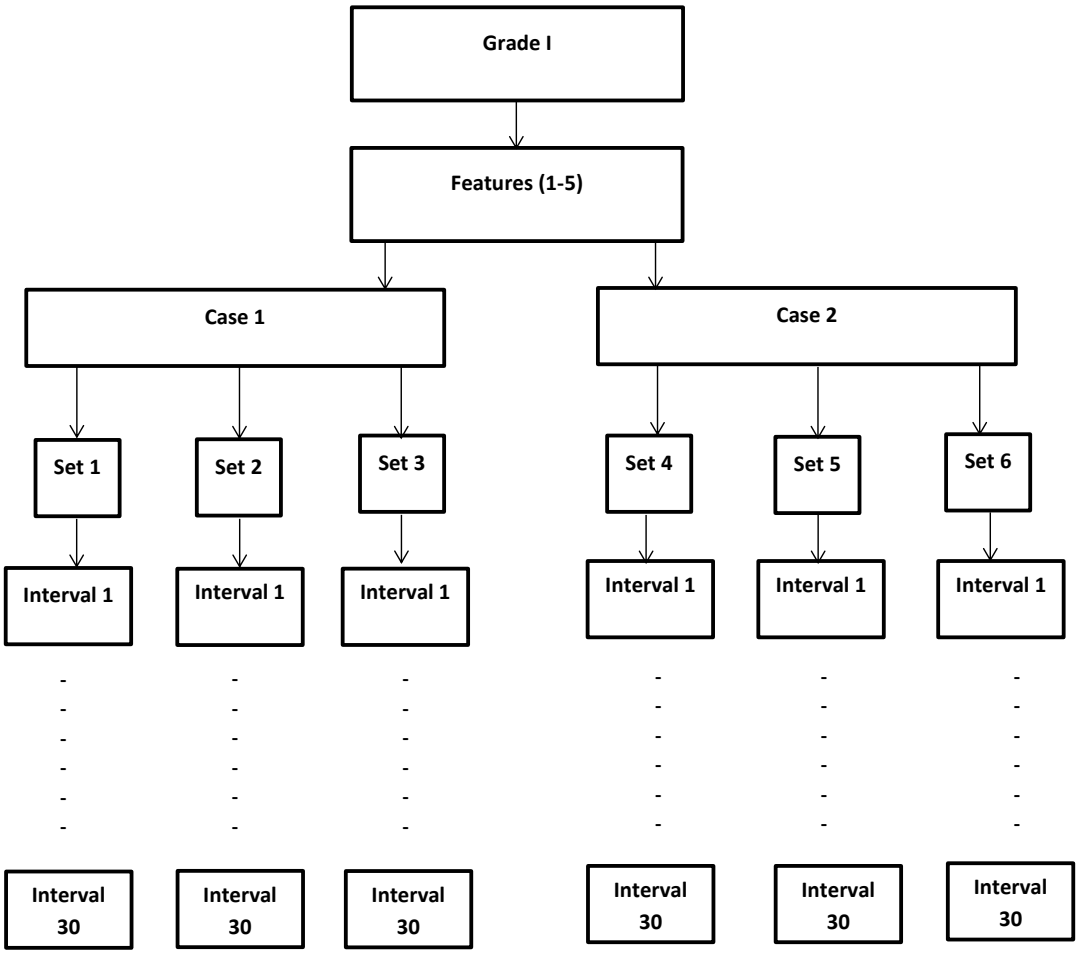


Figure 5.3: Block diagram of construction of fuzzy sets for G-I

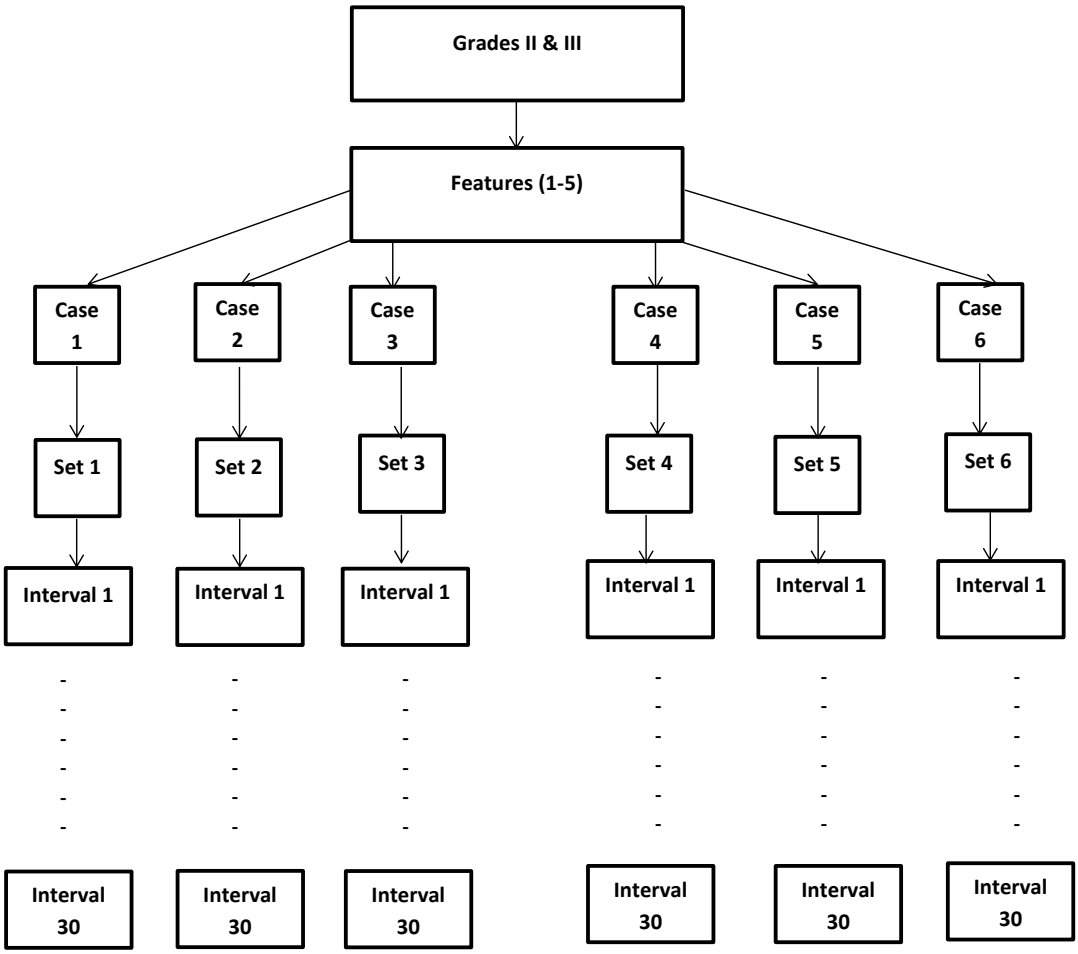


Figure 5.4: Block diagram of construction of fuzzy sets for G-II and G-III

The method used to create T-I fuzzy sets is the one described by Miller et al. [76] and explained in Chapter 2. The method was implemented using a script written in Matlab (version 7.02). As this method compares a number of combinations of values, as the number of input values increases, the number of combinations to be compared increases exponentially. While using spectral values as interval data, we observed that as number of spectra increases beyond 20, the computational time increases rapidly. To understand the scenario, we selected 20 random spectra and created a T-I fuzzy set with feature 1 and calculated the computational time. Figure 5.5 shows the computational time with

the increase in number of spectra for the original method for the creation of T-I fuzzy sets from 15 to 20 spectra. It can be seen that as the number of spectra reach 20, the time changes exponentially and it becomes computationally very expensive to create fuzzy sets beyond that as the number of combinations to be compared increases and system runs out of memory to handle such massive data combinations. To overcome this issue, the author of this thesis has proposed a method that gives an approximation of the original method while substantially reducing the time. The next subsection describes the proposed method in detail.

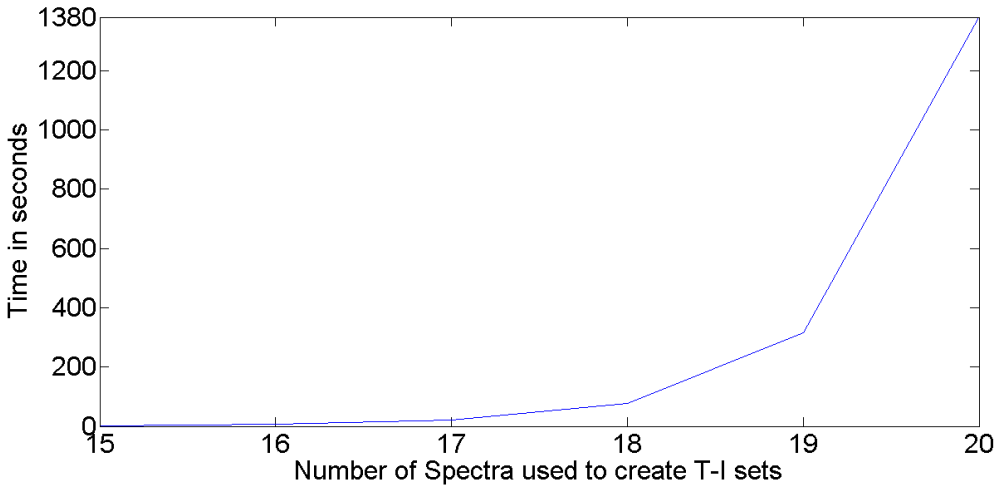


Figure 5.5: Processing time for methods shown in Miller et al. [76] with increasing spectra (intervals)

5.3.1 An Approximate Method to Create Fuzzy Sets from Interval data

Let $(x_1, y_1), ..., (x_n, y_n)$ be the interval data set to be used to create T-I fuzzy set. The method is as follows.

Step 1: Sort the (left) first elements of each interval in ascending order and store them as $(a_1 : a_n) = sort(ascend(x_1 .....x_n))$

Step 2: Sort the (right) second elements of each interval in descending order and save

them as

$(b_1 : b_n) = sort(descend(y_1.....y_n))$

Step 3: If any $(a \geq = b)$ then delete that entry of (a,b), this eliminates any empty intervals that result in $NULL$ with the original method intervals

Step 4: Combine both results as interval data and save them as final result=$(a_1,b_1),.....,(a_n,b_n)$

This algorithm is an approximation of the original algorithm in many cases giving identical result. The major advantage of the proposed method is that it is much faster than the original method making it practical in real world applications where high dimensional data of large volume needs to be processed in computationally efficient manner. Now we consider a few examples based on synthetic data to create T-I fuzzy sets with the proposed method and compare the results with the original method.

5.3.2 Examples from Synthetic Data

To explain the method, we consider some examples based on synthetic data. As we are dealing with interval data, it is possible that input data can be either completely overlapping, partially overlapping or mainly non-overlapping. Completely overlapping means that all intervals have some common values, partially overlapping means that some entries will not have any common values and non-overlapping means that all entries will be isolated and nothing will be common between them. We consider all of these scenarios and provide the details of the method.

First we take an example of interval data that is completely overlapping. Table 5.2 shows a set of three overlapping interval data. Figure 5.6 shows that all three entries of data have some area common in them. We apply both the original method and our proposed method on the data.

Table 5.2: Example of completely overlapping data

Interval Number	Interval data
1	(3,6)
2	(2,7)
3	(4,8)

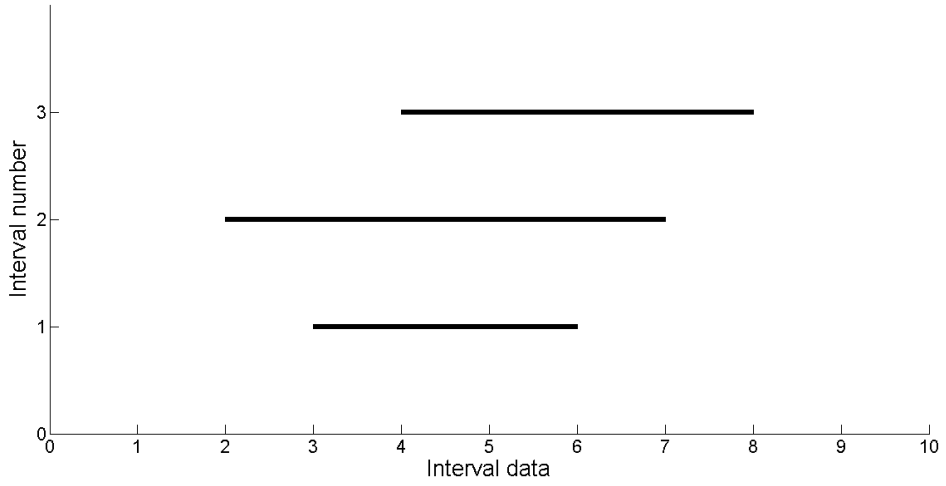


Figure 5.6: Plot of completely overlapping data

First we apply the original method described in Chapter 2 using Equation 2.27 for T-I fuzzy set creation.

$\mu(A) = y1/([3,6] \cup [2,7] \cup [4,8]) + y2/(([3,6] \cap [2,7]) \cup ([3,6] \cap [4,8]) \cup ([2,7] \cap [4,8])) +$
$y3/([3,6] \cap [2,7] \cap [4,8])$

$= y1/[2,8] + y2/[3,7] + y3/[4,6]$

Now we apply our proposed method.

Step 1: Sort $[3,2,4]$ in ascending order = $[2,3,4]$

Step 2: Sort $[6,7,8]$ in descending order = $[8,7,6]$

Step 3: Check if ($2 \geq 8$ or $3 \geq 7$ or $4 \geq 6$)= $Null$

Step 4: Combine both sorting results = $[2,8], [3,7], [4,6]$

The results obtained by both the methods are shown in Table 5.3. It can be seen

that both methods have produced equal results in case of completely overlapping data. It indicates that the proposed method is working well when all intervals overlap. Figure 5.7 shows the created T-I fuzzy set by applying both methods. The x-axis shows the domain values for the fuzzy set and y-axis shows the membership grade values.

Table 5.3: Result of overlapping data

Interval Number	Original Method	Proposed Method
1	(2, 8)	(2, 8)
2	(3, 7)	(3, 7)
3	(4, 6)	(4, 6)

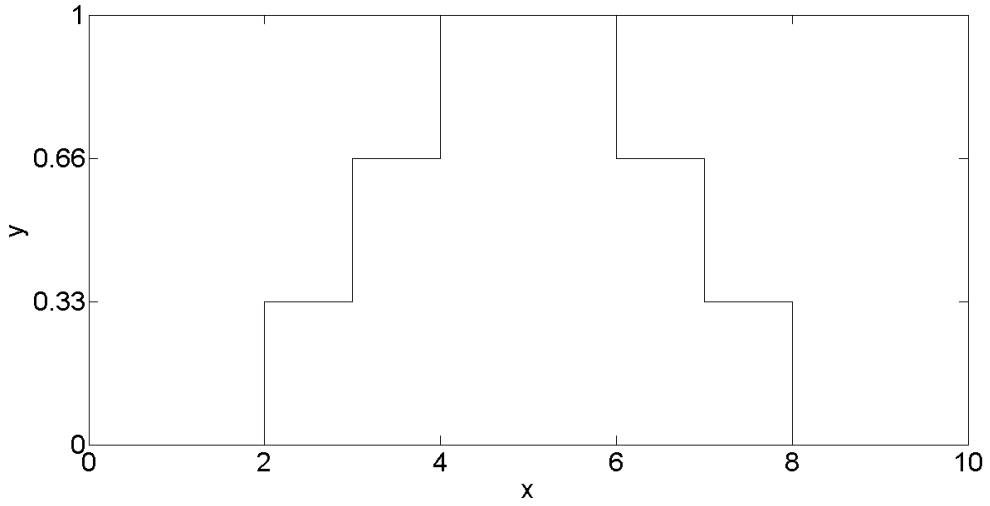


Figure 5.7: T-I fuzzy set for overlapping data

Now, we take an example of data where some of the entries do not overlap. Table 5.4 shows the data where interval one is completely non-overlapping while intervals 2 and 3 overlap. It can also be seen in Figure 5.8.

Table 5.4: Example of partially overlapping data

Interval Number	Interval data
1	(1, 3)
2	(5, 9)
3	(4, 8)

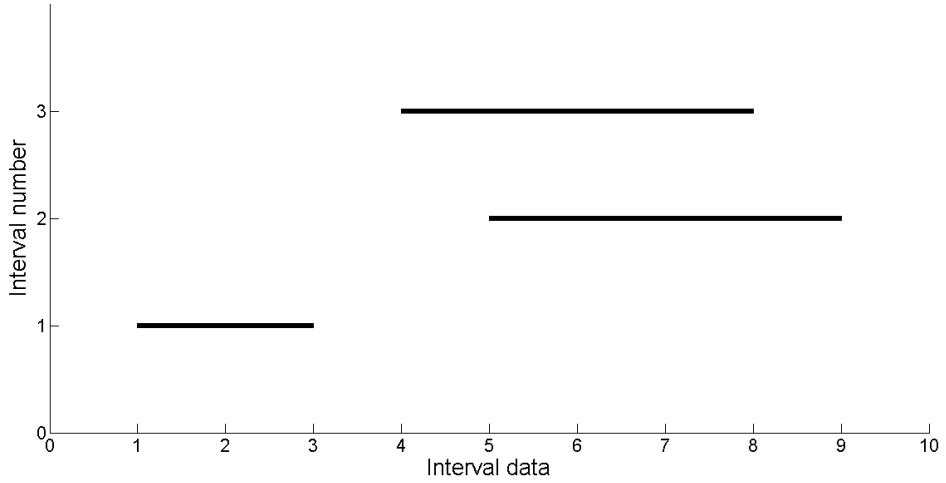


Figure 5.8: Plot of partially overlapping data

Firstly, we apply the original method.

$\mu(A) = y1/([1,3] \cup [5,9] \cup [4,8]) + y2/(([1,3] \cap [5,9]) \cup ([1,3] \cap [4,8]) \cup ([5,9] \cap [4,8])) +$
$y3/([1,3] \cap [4,8] \cap [5,9])$

$=y1/[1,3],[4,9] + y2/[5,8] + y3/NULL$

Now we apply our proposed method.

Step 1: Sort $[1,5,4]$ in ascending order = $[1,4,5]$

Step 2: Sort $[3,9,8]$ in descending order = $[9,8,3]$

Step 3: Check if ($1 \geq 9$ or $4 \geq 8$ or $5 \geq 3$)= delete $[5,3]$

Step 4: Combine both sorting results = $[1,9],[4,8]$

The result of the original and proposed method are shown in Table 5.5. It can be seen that the result is a sub-normal fuzzy set. Both methods are behaving differently in this example and this shows that if the interval data involves non-overlapping data then the proposed method does not provide the result equivalent to the original method.

Table 5.5: Result of data with some non-overlapped entries

Interval Number	Original Method	Proposed Method
1	(1, 3),(4, 9)	(1, 9)
2	(5, 8)	(4,8)
3	-	-

To further investigate a complex situation, we take an example of interval data where all of the data is non-overlapping. Table 5.6 shows the data. The plot of non-overlapping data is shown in Figure 5.9. It is evident from the figure that all three entries have no common area between them. Table 5.7 shows a comparison of result of the two methods.

Table 5.6: Example of completely non-overlapping data

Interval Number	Interval data
1	(2,3)
2	(6,7)
3	(4,5)

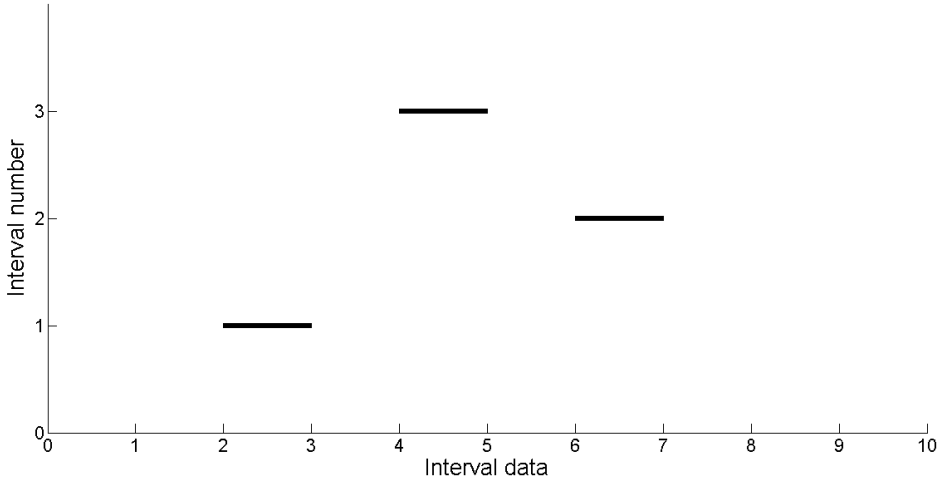


Figure 5.9: Plot of completely non-overlapping data

Firstly, we apply the original method.

$$\mu(A) = y1/([2,3] \cup [6,7] \cup [4,5]) + y2/(([2,3] \cap [6,7]) \cup ([2,3] \cap [4,5]) \cup ([6,7] \cap [4,5])) +$$

$y3/([2,3] \cap [6,7] \cap [4,5])$

$=y1/[2,3],[4,5],[6,7] + y2/NULL + y3/NULL$

Now we apply our proposed method.

Step 1: Sort $[2,6,4]$ in ascending order = $[2,4,6]$

Step 2: Sort $[3,7,5]$ in descending order = $[7,5,3]$

Step 3: Check if ($2 \geq 7$ or $4 \geq 5$ or $6 \geq 3$)= delete $[6,3]$

Step 4: Combine both sorting results = $[2,7],[4,5]$

The results of proposed and original methods are shown in Table 5.7. It can be seen that the proposed method results are completely different from the original method. It indicates that the proposed method is not suitable for completely non-overlapping data.

Table 5.7: Result of completely non-overlapping data

Interval data Number	Original Method	Proposed Method
1	(2,3),(4,5),(6,7)	(2,7)
2	-	(4,5)

The examples for overlapping, partial overlapping and non-overlapping synthetic data are shown to illustrate the working of the proposed method. Further experiments were conducted with more higher order synthetic data for three scenarios and on the data set used by Miller et al. [76] to have more confidence in the method. The results showed that the proposed method produces an acceptable approximation when data is completely overlapping. In case of non-overlapping data, different scenarios produce different results. As we are dealing with spectral data within a certain region or wave numbers, the chances are minimal that the data will be completely non-overlapped. We believe that for the majority of the time our proposed method will provide near equivalent results.

In the next section we present examples from the spectral data set.

5.3.3 Examples from Real Spectral Data

A main advantage of the proposed method is that it is computationally efficient when data is overlapping and provides near equivalent results when compared to the original method. We consider an example based on interval data extracted from G-I feature 1. We create T-I fuzzy sets varying from 15-20 spectra and compare the results to those produced with the original method with computation times. As an example, the interval data for 20 spectra is shown in Table 5.8. A fuzzy set created by the method is shown in Figure 5.10. The computational time for both methods varying from 15 to 20 spectra is shown in Table 5.9 and plotted in Figure 5.11. It can be observed that both methods provide equal results but in case of the proposed method, the computational time is significantly less. This shows that the proposed method can be used for higher order spectral data relatively easily.

Table 5.8: Comparison of result on Real spectral data

Interval Number	Interval data	Original Method	Proposed Method
1	(-0.0103, 0.0084)	(-0.0139, 0.0102)	(-0.0139, 0.0102)
2	(-0.0100, 0.0085)	(-0.0119, 0.0096)	(-0.0119, 0.0096)
3	(-0.0102, 0.0078)	(-0.0119, 0.0096)	(-0.0119, 0.0096)
4	(-0.0099, 0.0084)	(-0.0115, 0.0096)	(-0.0115, 0.0096)
5	(-0.0115, 0.0080)	(-0.0113, -0.0088)	(-0.0113, -0.0088)
6	(-0.0113, 0.0084)	(-0.0107, 0.0087)	(-0.0107, 0.0087)
7	(-0.0119, 0.0096)	(-0.0106, 0.0086)	(-0.0106, 0.0086)
8	(-0.0106, 0.0102)	(-0.0105, 0.0085)	(-0.0105, 0.0085)
9	(-0.0107, 0.0088)	(-0.0103 ,0.0084)	(-0.0103, 0.0084)
10	(-0.0097, 0.0074)	(-0.0103, 0.0084)	(-0.0103, 0.0084)
11	(-0.0105, 0.0087)	(-0.0103,0.0084)	(-0.0103, 0.0084)
12	(-0.0103, 0.0096)	(-0.0102, 0.0080)	(-0.0102, 0.0080)
13	(-0.0097, 0.0079)	(-0.0102, 0.0079)	(-0.0102, 0.0079)
14	(-0.0095, 0.0078)	(-0.0100, 0.0078)	(-0.0100, 0.0078)
15	(-0.0102, 0.0071)	(-0.0100, 0.0078)	(-0.0100, 0.0078)
16	(-0.0103, 0.0068)	(-0.0099, 0.0075)	(-0.0099, 0.0075)
17	(-0.0119, 0.0096)	(-0.0097, 0.0074)	(-0.0097, 0.0074)
18	(-0.0139, 0.0073)	(-0.0097, 0.0073)	(-0.0097, 0.0073)
19	(-0.0100, 0.0075)	(-0.0097, 0.0071)	(-0.0097, 0.0071)
20	(-0.0097, 0.0086)	(-0.0095, 0.0068)	(-0.0095, 0.0068)

Table 5.9: Comparison of computational time (in seconds)

Number of Spectra	Original method	Proposed method
15	1.651	0.950
16	5.692	0.960
17	22.096	0.970
18	75.875	0.980
19	314.656	0.990
20	1380.000	1.000

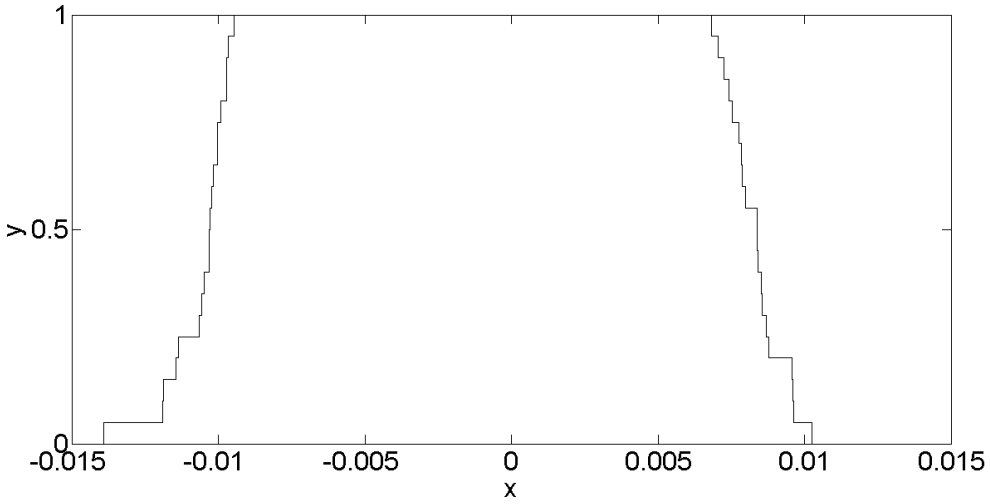


Figure 5.10: Fuzzy set for 20 spectra example

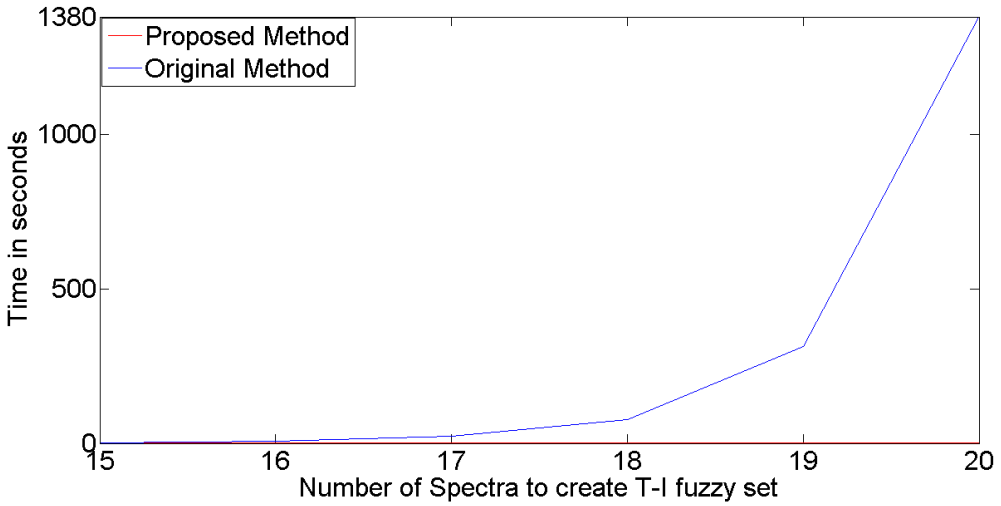


Figure 5.11: Computational time comparison

Figure 5.12 shows an example of T-I set created by the proposed method. This example has 100 spectral interval data set taken from feature 3 for case 1 of G-I. The computational time for set creation was 2.500 seconds. A comparison with the original method was not possible for these experiments as the computational power of the system (Core-i3 processor, 4GB RAM , 2.2 GHz) was not sufficient to handle the extraordinary number of comparisons required. Throughout this thesis, we assume that the proposed method provides a near equal approximation of the original method. In our examples with real spectral data, the results have been exactly the same, however, they may differ for different scenarios and confirmation is not practical because of computational limitations. The evidence indicates that the results generated by the proposed method is a practical method of achieving a close approximation of the original method.

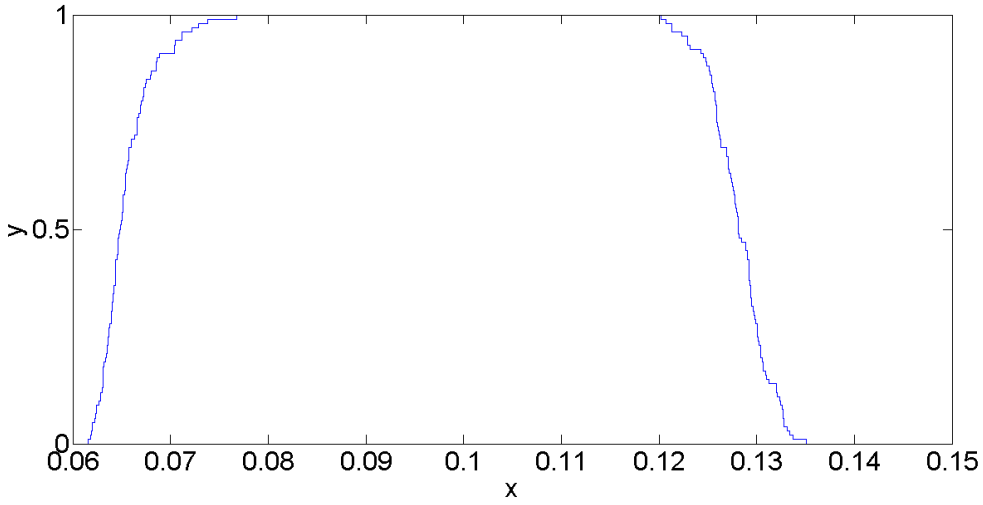
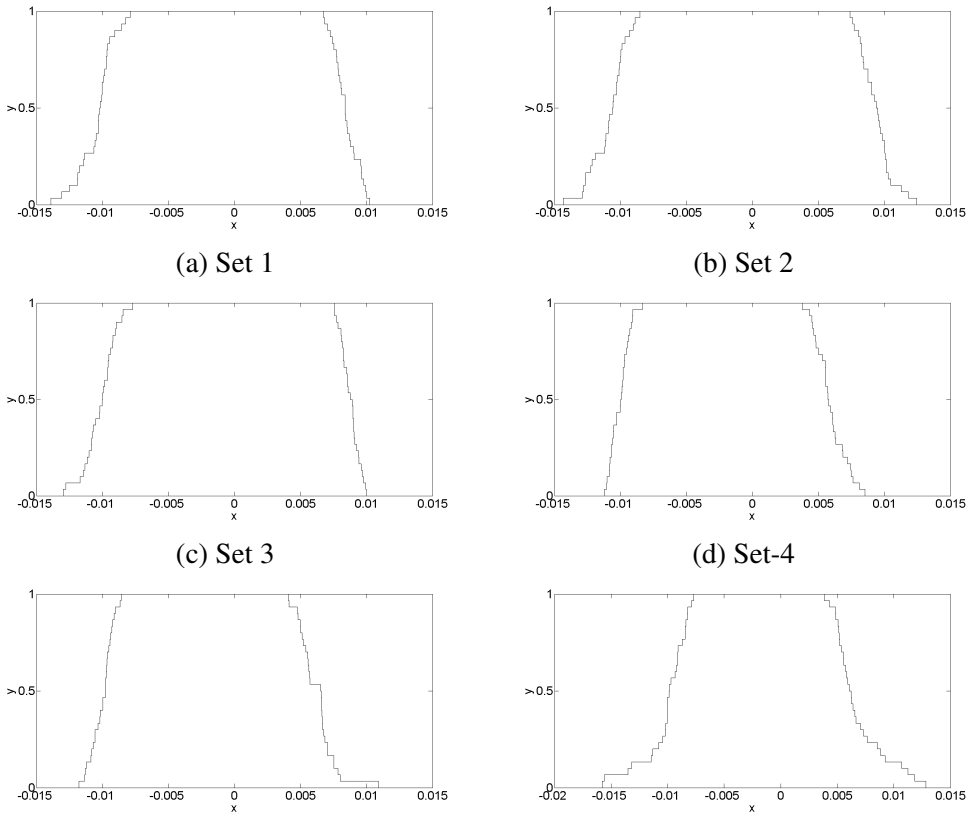


Figure 5.12: Example of creating a T-I fuzzy set from 100 spectra

Figure 5.13 shows 6 T-I fuzzy sets created for feature 1 of G-I, each consisting of interval data from 30 spectra. First three sets have been created from spectra from Case 1 and the last three from Case 2 of G-I. It can be seen that sets vary reflecting the intra-case uncertainty.



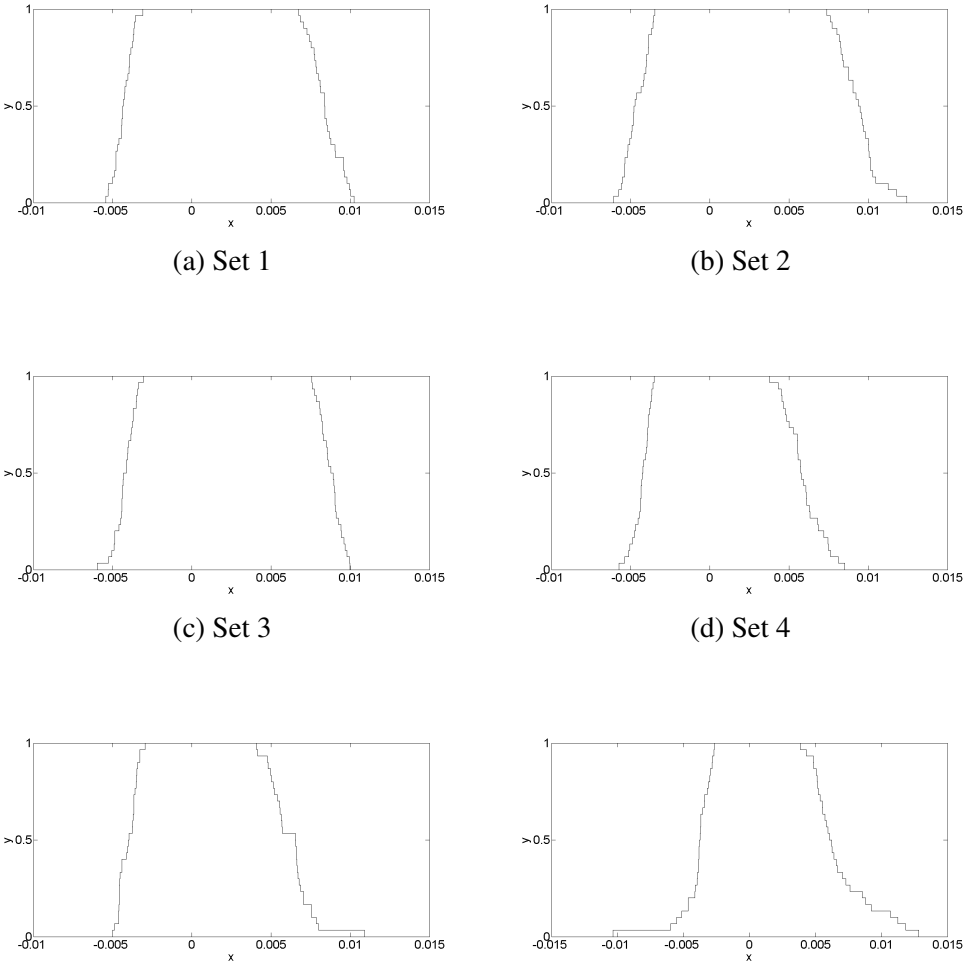
(a) Set 1 (b) Set 2

(c) Set 3 (d) Set-4

(e) Set 5 (f) Set 6

Figure 5.13: T-I fuzzy sets for feature 1 for G-I

In the same way, T-I fuzzy sets were created for the remaining four features for G-I. These sets are shown in Figures 5.14, 5.15, 5.16 and 5.17 with reference to features 2-5 respectively. There are a total of 30 T-I sets for G-I. For G-II and G-III, the same procedure was repeated and 30 T-I sets for each grade were created.



(a) Set 1

(b) Set 2

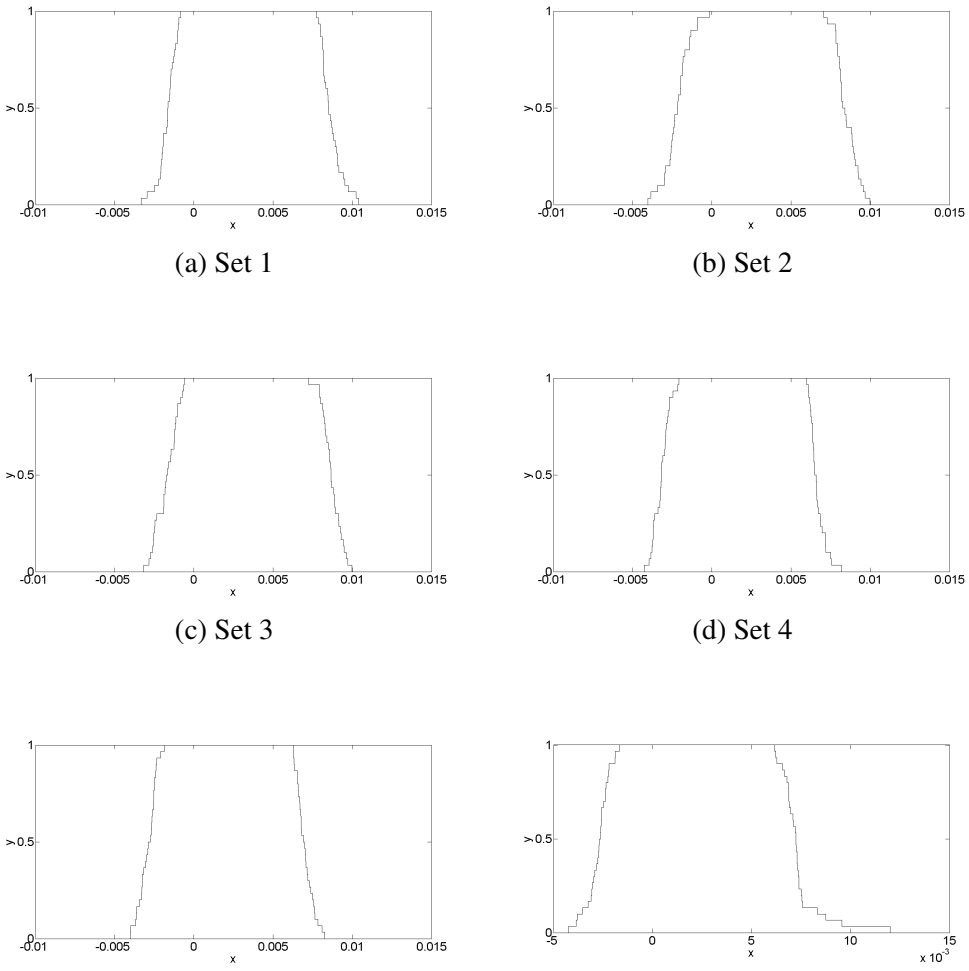
(c) Set 3

(d) Set 4

(e) Set 5

(f) Set 6

Figure 5.14: T-I fuzzy sets for feature 2 for G-I

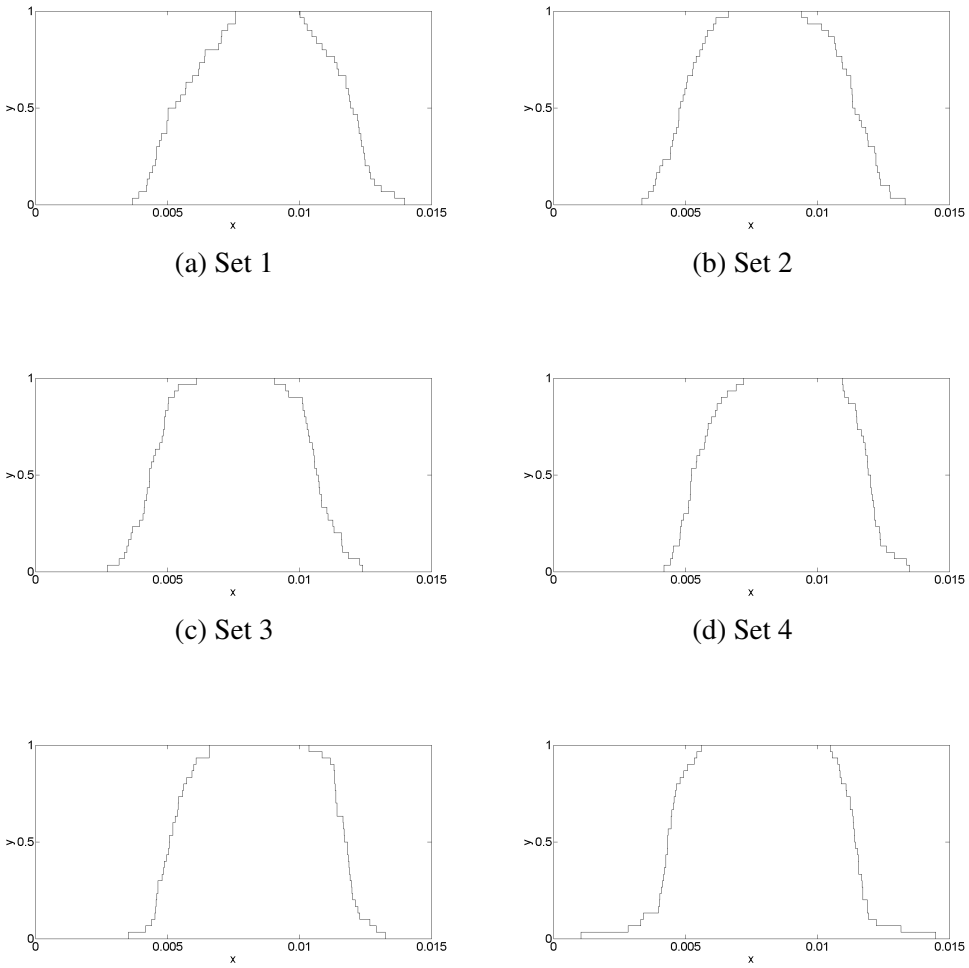


(a) Set 1 (b) Set 2

(c) Set 3 (d) Set 4

(e) Set 5 (f) Set-6

Figure 5.15: T-I fuzzy sets for feature 3 for G-I

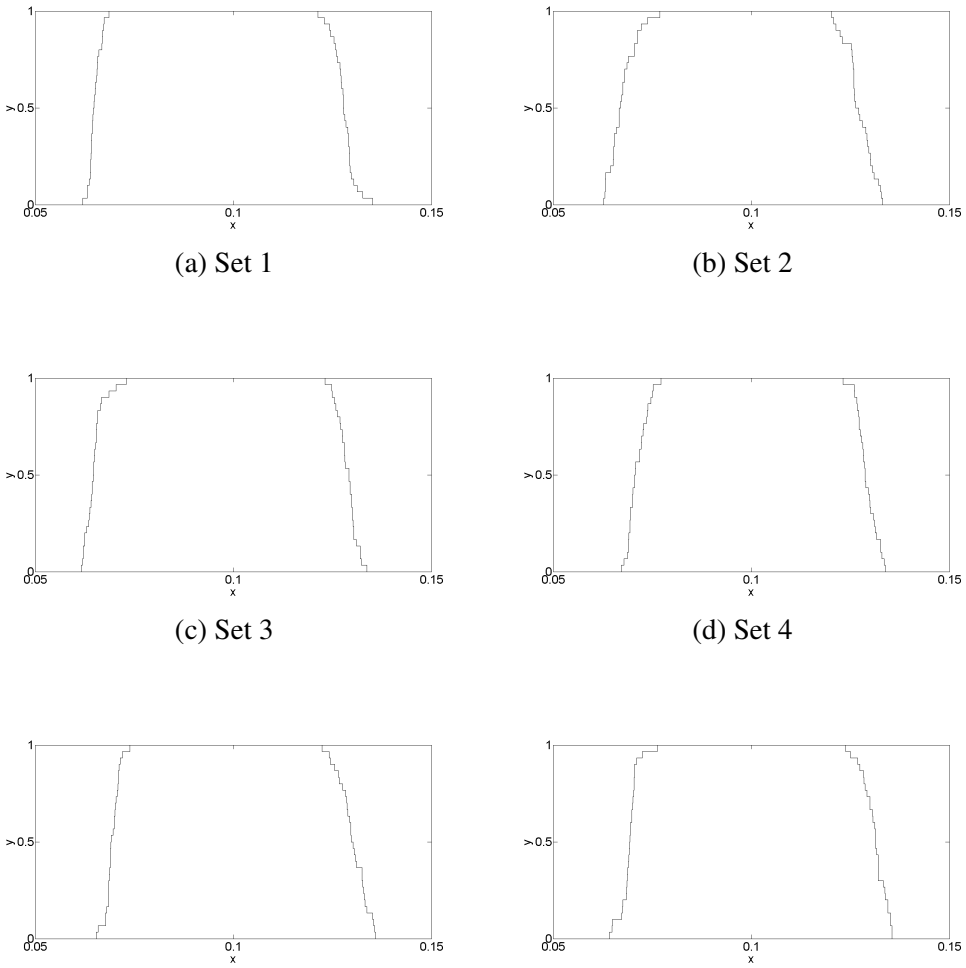


(a) Set 1 (b) Set 2

(c) Set 3 (d) Set 4

(e) Set 5 (f) Set 6

Figure 5.16: T-I fuzzy sets for feature 4 for G-I



(a) Set 1 (b) Set 2

(c) Set 3 (d) Set 4

(e) Set 5 (f) Set 6

Figure 5.17: T-I fuzzy sets for feature 5 for G-I

5.4 Construction of zGT-II Fuzzy Sets

For the construction of zGT-II fuzzy sets creation, we are using the method as described in Chapter 2 Equation 2.28. For the creation of the sets we use the similar approach as we used for the creation of T-I fuzzy sets. The algorithm for zGT-II set creation is as follows:

Let $([a_1,b_1],.....,[a_n,b_n])$ and $([c_1,d_1],....,[c_n,d_n])$ are two T-I fuzzy sets.

Step 1: Split data in groups with one group consisting of left entry of each interval data and second consisting of right entry of each interval data horizontally.

$([a_1,c_1],[a_2,c_2]),...([a_n,c_n])$ and $([b_1,d_1],.....,[b_n,d_n])$

Step 2: Sort elements of the first group in ascending order

$[a_{1s}, c_{1s}], [a_{2s}, c_{2s}], .., [a_{ns}, c_{ns}] = sort(ascend[a_1, c_1]), sort(ascend[a_2, c_2]), ..., sort(ascend[a_n, c_n])$

where $[a_{1s}, c_{1s}], ..., [a_{ns}, c_{ns}]$ is used to represent the result after sorting

Step 3: Sort elements of the second group in descending order

$([b_{1s}, d_{1s}]), ..., ([b_{ns}, d_{ns}]) = sort(descend[b_1, d_1]), sort(descend[b_2, d_2]), .., sort(descend[b_n, d_n])$

where $([b_{1s}, d_{1s}]), ..., ([b_{ns}, d_{ns}])$ is used to represent the result after sorting

Step 4: the zSlices for created zGT-II set are achieved by combining the results of step 2 and 3:

$z_1(z = 0.5) = ([a_{1s}, b_{1s}], ..., [a_{ns}, b_{ns}])$

$z_2(z = 1) = ([c_{1s}, d_{1s}], ..., [c_{ns}, d_{ns}])$

Note: The number of zSlices in zGT-II set is the number of T-I fuzzy sets to be combined.

To illustrate the algorithm we consider an example using synthetic data. As we are mainly concerned with overlapping data, therefore, the example is also of overlapping interval data. Suppose we have three sets of interval data as described in Table 5.10.

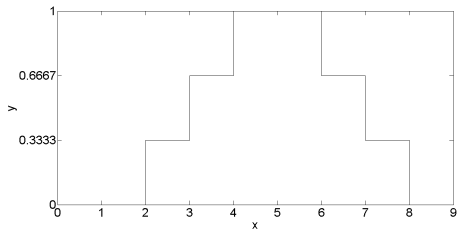
Table 5.10: Synthetic data example for zGT-II fuzzy set creation

Data set 1	Data set 2	Data Set 3
(3,6)	(2,5)	(4,7)
(2,7)	(1,6)	(1,5)
(4,8)	(3,7)	(3,8)

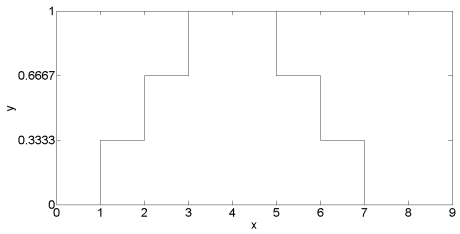
Firstly, we create the T-I fuzzy sets by the method explained before. The created T-I sets are shown in Table 5.11. The sets can also be viewed in Figure 5.18.

Table 5.11: T-I fuzzy sets for Synthetic data example for zGT-II fuzzy set creation

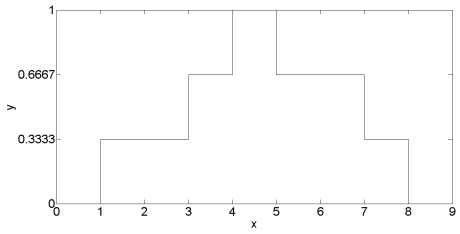
Data set 1	Data set 2	Data Set 3
(2,8)	(1,7)	(1,8)
(3,7)	(2,6)	(3,7)
(4,6)	(3,5)	(4,5)



(a) Data set 1



(b) Data Set 2



(c) Data Set 3

Figure 5.18: T-I fuzzy sets for synthetic data

Firstly, we apply the original method to create zGT-II fuzzy sets described in Chapter 2.

$\tilde{z_1} = 0.3333/(y1/([2,8] \cup [1,7] \cup [1,8]) + y2/([3,7] \cup [2,6] \cup [3,7]) + y3/([4,6] \cup [3,5] \cup [4,5])$

$= 0.25/(y1/[1,8] + y2/[2,7] + y3/[3,6])$

$\tilde{z_2} = 0.6667/(y1/(([2,8] \cap [1,7]) \cup ([2,8] \cap [1,8]) \cup ([1,7] \cap [1,8]) + y2/(([3,7] \cap [2,6]) \cup ([3,7] \cap [3,7])) \cup ([2,6] \cap [3,7])) + y3/((4,6] \cap [3,5]) \cup ([4,6] \cap [4,5]) \cup ([3,5] \cap [4,5])$

$= 0.25/(y1/[1,8] + [3,7] + [4,5])$

$\tilde{z_3} = 1/(y1/(([2,8] \cap [1,7] \cap [1,8]) + ([3,7] \cap [2,6] \cap [3,7]) + ([4,6] \cap [3,5] \cap [4,5]))$

$= 1/(y1/[2,7] + [3,5] + [4,5])$

Now, we apply our proposed approximation algorithm to create zGT-II fuzzy sets.

Step 1: Combine intervals in two groups with 3 intervals in each group:

$[(2,1,1),(3,2,3),(4,3,4)]$ and $[((8,7,8),(7,5,7),(6,5,5)]$

Step 2: Sort first group in ascending order:

$(1,1,2),(2,3,3),(3,4,4)$

Step 3: Sort second group in descending order:

$(8,8,7),(7,7,5),(6,5,5)$

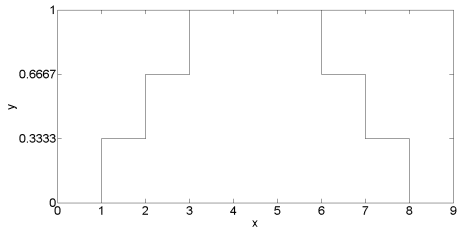
Step 4: Combine results to create zGT-II fuzzy set with 3 zSlices

$z_1(z = 0.3333) = [(1,8),(2,7),(3,6)]$

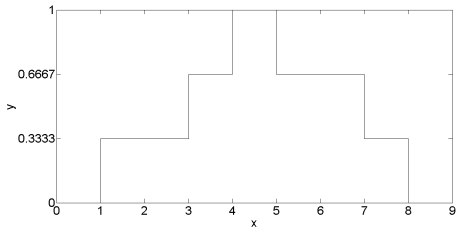
$z_2(z = 0.6667) = [(1,8),(3,7),(4,5)]$

$z_3(z = 1) = [(2,7),(3,5),(4,5)]$

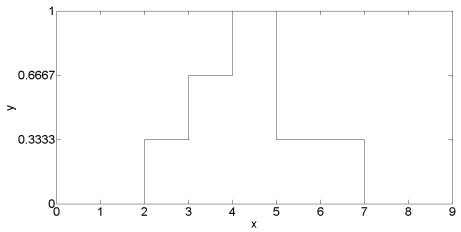
It can be seen that both the methods produce equivalent results. The resultant 2 dimensions plot of zGT-II fuzzy set can be viewed in Figure 5.19.



(a) zSlice 1 = 0.3333



(b) zSlice 2 = 0.6667



(c) zSlice 3 = 1

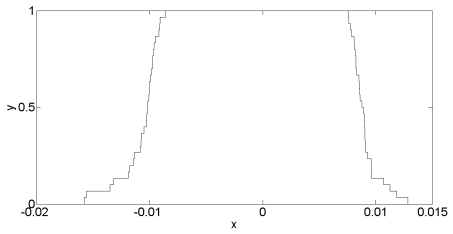
Figure 5.19: 2D plot of zGT-II fuzzy set for synthetic data

For our data set, for each feature, we create a zGT-II fuzzy set. Each zGT-II fuzzy set has 6 zSlices. For each grade, we have 5 zGT-II fuzzy sets each consisting of 6 zSlices.

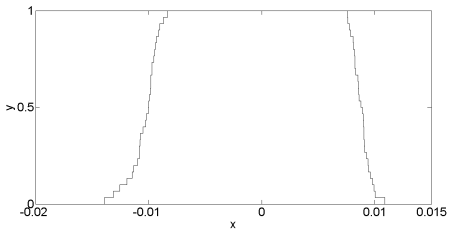
- Number of Features = 5
- Number of zGT-II fuzzy sets for each grade = 5
- Number of zSlices in each zGT-II fuzzy set = 6
- Number of Grades=3

- Total number of zGT-II fuzzy sets = 15

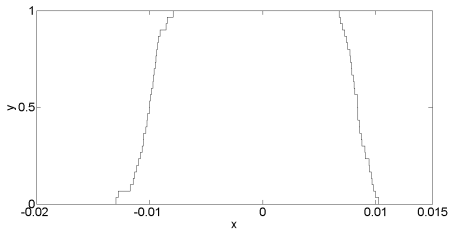
Figure 5.20 shows zGT-II fuzzy set with 6 zSlices created after combining the 6 T-I fuzzy sets for feature 1 of G-I as previously shown in Figure 5.13.



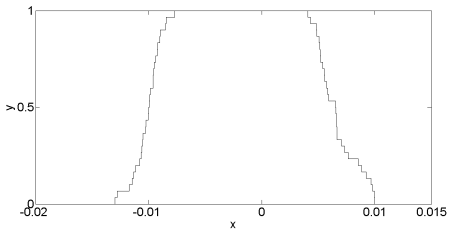
(a) zSlice 1 = 0.1667



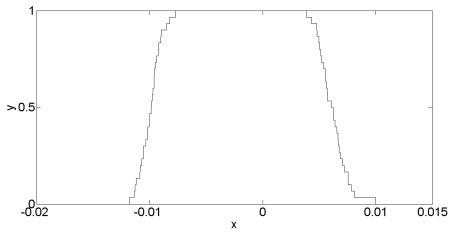
(b) zSlice 2 = 0.3333



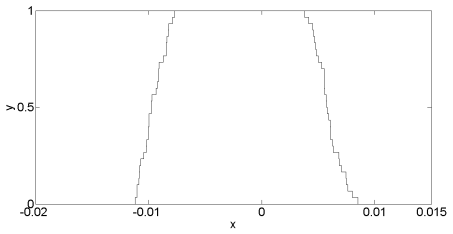
(c) zSlice 3 = 0.5



(d) zSlice 4 = 0.6667



(e) zSlice 5 = 0.8333



(f) zSlice 6 = 1

Figure 5.20: 2D plot of zGT-II fuzzy set for feature 1 for G-I

The combined plot showing the complete zGT-II fuzzy set is shown in Figure 5.21. The z-axes shows the 6 zSlices. It can be observed from the figure that these zSlices aim to cover both the intra-cases and inter-case uncertainties between interval data taken from spectra of the same case and from different cases. As in the first phase of creation, T-I fuzzy sets are created as shown in Figure 5.13. These sets cover the intra-case uncertainty found in the two cases of G-I. The zGT-II fuzzy set combines individual intra-case uncertainties and with the help of zSlices and in this way, the inter-case uncertainties are

modelled.

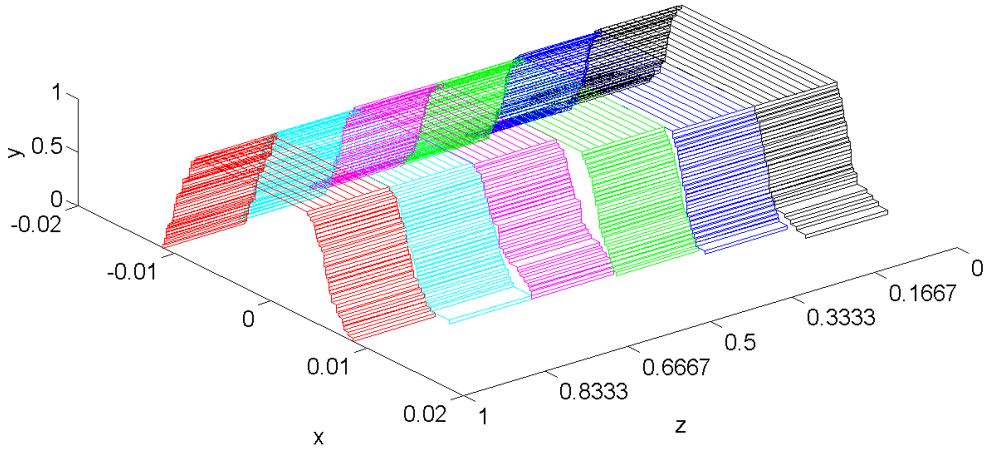
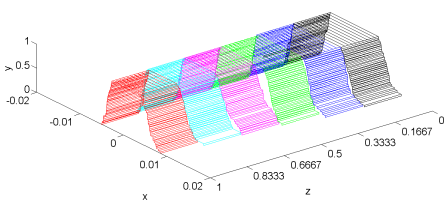


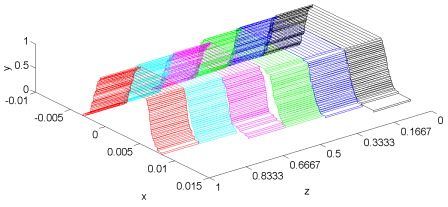
Figure 5.21: 3D plot of zGT-II fuzzy set for feature 1 of G-I

The zGT-II fuzzy sets for features 2-5 for G-I are shown in Figure 5.22. It can be seen that each feature has distinct zSlices. These zGT-II fuzzy sets are used as bench mark prototype for a particular feature of a grade.

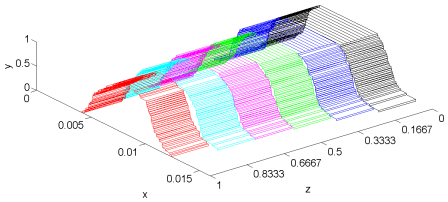
In the same way, zGT-II fuzzy sets are created for G-II and G-III. These zGT-II fuzzy sets are shown in Figures 5.23 and 5.24 for G-II and G-III respectively. Finally, we have 15 bench mark zGT-II sets with 5 zGT-II fuzzy sets for each grade.



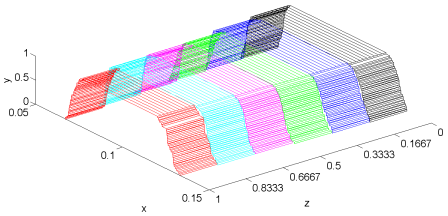
(a) Feature 2



(b) Feature 3

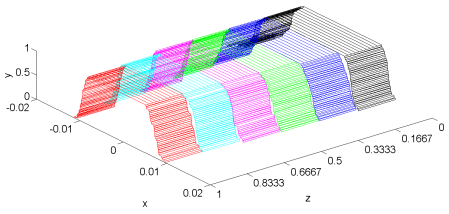


(c) Feature 4

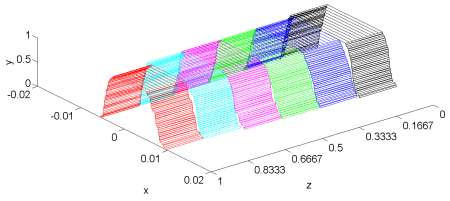


(d) Feature 5

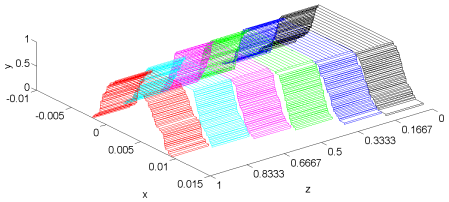
Figure 5.22: 3D plots for zGT-II fuzzy sets for features 2-5 for G-I



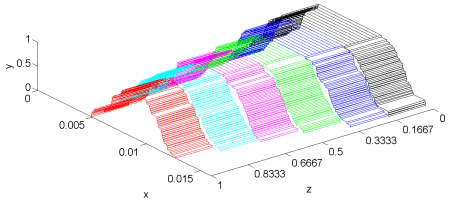
(a) Feature 1



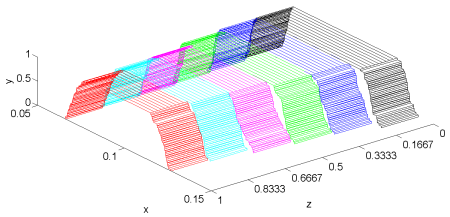
(b) Feature 2



(c) Feature 3

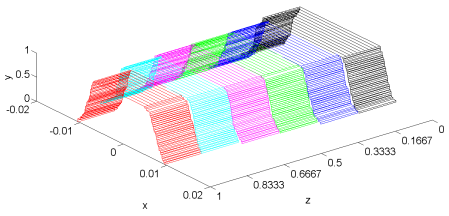


(d) Feature 4

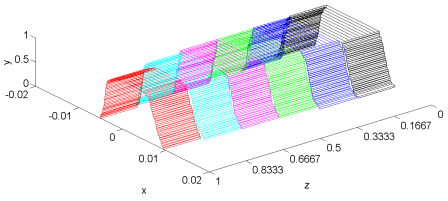


(e) Feature 5

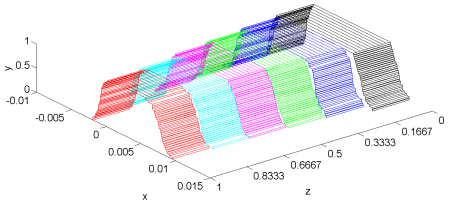
Figure 5.23: 3D plots for zGT-II fuzzy sets for features 1-5 for G-II



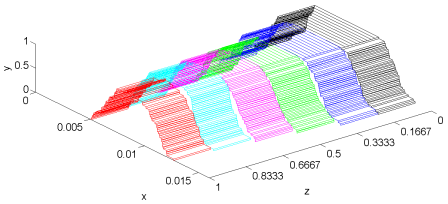
(a) Feature 1



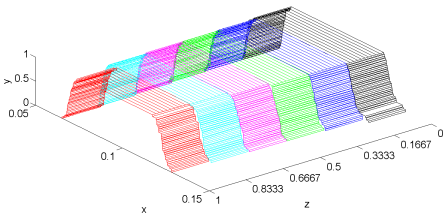
(b) Feature 2



(c) Feature 3



(d) Feature 4



(e) Feature 5

Figure 5.24: 3D plots for zGT-II fuzzy sets for features 1-5 for G-III

5.4.1 Similarity Measures for Type-II Fuzzy Sets

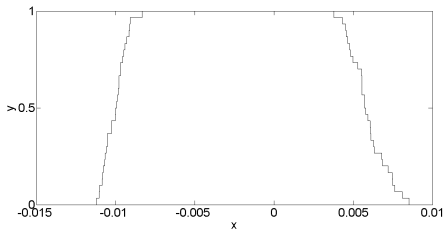
Similarity measures are commonly used in set theory to compare crisp, T-I and T-II fuzzy sets. For our work we are using the extended version of Interval T-II Jaccard method recently introduced by McCulloch et al [69] and explained in Chapter 2. After creating the zGT-II prototype fuzzy sets, we find similarity between prototype zGT-II sets with an unseen T-I fuzzy set. As each zGT-II set has 6 zSlices, we shall replicate the T-I fuzzy sets 6 times to compare it against each zSlice and then find the similarity measure.

The similarity is calculated by the following Equation:

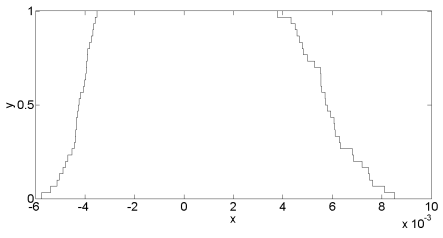
$$S(\tilde{P},U) = \frac{\sum_{i \in L} z_i S_\lambda(\tilde{P}_{z_i},U)}{\sum_{i \in L} z_i} \tag{5.1}$$

where S is a similarity function for the zGT-II fuzzy set $\tilde{P}$ and an unseen T-I fuzzy set U. S_λ is a similarity function applied to the IT-II fuzzy set at zLevel i shown as $\tilde{P}_{z_i}$ and the unseen T-I fuzzy set U. L is the set of zLevels used in $\tilde{P}$, and z_i represents a particular zLevel (secondary degree of membership). A value of 0 indicates disjoint sets where as a value of 1 means the sets are identical.

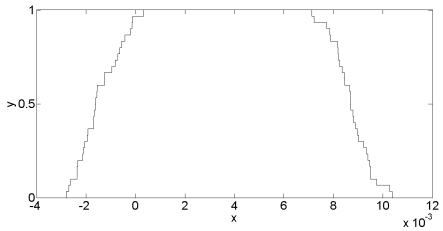
To describe the method, we consider an example with unseen data for G-I. Firstly, we create T-I fuzzy sets with 30 spectral interval data for 5 features. These T-I fuzzy set are shown in Figure 5.25.



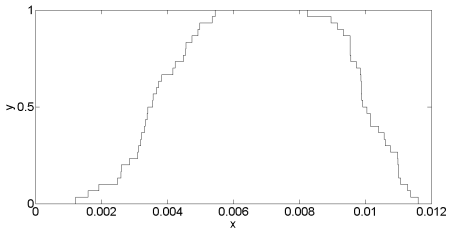
(a) Feature 1



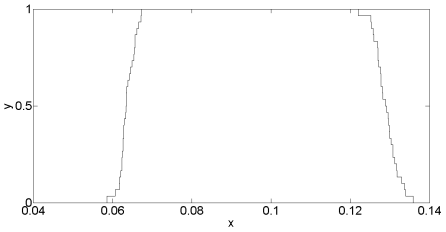
(b) Feature 2



(c) Feature 3



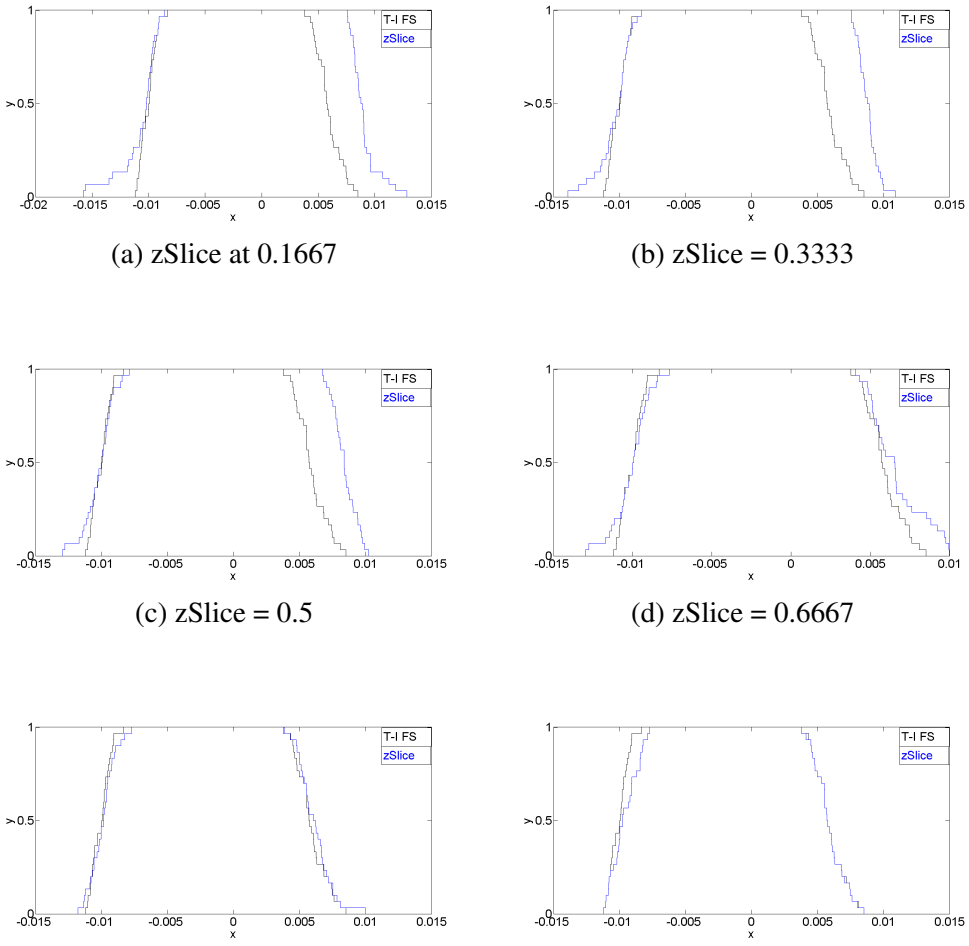
(d) Feature 4



(e) Feature 5

Figure 5.25: T-I fuzzy sets for unseen data of G-I

Now, these T-I fuzzy sets are compared against the model prototype zGT-II fuzzy sets for each feature for each grade. To illustrate this, we show the similarity of unseen T-I fuzzy set with zGT-II fuzzy set for feature 1. The comparison of unseen T-I fuzzy set with each zSlice of feature 1 of zGT-II set is shown in Figure 5.26. The first similarity values calculated by Equation 5.1 are shown in Table 5.12.



(a) zSlice at 0.1667

(b) zSlice = 0.3333

(c) zSlice = 0.5

(d) zSlice = 0.6667

(e) zSlice = 0.8333

(f) zSlice = 1

Figure 5.26: T-I fuzzy sets for unseen data of G-I

Table 5.12: First Similarity scores for feature 1 for unseen T-I fuzzy set example

zSlice level	Similarity score
0.1667	0.8083
0.3333	0.8313
0.5	0.8503
0.6667	0.9371
0.8333	0.9704
1	0.9719

Now these similarity scores are multiplied by their corresponding weighted zSlice value and then divided by the sum of all weights as defined in Equation 5.1.

$$FS = \frac{0.1667*0.8083+0.3333*0.8313+0.5*0.8503+0.6667*0.9371+0.8333*0.9704+1*0.9719}{0.1667+0.3333+0.5+0.6667+0.8333+1}$$

Where FS is the final similarity for this feature, that after evaluation comes at 0.9264. In the same way similarity for all features is calculated for all grades.

In the next section we test the prototype model with unseen data in order to assess its performance.

5.5 Model Testing with Unseen Data

In this section, we consider examples of unseen spectral data from our data set. We create 10 T-I fuzzy sets each created using 30 spectra, for each feature. Two sets are from two cases of G-I, 6 sets from G-II cases and 6 sets from 6 cases of G-III. All of these T-I fuzzy sets are compared against the prototype zGT-II fuzzy sets for each feature of each grade. For the classification grade of unseen data, we use two methods.

1. Summation over similarities

2. Majority vote

In the first method, we record a similarity value for each feature against each grade and then compute the sum of all similarities for each grade and report the maximum value as the predicted grade.

In the second method we take the maximum value for each feature as a vote. In the end the grade with maximum number of votes is reported as the classified grade.

This model testing scheme is also shown in Figure 5.27.

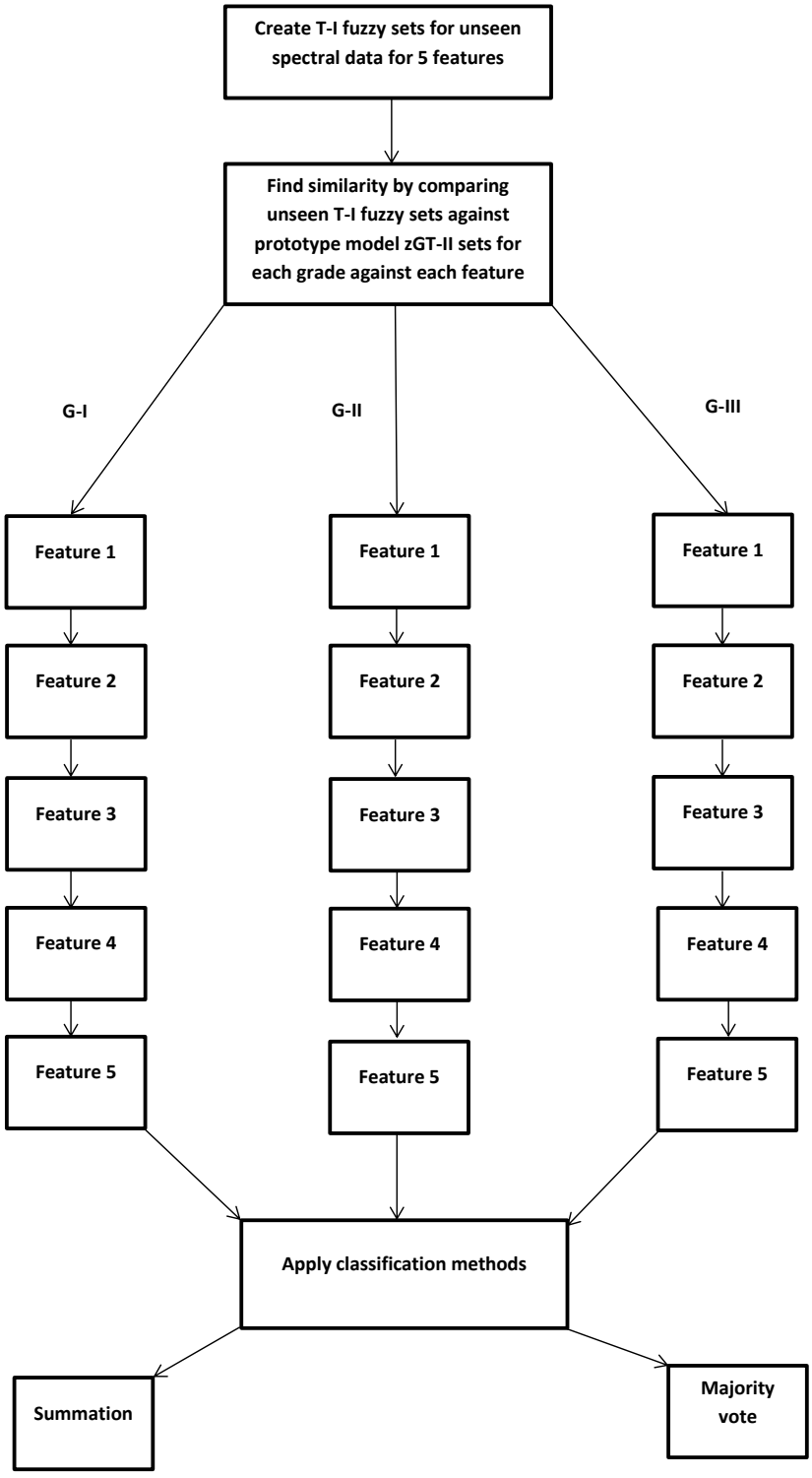


Figure 5.27: Model testing scheme

The method has been tested on all cases used for the creation of the prototype model. Table 5.13 shows the similarity scores for the two cases of G-I compared against all grades along with results determined by the both classification methods. **W** in majority vote column indicates the winning grade where as **L** indicates a losing grade. **T** indicates a tie when votes are equal for a certain feature. The maximum similarity score for a feature is highlighted in the tables and is also considered the winner for that feature. It can be seen that both methods correctly classified the grade. In case of majority vote, case 1 won by 4-1 and case 2 by 5-0.

Table 5.13: Similarity scores for G-I with test data

(a) Case 1

Feature	G-I	G-II	G-III
1	**0.9264**	0.8397	0.8235
2	**0.8938**	0.8035	0.7905
3	0.8122	0.8452	**0.8790**
4	**0.6407**	0.5194	0.5347
5	**0.9001**	0.8391	0.8719
Sum	**4.1732**	3.8469	3.8996
Majority Vote	**W**	L	L

(b) Case 2

Feature	G-I	G-II	G-III
1	**0.9047**	0.8424	0.8235
2	**0.8681**	0.7935	0.7905
3	**0.7816**	0.6838	0.7319
4	**0.7653**	0.7089	0.7102
5	**0.9283**	0.8684	0.8617
Sum	**4.248**	3.897	3.9421
Majority Vote	**W**	L	L

The similarity scores with both methods for G-II are presented in Table 5.14. It can be seen that in the case of the sum of similarities method, only case 1 was classified as G-II where as cases 2-5 were classified as G-III. In case of the majority vote, two cases (case 1 and case 6) were classified as G-II, cases 2, 4 and 5 were classified as G-III, case 3 was tied between G-I and G-II. The results indicate that performance was very poor in the case of classifying G-II spectra. This reiterates the point discussed earlier that G-II is regarded as a difficult grade to distinguish from other grades.

Table 5.14: Similarity scores for G-II with test data

(a) Case 1

Feature	G-I	G-II	G-III
1	**0.8140**	0.7370	0.7232
2	**0.7098**	0.6733	0.6532
3	0.3543	**0.4593**	0.4163
4	0.4513	**0.5545**	0.5279
5	0.8300	**0.8922**	0.8804
Sum	3.1594	**3.3163**	3.201
Majority Vote	L	**W**	L

(b) Case 2

Feature	G-I	G-II	G-III
1	0.8723	**0.8938**	0.8847
2	0.8122	0.8947	**0.9032**
3	0.8606	0.8165	**0.8919**
4	**0.8773**	0.7602	0.7872
5	0.8666	0.9111	**0.9390**
Sum	4.289	4.2763	**4.4060**
Majority Vote	L	L	**W**

(c) Case 3

Feature	G-I	G-II	G-III
1	0.8820	**0.9175**	0.9077
2	0.8044	**0.8889**	0.8795
3	0.8606	0.8177	**0.8886**
4	**0.8547**	0.7804	0.7973
5	**0.9422**	0.8802	0.8928
Sum	4.3336	4.2847	**4.3659**
Majority Vote	T	T	L

(d) Case 4

Feature	G-I	G-II	G-III
1	0.7915	**0.8728**	0.8556
2	0.7877	0.8893	**0.8910**
3	0.7877	0.8203	**0.8374**
4	0.7847	0.8037	**0.8060**
5	**0.9518**	0.8977	0.9001
Sum	4.1034	4.2838	**4.2901**
Majority Vote	L	L	**W**

(e) Case 5

Feature	G-I	G-II	G-III
1	0.8902	0.9028	**0.9040**
2	0.8488	0.8959	**0.9092**
3	0.5784	**0.7176**	0.6603
4	**0.8394**	0.7005	0.7223
5	0.8719	0.9066	**0.9513**
Sum	4.0287	4.1234	**4.1471**
Majority Vote	L	L	**W**

(f) Case 6

Feature	G-I	G-II	G-III
1	0.8268	**0.8887**	0.8793
2	0.7615	**0.8549**	0.8491
3	**0.8100**	0.7337	0.8075
4	0.3884	**0.4778**	0.4518
5	0.8950	**0.9105**	0.9094
Sum	3.6817	3.8656	**3.8971**
Majority Vote	L	**W**	L

Table 5.15 represents the results obtained from testing unseen spectra taken from 6 cases of G-III. It can be seen that in case of maximum sum of similarity method, all six cases were classified as G-III. In the case of Majority vote, three cases (3, 4 and 5) were classified correctly where as case 1 was falsely classified as G-II and there was a tie between G-II and G-III for case 2 and between G-I and G-III for case 6. The results

indicate that both methods performed reasonably well for classification of unseen spectra from G-III cases.

Table 5.15: Similarity scores for G-III with test data

(a) Case 1

Feature	G-I	G-II	G-III
1	0.7970	**0.8803**	0.8621
2	0.7101	**0.8147**	0.8131
3	**0.7737**	0.6343	0.6968
4	0.3061	**0.3890**	0.3641
5	0.8296	**0.8779**	0.8618
Sum	3.4165	3.5962	**3.5979**
Majority Vote	L	**W**	L

(b) Case 2

Feature	G-I	G-II	G-III
1	0.8981	**0.9199**	0.9131
2	0.8198	**0.9234**	0.9153
3	0.7767	0.8246	**0.8352**
4	**0.8004**	0.7767	0.7987
5	0.8860	0.8901	**0.9463**
Sum	4.1810	4.3347	**4.4086**
Majority Vote	L	T	T

(c) Case 3

Feature	G-I	G-II	G-III
1	0.8674	**0.9168**	0.9088
2	0.8284	**0.9120**	0.9036
3	0.7827	0.8261	**0.8387**
4	0.6564	0.7321	**0.7392**
5	0.8860	0.9195	**0.9499**
Sum	4.0209	4.3065	**4.3402**
Majority Vote	L	L	**W**

(d) Case 4

Feature	G-I	G-II	G-III
1	0.8242	**0.8848**	0.8755
2	0.8128	0.9075	**0.9085**
3	0.8227	0.8260	**0.8736**
4	0.6840	**0.7516**	0.7318
5	0.8819	0.9219	**0.9405**
Sum	4.0256	4.2918	**4.3299**
Majority Vote	L	L	**W**

(e) Case 5

Feature	G-I	G-II	G-III
1	0.8695	**0.9150**	0.9023
2	0.8331	0.8930	**0.9109**
3	0.8151	0.8397	**0.8746**
4	**0.8388**	0.7214	0.7362
5	0.9069	0.8853	**0.9313**
Sum	4.2634	4.2544	**4.3553**
Majority Vote	L	L	**W**

(f) Case 6

Feature	G-I	G-II	G-III
1	0.8151	**0.8415**	0.8363
2	0.7523	0.8437	**0.8449**
3	0.8476	0.8286	**0.8859**
4	**0.8019**	0.6984	0.7087
5	**0.9409**	0.8888	0.9190
Sum	4.1578	4.101	**4.1948**
Majority Vote	T	L	T

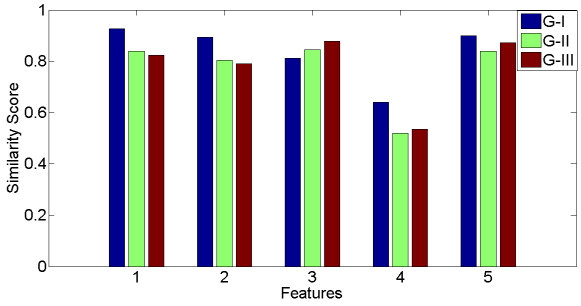
In general we can say that both methods have their advantages and disadvantages. With maximum of summation, we always get a winner no matter how close the other values have been as it mathematically declares a winner. It means that even if there is

a difference of a very small fraction between summation scores of grades, the one with larger value will be declared and it will not indicated the closeness of the competition. In case of the majority vote, if different features have different winners then it may result in a tie as we have seen in case 3 of G-II and case 6 of G-III. A tie reflects that features from multiple grades behave differently. So majority vote does not always have a winner but reflects competition among grades for certain cases; in case of a it shows the complexity involved in classifying a grade. cases.

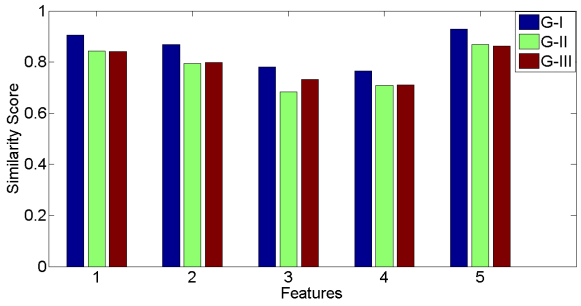
In the next section, we create grade profiles based on the similarity scores and discuss which features have been able to perform well in predicting the correct grade.

5.6 Discussion

Figure 5.28 shows a grade profile for two cases of G-I testing data plotting similarity scores for features as described in previous section. It can be seen that features 1, 2 and 5 provide high similarity scores for both cases with the correct grade where as feature 4 provides the lowest similarity score. For feature 4, G-I similarity was comparatively higher with the correct grade. Feature 3 is the most inconsistent feature as it was able to classify case 2 but classified case 1 as G-III. It can also be observed that there is significant differences between G-I similarity scores compared to G-II and G-III in general for all features where G-I was chosen. That is why both maximum sum of similarity and majority vote performed well for G-I. Another observation is that scores for G-II and G-III remained very close to each other in more features. We conclude that features 1, 2 and 5 are the most useful as bench mark features to distinguish G-I from other grades.



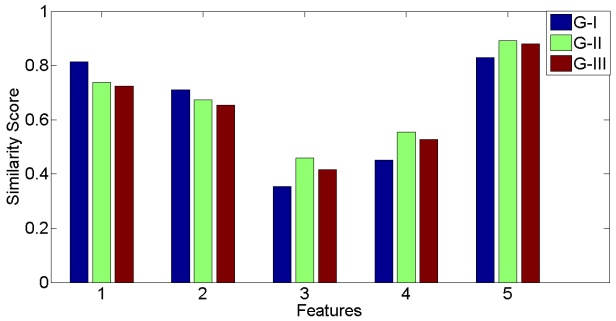
(a) Case 1



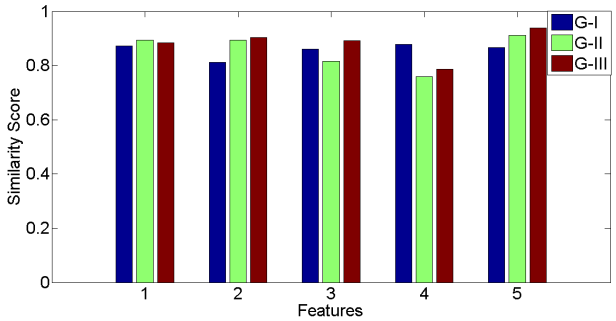
(b) Case 2

Figure 5.28: Grade profile for two cases of G-I

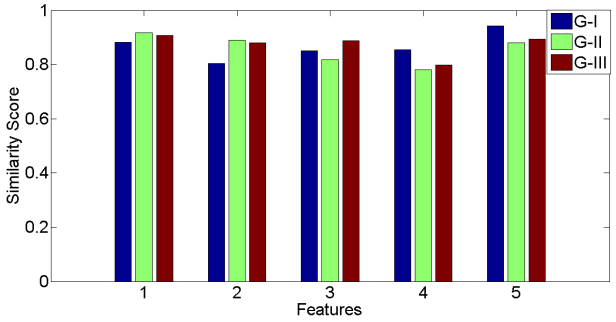
Figure 5.29 shows grade profiles for six test cases of G-II. Previously we have seen that both classification methods perform poorly and G-II is not clearly distinguishable from other grades. However, there are some interesting observations that we can make by looking at the Figure 5.29. Feature 1 is able to classify the correct grade for cases 2, 3, 4 and 6 and case 5 is narrowly mistaken as G-III. Feature 2 correctly classified the grade for cases 3 and 6 where as for cases 2, 4 and 5 it was very close to classifying the correct grade. Feature 5 only classified correctly for case 1. In the majority of the cases where G-II was not classified correctly, it was classified as G-III. This is generally the case in real world scenarios, as G-II and G-III are considered very close to each other and chances of false classification remain high. We conclude that only feature 1 is able to classify the correct for majority of the G-II cases (4 out of 6) while other features remain inconsistent so only feature 1 is useful as a bench mark feature for identifying G-II from other grades.



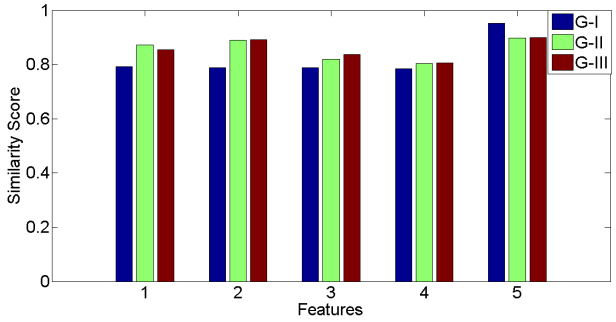
(a) Case 1



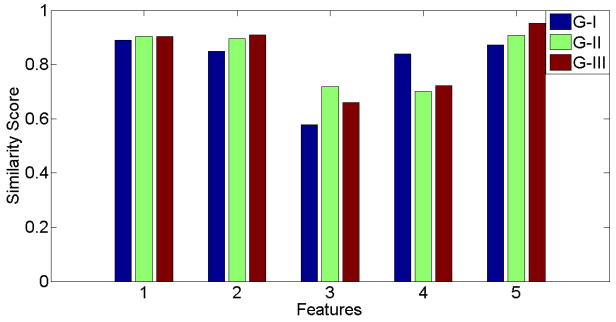
(b) Case 2



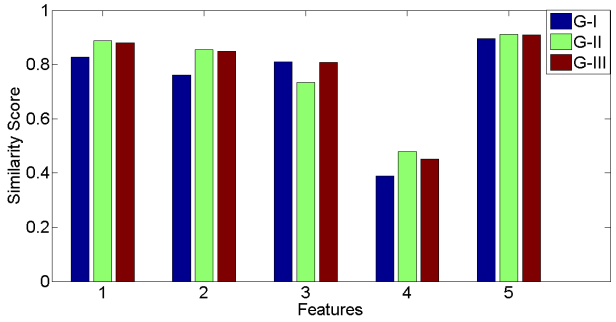
(c) Case 3



(d) Case 4



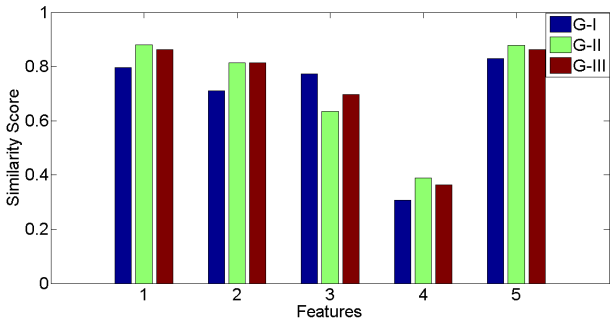
(e) Case 5



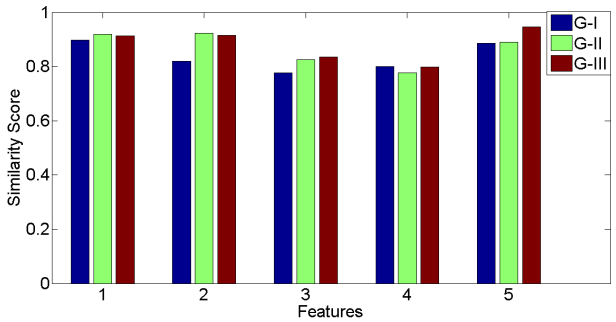
(f) Case 6

Figure 5.29: Grade profile for six cases of G-II

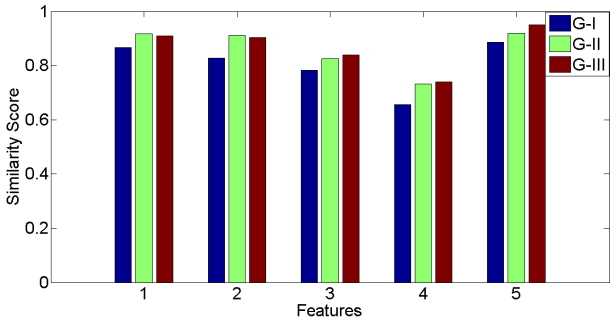
Figure 5.30 shows grade profile for six test cases of G-III with their similarity scores for each feature of each grade. It can be seen that feature 3 always classified correctly except for case 1. Feature 5 also classified correct grade for 4 out of 6 cases (cases 2,3,4, and 5). In case of features 1 and 2, G-III scores were slightly less than G-II where as in case of feature 4, G-III was also falsely classified as G-I for cases 2, 5 and 6. We conclude that features 3 and 5 are best suited for classifying G-III correctly for the majority of the cases and may be used as bench mark features for G-III classification.



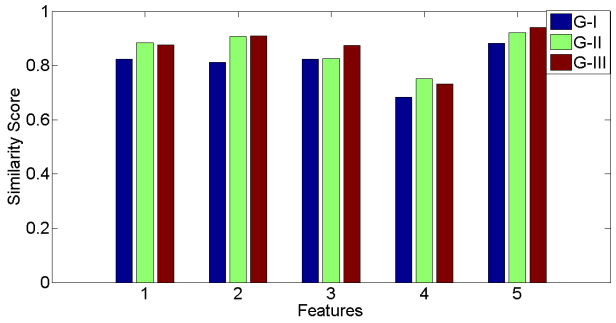
(a) Case 1



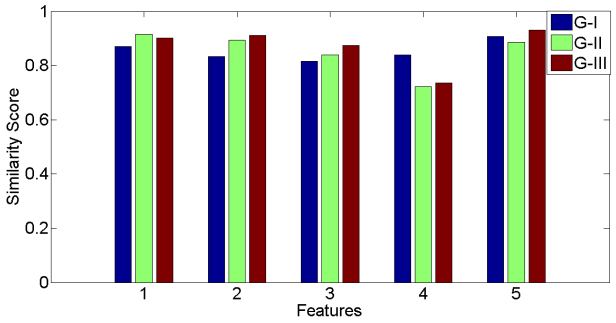
(b) Case 2



(c) Case 3



(d) Case 4



(e) Case 5

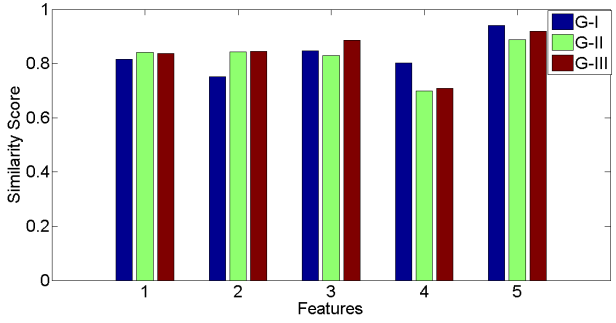


Figure 5.30: Grade profile for six cases of G-III

Table 5.16 shows a summary of the features against correctly classified grades. The features that were able to classify the grade correctly for test cases have been highlighted. The results indicate that features performed differently for the three grades. Feature 3 is only significant in case of G-III classification and did not perform well for any other grade. Similarly, feature 2 only performed well in case of G-I and classified false grade in all other cases. Our results indicate that various features based on different regions of the same spectra may provide different information and some may be helpful in classifying a grade correctly while others may not be useful as explained before. It can also be seen that zGT-II fuzzy sets based on interval data from spectral regions may be useful in extracting important information regarding grade classification problems where both inter and intra variabilities are involved.

Features 1 and 2 have the same peak value but they do not behave identically, although they have similar results in some cases. This shows that a feature with a common value can still be useful and may provide useful information for classification.

Table 5.16: Summary of grade profiles

Grades	Feature 1	Feature 2	Feature 3	Feature 4	Feature 5
	Correctly classified / Total test cases				
I	**2/2**	**2/2**	1/2	**2/2**	**2/2**
II	**4/6**	2/6	2/6	2/6	2/6
III	0/6	3/6	**5/6**	1/6	**4/6**

Table 5.17 shows the summary of the results in terms of correct classification of grades by both the summation and the majority vote method for all test cases. It can be seen that summation method has performed well for G-I and G-III test cases. The majority vote method has performed well for G-I test cases and has shown reasonable results for G-III test cases as well. Both methods have not been able to classify G-II.

Table 5.17: Summary of results with test cases by the summation and majority vote method

Type	Test Cases	Correct Classification Summation	Incorrect Classification Summation	Correct Classification Majority vote	Incorrect Classification Majority vote
G-I	2	2	0	2	0
G-II	6	1	5	2	3 (1 Tie)
G-III	6	6	0	3	1 (2 Tie)

5.7 Model Testing with an Alternative Configuration

The model has also been tested with a similar configuration as that used for creating the FIS in Chapter 4. To allow comparison, we create 26 sets for G-II each consisting of 4 spectra making a total of 104 spectra, the same number of spectra as used in Chapter 4. For our model, the number of sets needs to be same for all grades, therefore, we create 26 sets for G-I and G-III. We have used the following approach to achieve this.

G-I: 13 sets each for 2 cases making 26 sets in total per feature each consisting of 4 spectra.

G-II: 26 sets from 26 cases

G-III: 4 sets for first 4 cases each consisting of 4 spectra: 5 sets for last 2 cases each consisting of 4 spectra each completing a total of 26 sets.

For all features for all grades, at first T-I fuzzy sets were created. Each feature has 26 sets. After that zGT-II sets were created for each grade by combining the 26 sets per feature. Each zGT-II set has 26 zSlices. These zGT-II fuzzy sets serve as prototype for the unseen data to be compared against.

We have used 2 unseen sets of each grade to test the system.

G-I: 2 sets from 2 cases

G-II: 2 sets from 2 cases out of 26

G-III: 2 sets from 2 cases out of 6

Table 5.18 shows the results obtained after testing the model with two G-I cases. The tables show the similarity scores as well as classified grade by the maximum sum

of similarity and the majority vote method. It can be seen that case 1 was incorrectly classified as G-III by both methods and case 2 was classified correctly as G-I by both methods.

Table 5.18: Similarity scores for G-I (Alternative Configuration)

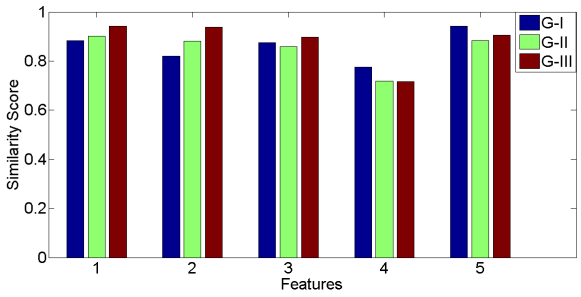
(a) Case 1

Feature	G-I	G-II	G-III
1	0.8823	0.9005	**0.9430**
2	0.8194	0.8814	**0.9380**
3	0.8745	0.8580	**0.8964**
4	**0.7755**	0.7188	0.7167
5	**0.9420**	0.8838	0.9056
Sum	4.2937	4.2425	**4.3997**
Majority Vote	L	L	**W**

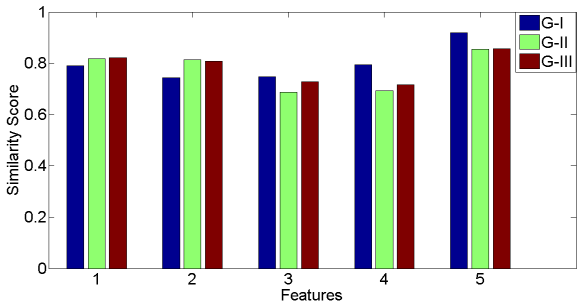
(b) Case 2

Feature	G-I	G-II	G-III
1	0.7907	0.8183	**0.8220**
2	0.7433	**0.8129**	0.8076
3	**0.7479**	0.6867	0.7272
4	**0.7941**	0.6930	0.7166
5	**0.9194**	0.8548	0.8562
Sum	**3.9954**	3.8657	3.9296
Majority Vote	**W**	L	L

Figure 5.31 shows a grade profile plot of the similarity scores for both test cases of G-I with all features. It can be seen that features 4 and 5 correctly classified the grade for both cases where as features 1 and 2 never classified the correct grade, falsely classifying it as G-III or G-II with a significant difference in scores as G-I scores remained the lowest for these features. In terms of low scores, feature 4 scores remained low and feature 5 scores remained consistently high. We conclude that for this configuration, features 4 and 5 are the most appropriate for classification of G-I.



(a) Case 1



(b) Case 2

Figure 5.31: Grade profile for two cases of G-I (Alternative configuration)

Table 5.19 shows the results obtained after testing the model on two cases of G-II. The tables show the similarity scores as well as grade class by the maximum sum of similarity and the majority vote method. It can be seen that case 1 was falsely classified as G-III in terms of maximum sum of similarity method where as the majority vote resulted in a tie between G-II and G-III. Case 2 was also not classified correctly by both methods as both resulted in classification as G-III.

Table 5.19: Similarity scores for G-II (Alternative Configuration)

(a) Case 1

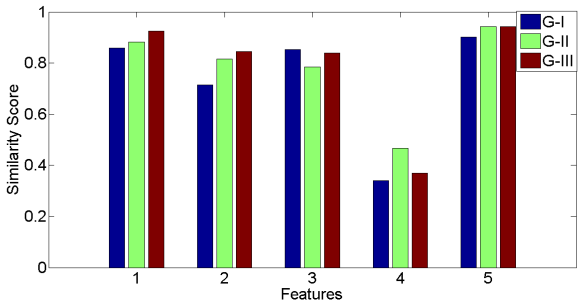
Feature	G-I	G-II	G-III
1	0.8580	0.8825	**0.9246**
2	0.7140	0.8154	**0.8452**
3	**0.8537**	0.7838	0.8386
4	0.3392	**0.4661**	0.3702
5	0.9018	**0.9432**	0.9427
Sum	3.6667	3.891	**3.9303**
Majority Vote	L	T	T

(b) Case-2

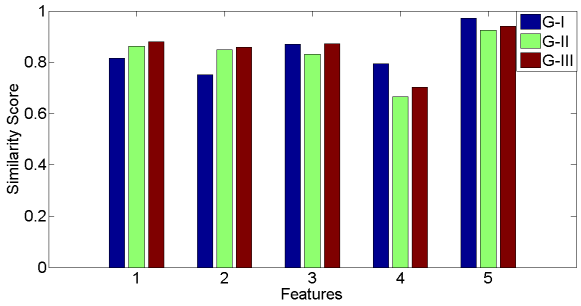
Feature	G-I	G-II	G-III
1	0.8165	0.8619	**0.8796**
2	0.7519	0.8485	**0.8581**
3	0.8697	0.8314	**0.8719**
4	**0.7935**	0.6649	0.7019
5	**0.9719**	0.9255	0.9412
Sum	4.2035	4.1322	**4.2527**
Majority Vote	L	L	**W**

Figure 5.32 shows the grade profile for G-II for this configuration. It can be observed

that in case 1, features 4 and 5 classified the correct grade where as features 1 and 2 falsely classified it as G-III though the values of similarity scores were close to G-II as well. In test case 2 for this configuration, features 1 and 2's similarity scores for G-II and G-III were close although G-III was the highest. Features 3 falsely classified case 2 as G-III and features 4 and 5 classified it as G-I. We conclude that the features make inconsistent classifications for these cases but features 1 and 2 have been a close and may be used for testing with other configurations to make correct classifications. As these two features falsely classified G-II as G-III but for both cases, their scores remained competitive.



(a) Case 1



(b) Case 2

Figure 5.32: Grade profile for two cases of G-II (Alternative configuration)

Table 5.20 shows the results obtained after testing the model with 2 randomly selected cases of G-III. It can be clearly noted that both methods correctly classified the grade for both cases. This shows that the model performed well for these cases and was able to produce correct results.

Table 5.20: Similarity scores for G-III (Alternative Configuration)

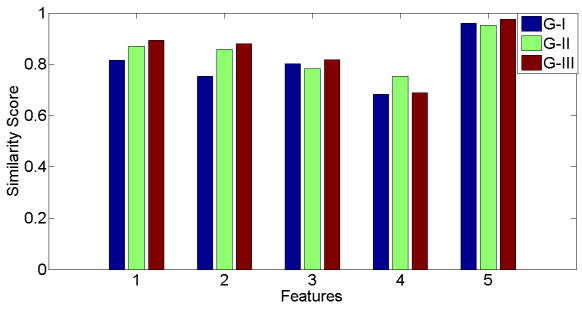
(a) Case 1

Feature	G-I	G-II	G-III
1	0.8164	0.8703	**0.8930**
2	0.7541	0.8584	**0.8810**
3	0.8017	0.7827	**0.8174**
4	0.6834	**0.7533**	0.6886
5	0.9607	0.9523	**0.9764**
Sum	4.0163	4.217	**4.2564**
Majority Vote	L	L	**W**

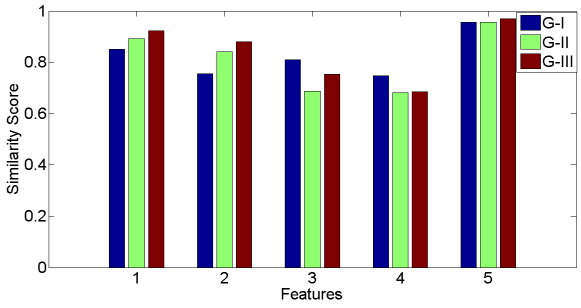
(b) Case 2

Feature	G-I	G-II	G-III
1	0.8507	0.8912	**0.9283**
2	0.7561	0.8417	**0.8798**
3	**0.8090**	0.6873	0.7534
4	**0.7467**	0.6814	0.6844
5	0.9561	0.9555	**0.9701**
Sum	4.1186	4.0571	**4.216**
Majority Vote	L	L	**W**

Figure 5.33 shows the grade profile for G-III for this configuration. It can be seen that for both cases, features 1, 2 and 5 classified the correct grade where as feature 4 never classified the correct grade and also produced lower similarity scores. Feature 3 was inconsistent as it classified correct grade for case 1 only. We conclude that features 1, 2 and 5 are the most suitable to be used as bench mark features to classify the correct grade for this configuration.



(a) Case 1



(b) Case 2

Figure 5.33: Grade profile for two cases of G-III (Alternative configuration)

Table. 5.21 shows a summary of all grade profiles plotting grades and features that

classify a grade correctly. In general, it can be observed that features 4 and 5 represent G-I clearly. In case of G-II, no feature was able to consistently classify correctly and for G-III, features 1, 2 and 5 were able to perform well for the configuration.

Table 5.21: Summary of grade profiles with the alternative configuration

Grades	Feature 1	Feature 2	Feature 3	Feature 4	Feature 5
	Correctly classified / Total test cases				
I	0/2	0/2	1/2	**2/2**	**2/2**
II	0/2	0/2	0/2	1/2	1/2
III	**2/2**	**2/2**	1/2	0/2	**2/2**

If we compare Table. 5.16 and Table. 5.21, we can observe that for both the configurations, feature 5 commonly classified G-I correctly. For G-II, features 1 and 2 were found to be the closest for classifying it from others. For G-III, feature 5 was the common feature for both configurations for correct classification.

Table 5.22 shows a summary in terms of correct classification of each test case by both methods. It can be observed that for G-I cases, both summation and majority vote perform fairly (1 out of 2 correct). For G-II, both method's results are poor and they are not able to classify a single test case correctly. For G-III, both methods have shown good results. The results indicate that G-II has been the most complex one to classify and no method has been able to classify this grade correctly.

Table 5.22: Summary of results with test cases by the summation and majority vote method (alternative configuration)

Type	Test Cases	Correct Classification Summation	Incorrect Classification Summation	Correct Classification Majority vote	Incorrect Classification Majority vote
G-I	2	1	1	1	1
G-II	2	0	2	0	1 (1 Tie)
G-III	2	2	0	2	0

In brief, we have seen that for both configurations, G-I and G-III were distinguishable from other grades based on certain features where as classification of G-II is found to be very complex. Our experiments have shown that selection of different features from same

spectra classify for the same grade differently, and zGT-II fuzzy sets with interval data may provide guidance in classification of breast cancer grading.

5.8 Summary

In this chapter we have described a model to create zGT-II fuzzy sets from interval data and use them as prototypes for classification of cancer grades from unseen data. Features were extracted from spectral data and then used to create first T-I fuzzy sets, and then zGT-II fuzzy sets from interval data. An approximate method to create fuzzy sets in a computationally efficient way has also been described. Two different test data configurations were tested on the model. This was done by creating T-I fuzzy sets from unseen data and then measuring the similarity between the T1 sets and prototype zGT-II fuzzy sets using a new method found in the literature. A discussion on results obtained by testing data in terms of grade profiles was also described. Results indicate that features perform differently for different grades and different configurations. G-III was found to be consistently classified followed by G-I. G-II was found to be the most difficult to classify and is generally confused with G-III. Results also indicate that the prototype was able to provide useful information even in case of false classification of grade. In the next chapter, we further evaluate the model on a new data set from the literature and analyse the performance of the model.

Chapter 6

Model Evaluation

This chapter includes the testing of the model proposed in Chapter 5 on a new FTIR data set of oral cancer patients to differentiate between tumour and stroma cells from 3 patients. Five features have been created to construct the prototype model, it has been tested against unseen data to evaluate its performance.

6.1 Data set Description

This data set was originally used by Wang [109] to distinguish between tumour and stroma cells using clustering algorithms. The data set is a combination of 7 individual spectral data sets obtained from three different oral cancer patients. The spectra have been collected from the 900-1800 cm^{-1} spectral region. The spectral data has been pre-processed with base line correction and normalisation by standard procedures. A brief summary of these data sets is given in Table 6.1. It can be seen that data sets 4 and 7 have additional spectral classifications other than tumour and stroma. It is also important that in the original work all data sets were used separately with clustering algorithms.

Table 6.1: Original Oral cancer data Set

Data set No.	Total Spectra	Tumour	Stroma	Any other
1	15	10	5	0
2	18	10	8	0
3	11	8	3	0
4	31	12	7	12
5	30	18	12	0
6	15	10	5	0
7	42	21	14	7

We have made slight modifications to the data set for our work. They are

- We have excluded the data for additional classifications from our experiments (in data set 4 and data set 6 because of availability of lower number of spectra available to build the system)
- We have set a minimum of 5 spectra for a data set to be included so data set 3 has been excluded
- We shall be using 10 spectra each for tumour and stroma for each data set where available (5 for training and 5 for testing) and the remaining spectra will not be included

After these modifications, we have 6 data sets. We have re-numbered them for our convenience. The data set used for evaluation is shown in Table 6.2. It can be seen that data sets 1, 2, 3, and 5 have only 5 spectra for stroma cells. The reason is that another group of 5 spectra (the number required to create a data set) was not possible from the original data sets. We shall be testing with unseen stroma cells for data sets 4 and 6 while unseen tumour cells will be tested for all 6 data sets. The final spectral data sets used for both training and testing of tumour and stroma cells are described in Tables 6.3 and 6.4 respectively. As this data set consists of data from 3 different patients and for two types of cells namely tumour and stroma, it contains both intra-patient and inter-patient variability for tumour and stroma cell classification. We shall be using our developed zGT-II fuzzy model for the classification of these two cell types.

Table 6.2: Data set for evaluation

Data set No.	Total Spectra	Tumour	Stroma
1	15	10	5
2	15	10	5
3	15	10	5
4	20	10	10
5	15	10	5
6	20	10	10

Table 6.3: Final data set for tumour data

Data Set	Training	Testing
1	5	5
2	5	5
3	5	5
4	5	5
5	5	5
6	5	5

Table 6.4: Final data set for stroma data

Data Set	Training	Testing
1	5	0
2	5	0
3	5	0
4	5	5
5	5	0
6	5	5

6.2 Evaluation of Model Frame Work

Evaluation of the model will follow the same steps as described in Chapter 5. Now we describe each of them for this data set.

6.2.1 Feature Extraction

As in previous experiments, a set of 5 features has been selected for this data set. A sample spectrum from the data set with approximate locations of features is shown in

Figure 6.1. The bar indicates the approximate area covered by each feature. The minimum and maximum absorbance values obtained from the features serve as interval data for the model. The features have been selected from similar locations to our previous models. The only change is that this data set starts at 900 cm $^{-1}$ wave number rather than 1000 cm^{-1} wave number, therefore, for feature 1, the minimum value of interval data has been calculated from 900 cm^{-1}. A brief description of these features is as follows.

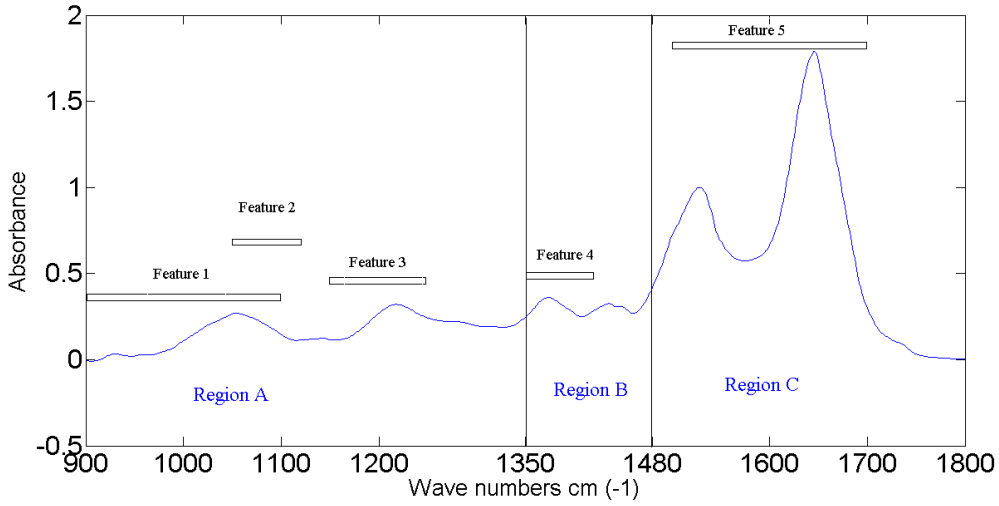


Figure 6.1: An example of a sample spectrum with regions and approximate locations of features

Feature 1: This feature consists of the minimum absorbance values from the region 900-1020 cm^{-1} and maximum peak spectral absorbance values between region 1000-1100 cm^{-1}. This feature has been selected to cover the highest distinct left peak available in the region with the left most lowest value in the region.

Feature 2: This feature consists of the peak height absorbance values 1000-1100 cm^{-1} and minimum absorbance values in region 1050-1120 cm^{-1}. The aim is to cover maximum left side peak in the region and associate it with a negative peak in the region.

For features 1 and 2, the peak height value is the same. The aim is to see how these two features whose one value is common respond in the model.

Feature 3: This feature consists of minimum absorbance values in the region 1150-1220 cm^{-1} and peak absorbance values of 1200-1250 cm^{-1}. The feature was selected to

cover second main peak in region A.

For region B, one feature has been selected.

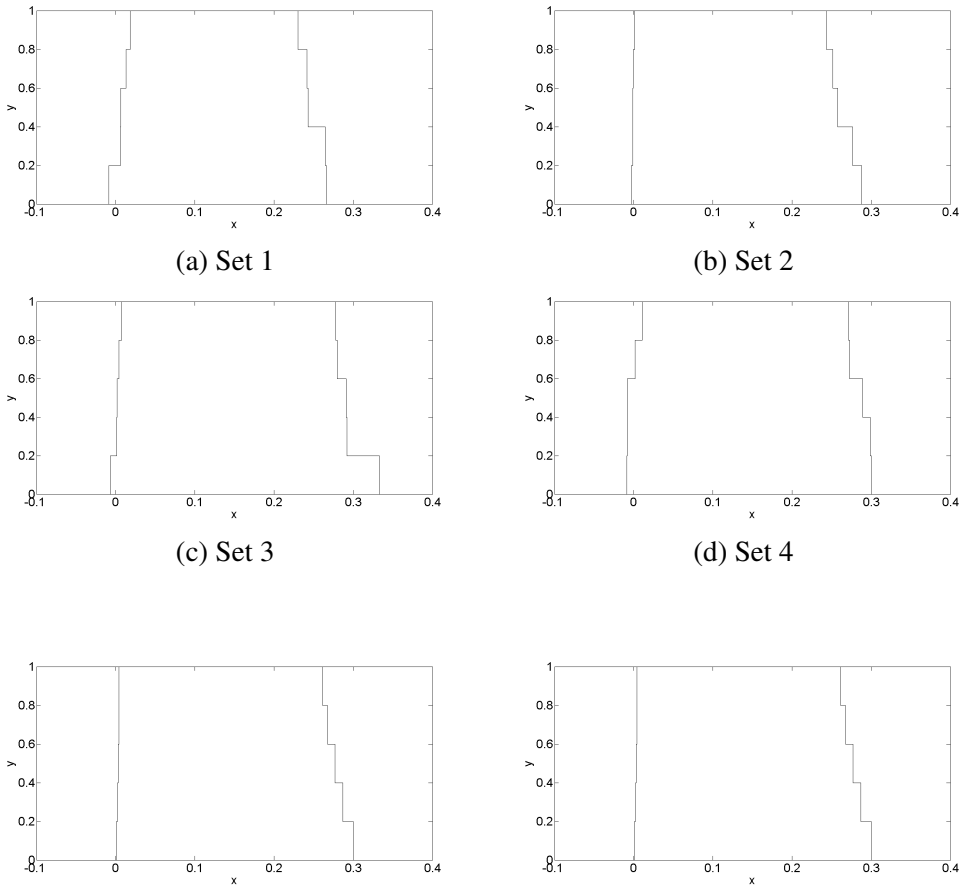
Feature 4: This feature consists of minimum absorbance value in the region 1350-1400 cm^{-1} and peak absorbance value in the region 1350-1420 cm^{-1}.

For region C, one feature has been selected.

Feature 5: This feature consists of the peak heights of Amide-I and Amide-II regions as interval data. This feature has been selected to cover the two most distinct peaks in the spectra. Amide-II peak height is the maximum absorbance values in the region 1500-1600 cm^{-1} and Amide-I peak is the maximum peak absorbance value in the region 1600-1700 cm^{-1} wave numbers.

6.2.2 Construction of Type-I Fuzzy Sets

Five spectra from each data set were used to create T-I fuzzy sets for each feature. 6 T-I fuzzy sets were created for each feature for tumour and stroma cells for each data set. Figure 6.2 shows the 6 T-I fuzzy sets for feature 1 for stroma cells. The x-axes shows the domain values for the fuzzy set and y-axes shows the membership grade values.



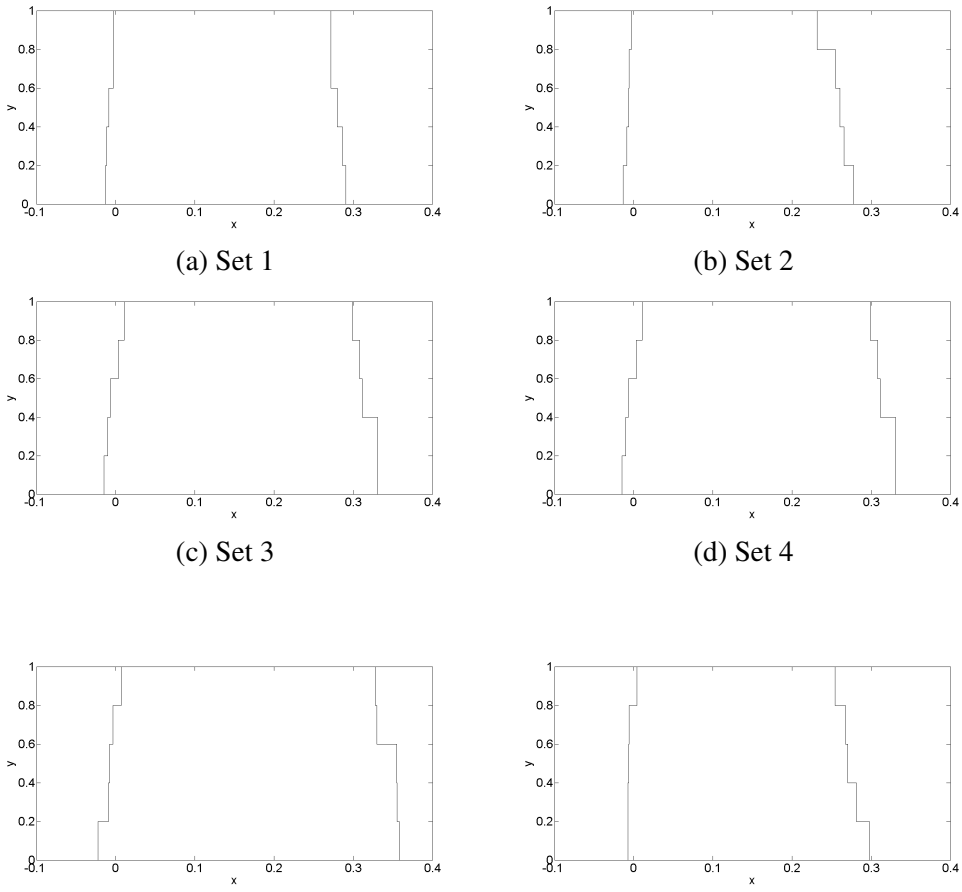
(a) Set 1 (b) Set 2

(c) Set 3 (d) Set 4

(e) Set 5 (f) Set 6

Figure 6.2: T-I fuzzy sets for stroma cells with feature 1

These fuzzy sets incorporate the intra-patient variability found in the stroma cells of each patient. Similarly for tumour cells, intra-patient variability between spectra can be seen in Figure 6.3 with an example of 6 T-I fuzzy sets for feature 1 for 6 data sets.



(a) Set 1 (b) Set 2

(c) Set 3 (d) Set 4

(e) Set 5 (f) Set 6

Figure 6.3: T-I fuzzy sets for tumour cells with feature 1

Similarly, T-I fuzzy sets were created for other 4 features for the tumour and stroma cells. In total, we have 30 T-I sets for five features for tumour and stroma cells each consisting of 5 spectral interval values.

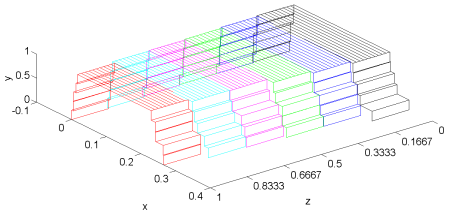
6.2.3 Construction of zGT-II Fuzzy Sets

These 6 T-I sets have been combined to create a zGT-II fuzzy set for each feature for tumour and stroma cells as described below.

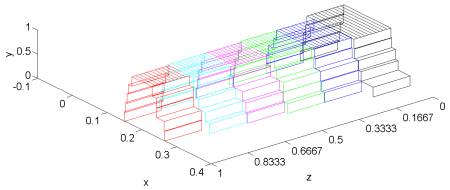
- Total Number of zGT-II fuzzy sets: 10 (2 for each feature for tumour and stroma)
- Number of zSlices in each Set: 6

- Number of classifications to be made : 2 (tumour and stroma)

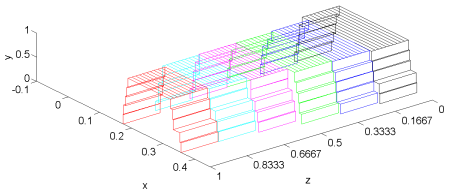
These zGT-II fuzzy sets were created using the same method as described in Chapter 5. The created zGT-II fuzzy sets serve as a bench mark prototype and are used for the testing of unseen data. The zGT-II fuzzy sets for stroma cells for all five features are shown in Figure 6.4 where z-axes shows the secondary membership grades and x and y-axes have got domain values and primary membership grades respectively.



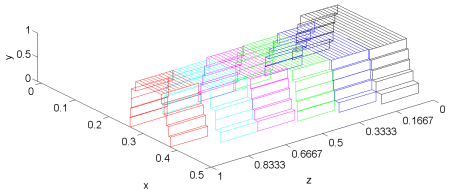
(a) Feature 1



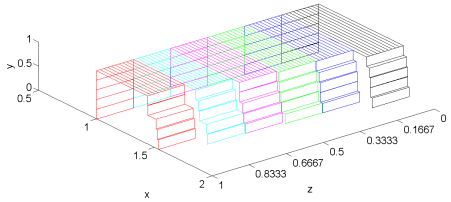
(b) Feature 2



(c) Feature 3



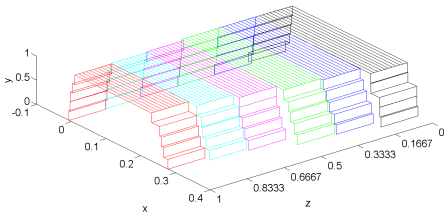
(d) Feature 4



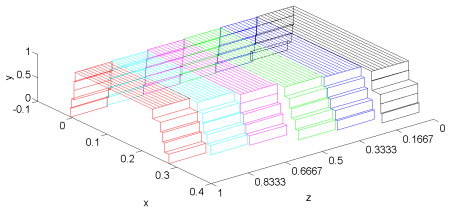
(e) Feature 5

Figure 6.4: zGT-II fuzzy sets for stroma cells for five features

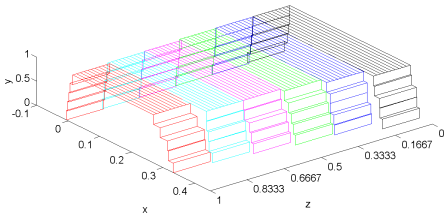
The zSlices on z-axes represent the variability found in the T-I fuzzy sets in terms of inter-patient variability, that is, variability found between different patients stroma spectra. The more zSlices a zGT-II fuzzy set has, the more variability it captures. The zGT-II fuzzy sets created for tumour cells can be seen in Figure 6.5.



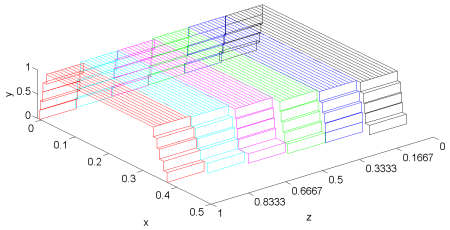
(a) Feature 1



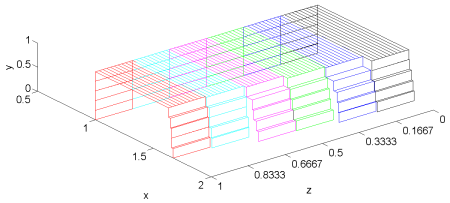
(b) Feature 2



(c) Feature 3



(d) Feature 4

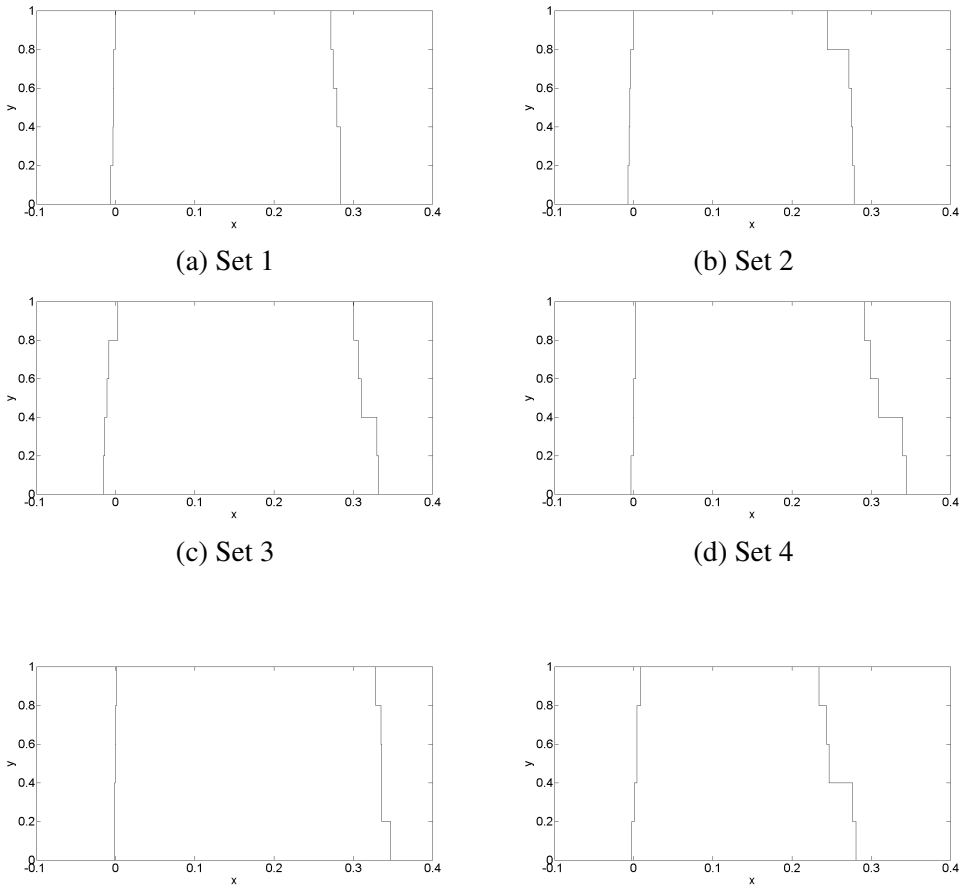


(e) Feature 5

Figure 6.5: zGT-II fuzzy sets for tumour cells for five features

6.2.4 Model Testing with Unseen Data

We have tested this prototype model on 6 T-I fuzzy sets for unseen tumour spectra and 2 T-I unseen sets for stroma spectra for each feature against the prototype zGT-II fuzzy sets with the similarity measure described in Chapter 5. The data for testing was unseen. The created T-I fuzzy sets for testing data for feature 1 for tumour cells data are shown in Figure 6.6 and for stroma cells in Fig. 6.7. Similarly T-I fuzzy sets have been created for the other 4 features for tumour and stroma cells.

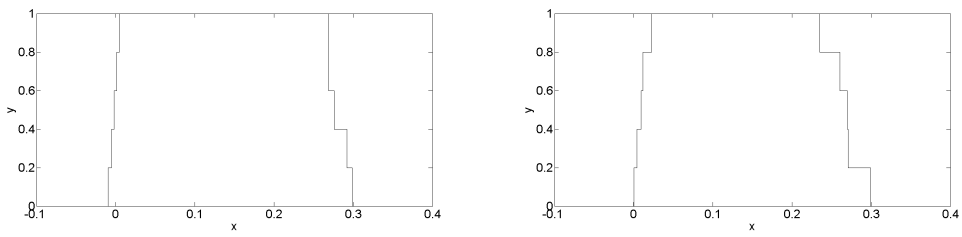


(a) Set 1 (b) Set 2

(c) Set 3 (d) Set 4

(e) Set 5 (f) Set 6

Figure 6.6: T-I fuzzy sets for tumour cells for testing data with feature 1



(a) Set 4 (b) Set 6

Figure 6.7: T-I fuzzy sets for stroma cells for testing data with feature 1

The results using two methods of majority vote and summation method for tumour test cases with their similarity scores are shown in Table 6.5 for 6 testing data sets. The highest similarity scores have been highlighted in bold. For majority vote, **W** indicates the winner and **L** indicates the losing category. It can be observed from the tables that both methods

perform well on tumour test cases. All the test data sets have been classified correctly by both methods. Although scores for features 1 and 5 are relatively very close, the majority vote correctly classifies the type of cell with these two features. The Summation method produces accurate results because the difference of similarity scores between tumour and stroma cells for features 2-4 is very large and the tumour scores are substantially higher making the summation value larger than the stroma similarity summation values.

Table 6.5: Similarity scores for tumour data sets

(a) Data Set 1

Feature	Tumour	Stroma
1	**0.9200**	0.9000
2	**0.9220**	0.3920
3	**0.9319**	0.4643
4	**0.9472**	0.2572
5	**0.8737**	0.7322
Sum	**4.5948**	2.7457
Majority Vote	**W**	L

(b) Data Set 2

Feature	Tumour	Stroma
1	**0.9180**	0.9157
2	**0.9187**	0.3893
3	**0.9295**	0.4376
4	**0.9443**	0.2730
5	**0.8954**	0.7564
Sum	**4.6059**	2.7720
Majority Vote	**W**	L

(c) Data set 3

Feature	Tumour	Stroma
1	**0.8733**	0.8070
2	**0.8719**	0.3517
3	**0.9119**	0.4844
4	**0.9288**	0.2894
5	**0.9044**	0.7650
Sum	**4.4903**	2.6975
Majority Vote	**W**	L

(d) Data set 4

Feature	Tumour	Stroma
1	**0.8659**	0.8251
2	**0.8626**	0.3646
3	**0.8798**	0.4977
4	**0.9164**	0.3170
5	**0.9394**	0.8477
Sum	**4.4641**	2.8521
Majority Vote	**W**	L

(e) Data set 5

Feature	Tumour	Stroma
1	**0.8253**	0.7738
2	**0.8292**	0.3375
3	**0.8643**	0.4907
4	**0.9074**	0.3245
5	**0.9381**	0.8422
Sum	**4.3643**	2.7687
Majority Vote	**W**	L

(f) Data set 6

Feature	Tumour	Stroma
1	0.8797	**0.9345**
2	**0.8893**	0.3785
3	**0.9184**	0.4721
4	**0.9256**	0.2703
5	**0.9357**	0.8326
Sum	**4.5487**	2.8880
Majority Vote	**W**	L

Table 6.6 shows the similarity scores and their results using the two methods for two test cases of stroma cells for data sets 4 and 6 respectively. It can be observed from the tables that both methods produced the correct classification for stroma cells. Features 1 and 2 behave differently for the two test cases of stroma cells but features 3-5 produce the same results for both methods. For the test case for data set 4, a majority vote classifies stroma cells as tumour cells for first two features but classified them correctly for remaining 3 features winning the majority vote for stroma cells by 3 votes to 2 votes. Similarly, for data set 6, feature 2 classifies incorrectly but the rest of the features made the correct classification and in this case, the majority vote won by 4 votes to 1 vote. The Summation method classifies correctly for both test cases of stroma because of the large difference between the similarity values for features 3 and 4 for stroma cells.

Table 6.6: Similarity scores for stroma for data sets

(a) Data set 4

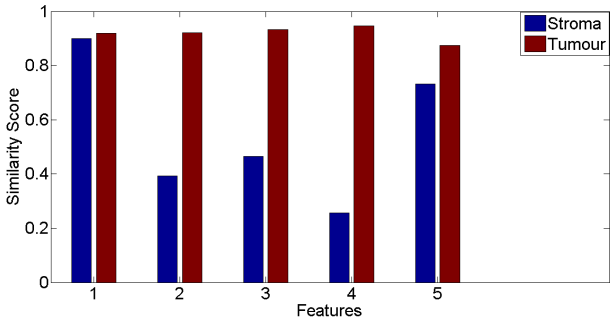
Feature	Tumour	Stroma
1	**0.9223**	0.9148
2	**0.9241**	0.3934
3	0.4256	**0.8570**
4	0.2594	**0.8065**
5	0.8710	**0.9070**
Sum	3.4024	**3.8787**
Majority Vote	L	**W**

(b) Data set 6

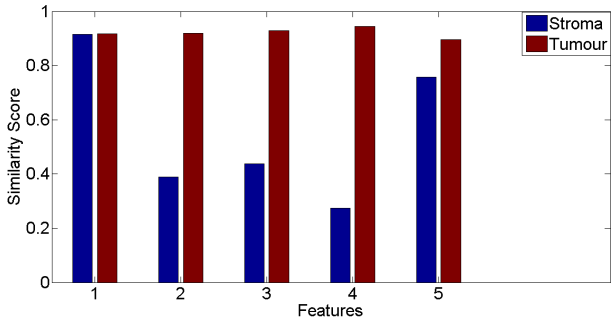
Feature	Tumour	Stroma
1	0.8799	**0.9255**
2	**0.8875**	0.4030
3	0.4680	**0.8604**
4	0.2962	**0.7990**
5	0.8580	**0.9059**
Sum	3.3896	**3.8938**
Majority Vote	L	**W**

6.3 Discussion

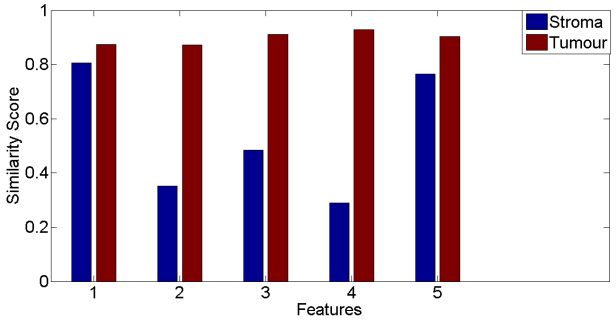
Figure 6.8 shows tumour test case profiles with similarity scores for tumour and stroma cells. It can be observed from the figure that feature 1 does not indicate a clear distinction between tumour and stroma cells and scores remain very close to each other. Feature 5 also produce high scores for both tumour and stroma cells. Features 2-4 show a clear distinction between these two types and tumour scores remain very high for tumour cells as compared to stroma cells and classify the tumour test cases correctly. We can conclude that features 2-4 can be used as bench mark features for the classification of tumour cells.



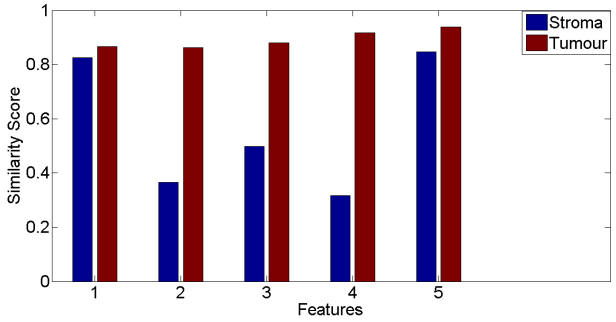
(a) Data set 1



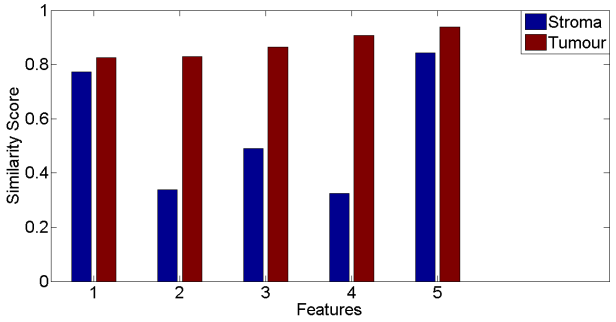
(b) Data set 2



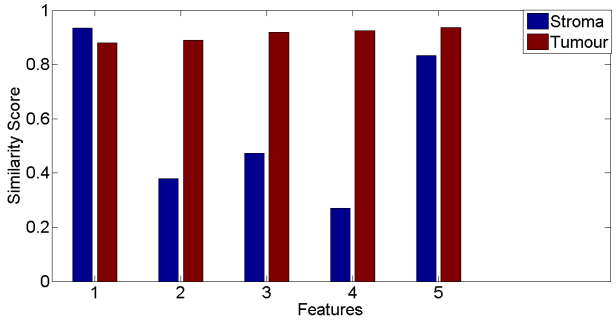
(c) Data set 3



(d) Data set 4



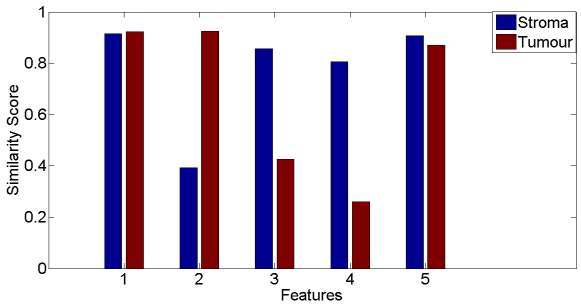
(e) Data set 5



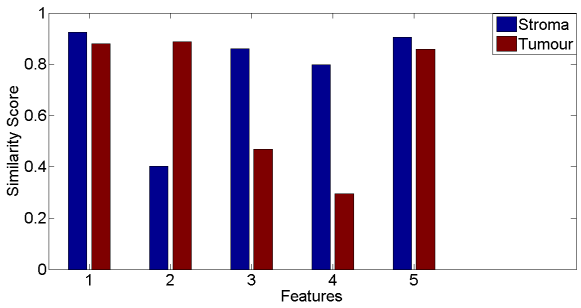
(f) Data set 6

Figure 6.8: Classification profiles for tumour cells test cases

Figure 6.9 shows profiles for two test cases of stroma cells (for data sets 4 and 6) with their similarity scores for each feature for tumour and stroma cells. It can be seen that features 1 and 2 behave differently for two test cases. In the case of data set 4, both features 1 and 2 classify stroma cells as tumour cells. Feature 5 classifies stroma cells correctly but the scores for tumour cells also remain high. In the case of data set 6, feature 2 produce significantly lower scores for stroma cells. Feature 1 classifies stroma cells correctly but scores for tumour cells also remain high. For both test cases, features 3 and 4 provide significantly higher scores for stroma cells as compared to tumour cells. We conclude that features 3 and 4 can be used as bench mark features to classify tumour cells from stroma cells.



(a) Data set 4



(b) Data set 6

Figure 6.9: Classification profiles for stroma cells test cases

Table 6.7 shows a break down of correctly classified cells of tumour and stroma by each feature's individual similarity score. It can be observed that except for feature 1, all other features are able to classify tumour cells correctly. In case of stroma cells, features 1 and 2 are not consistent but features 3-5 provide correct classification.

Table 6.7: Summary of grade profiles

Grades	Feature 1	Feature 2	Feature 3	Feature 4	Feature 5
	Correctly classified / Total test cases				
Tumour	5/6	**6/6**	**6/6**	**6/6**	**6/6**
Stroma	1/2	1/2	**2/2**	**2/2**	**2/2**

Table 6.8 shows a summary of test cases for both tumour and stroma cells using majority vote and summation methods. Both methods produce the same results for the test cases and all the test data sets have been classified correctly. The number of spectra for the creation of sets is relatively small (5) and we have only tested 2 data sets for stroma cells. Still, the results indicate that method has the capability to be used for classification of spectral data sets that have a high level of uncertainty.

Table 6.8: Summary of results with test cases by Majority vote and Summation Method

Type	Test Cases	Correct Classification	Incorrect Classification
Tumour	6	6	0
Stroma	2	2	0

Features 1 and 2 have the same peak value but do not behave identically, feature 1 is not able to provide clear distinctive scores for both tumour and stroma cell classification, while feature 2 has performed well for tumour cell classification. As in Chapter 5, it shows that a feature with a common value can still be useful in providing useful information for classification.

6.4 Comparison with Original Results

Table 6.9 shows the results obtained in the original study by Wang in differentiating cancer and stroma cells [109]. In this study, HCA, k-means and FCM clustering algorithms were used by setting number of clusters as 2 and matching it with the clinical study. Table 6.9 shows the best results obtained by each clustering algorithm in the original study. 'Data

set No. modified' is the reference number for each data set used in this study where as 'Data set No. original' is the data set number used in the original study. It can be seen from the table that generally all three clustering algorithms have produced good results except for the data set 7 of the original study where results are not good as compared to other data sets. Our results with the newly proposed model have also produced equally good results as discussed in the previous section. It shows that our proposed method is able to compete well in terms of providing near equivalent results of the original study. Although in our study, the number of spectra is reduced because of the requirement of the model, still results are generally as good as in the original study.

Table 6.9: Results of original study

Data set No. Modified	Data set No. Original Study	Tissue Type	Clinical Study	HCA	k-means	FCM
1	1	Tumour	10	10	10	10
		Stroma	5	5	5	5
2	2	Tumour	10	9	9	9
		Stroma	8	9	9	9
3	4	Tumour	12	12	11	11
		Stroma	7	7	8	8
4	5	Tumour	18	18	17	14
		Stroma	12	12	13	16
5	6	Tumour	10	10	10	10
		Stroma	5	5	5	5
6	7	Tumour	21	28	17	18
		Stroma	14	13	18	16

Our experiments have shown that zGT-II fuzzy sets can be created with the help of spectral data sets by extracting features as interval data. Results with Oral cancer achieved by the proposed method are as good as in the original study with clustering algorithms. Although results are based on a smaller data set but they indicate that the proposed model can be applied on any FTIR spectral data set involving higher level of complexities and with higher order data sets, statistical significance of the model can be evaluated.

6.5 Summary

In this Chapter, we have evaluated the model prototype from Chapter 5 on a different data set of oral cancer patients. Five features were selected, and 5 spectra were used to create T-I fuzzy sets for differentiating between tumour and stroma cells. These T-I fuzzy sets were combined to create zGT-II fuzzy sets for all features for both types of cells. The prototype model zGT-II fuzzy sets were tested against unseen T-I fuzzy sets for both tumour and stroma cells and the results were obtained by two methods, majority vote and summation of scores of features. Profiles based on similarity scores for both tumour and stroma cells were also created to get in depth information regarding the behaviour of features. The results indicate that the proposed prototype model is able to produce appreciable results in classifying unseen tumour and stroma cells. Although the number of spectra is low, the results show that the proposed model can be created with independent spectral data sets and bench mark zGT-II fuzzy sets can be created that include various types of complex uncertainties involved in real spectral data sets, and may become a tool for solving real world classification problems, especially with cancer spectral data sets. In the next chapter, we conclude the findings contributed by this thesis and report directions towards future work.

Chapter 7

Conclusions and Future Work

This chapter concludes the research work done during the course of this PhD project. It also identifies the contributions made to the knowledge during the research. A list of possible future work is also provided in this chapter. The chapter ends with a list of papers published or in preparation for submission from the research work.

7.1 Conclusions

Breast cancer has become a major cause of death among women throughout the world. Use of Computer based technique to help in the diagnosis and prognosis of cancer is a common area of world wide research. Fourier Transform Infra-red Spectroscopy (FTIR) is one of the techniques that has been frequently used for cancer data. A main reason is that this technique has ability to identify small molecular changes found in the cell relatively easily which is not possible by microscopic evaluation of cancer cells. After the diagnosis, predicting the long term survival for the patients is important and cancer grading is a critical parameter in it which is part of world wide accepted Nottingham Prognostic Index(NPI). Cancer grade is found with the help of Nottingham Grading System (NGS) method accepted world wide. Manual classification of cancer grade is done by experts by observing a cancer sample under microscope to assign it as Grade-I (G-I), Grade-II (G-II) or Grade-III (G-III). This manual method has a higher probability of

errors and automated methods can assist the experts in grade classification.

In this thesis, we have investigated the use of advanced computational methods with FTIR based spectral data for classification of breast cancer grading. In the real world, cancer spectral data is complex and obtained from multiple patients of the 3 grades. There are two types of variabilities involved in it, one is between spectra obtained from one case of a patient (intra-case) and other is between different patients of different cases (inter-case) for each of three grades. We have used standard clustering algorithms followed by the use of a supervised learning method to create a Fuzzy Inferencing System (FIS) for grade classification and have shown that traditional methods are unable to address the complicated classification problem. We have shown a new method using zSlices based General Type-II fuzzy sets (zGT-II) fuzzy sets to create bench mark prototype models that can be saved in a data base and compared against unseen spectral data for grade classification. To the best of our knowledge, it is the first attempt of its kind to create interval data from different features from spectral regions and then create prototype zGT-II fuzzy models from it as a bench mark for grade classification. Now, we describe the summary of the work carried out.

7.2 Summary

In Chapter 2, a comprehensive literature review on spectral data is given. It covers topics including breast cancer, NPI, NGS, Spectral pre-processing, spectral features extraction, the standard clustering algorithms of k-means and fuzzy c-means clustering (FCM), Type-I (T-I) fuzzy logic and FIS. This chapter also covers Type-II (T-II) fuzzy logic with a special emphasis on zGT-II fuzzy sets and similarity measures used with T-II fuzzy sets.

In Chapter 3, three different types of spectral data sets have been used with unsupervised learning methods using the standard clustering algorithms of k-means and FCM. Each data set increases the complexity of the data. We have shown that in the case of data set 3 where both intra-grade and inter-grade variabilities were involved, standard

clustering algorithms are not able to classify the grade successfully.

In Chapter 4, a supervised learning mechanism has been used for grade classification for data set 3. A Mamdani type FIS has been created with three outputs each representing a grade. Principal Component Analysis (PCA) was used to reduce the dimensionality of the data sets and Hill Climbing (HC) and Simulated Annealing (SA) methods along with the first three PCs have been used to train the membership functions and rules of a FIS that can predict the breast cancer grade. The developed FIS was tested on unseen data. Results indicated that the proposed method was able to distinguish between G-I and G-III reasonably well, but was unable to classify the G-II. The results also showed that this method was not good enough for dealing with the complexities involved in the data set.

In Chapter 5, a novel model based on zGT-II fuzzy sets has been described. It is the first reported attempt to the best of our knowledge to create such model based on zGT-II fuzzy sets on spectral data sets. The model starts by extracting 5 key features based on certain peak heights and trough values selected from spectral regions to create interval data. T-I fuzzy sets are created for each feature for each grade. These T-I fuzzy sets incorporate the variabilities involved with in the spectra of a case (intra-case). An approximate method has been developed that creates fuzzy set from interval data with substantial reduction in computation time. From examples it has been shown that this approximate method works well for highly overlapped data but as overlapping reduces, the results deviate lot from those produced by the original algorithm to create fuzzy sets from interval data. zGT-II fuzzy sets have been created for each grade by combining the T-I fuzzy sets. These zGT-II fuzzy sets contain the inter-case variabilities found in the data set. These zGT-II fuzzy sets serve as bench mark for unseen data and T-I fuzzy sets from unseen data were compared against these bench mark prototypes and similarity scores were used to classify the grade. Majority vote and summation of similarity methods were used for final classification. Results indicated that proposed model was able to work well for G-I and G-III but for G-II, it did not perform well. Profiles for each grade were created for in depth analysis of their similarity scores with a discussion to find the complexities

that resulted in this performance. An alternative data set created with a different set of G-II cases was also used for the analysis of the model and results were discussed. Overall, the model did not work well for G-II but showed the potential that it can be used as an alternative method for extraction of key information from complex cancer spectral data set.

In Chapter 6, the model created in Chapter 5 is evaluated on a new data set. This data set is a spectral data set of 3 oral cancer patients and consists of 6 data sets. The model was used to classify tumour and stroma cells. The results show that the proposed model is able to classify between tumour and stroma cells on a small sample data set. A draw back of this evaluation was that number of spectra was low. 5 spectra were used to create a fuzzy set. As such, the evaluation can not be proven statistically but the model creation indicates the potential of this method to be used for real spectral data sets where various types of uncertainties and variabilities are involved. We have tried to obtain novel breast cancer spectral data sets from various patients for all three grades from Nottingham Breast Cancer Research Group and School of Chemistry, University of Nottingham and are still in the process of obtaining those data sets as they may be used to find statistically significant results and further investigate for the improvement of the model.

7.3 Contributions to the Knowledge

Following are the main contributions to the knowledge made by this PhD project.

- Creating interval data from spectral data set
- A newly developed method to create fuzzy sets from interval data in a computationally efficient manner
- Development of fuzzy sets (Type-I & II) based on interval data extracted from features from spectral data set
- Development of a step wise model for classification of unseen spectra by using a

similarity measure with known benchmark zGT-II fuzzy sets

7.4 Limitations

Following is a list of limitations of the work carried out.

- The proposed model can only work with interval data and does not support any other type of data
- The proposed model does not work when interval data is non-overlapping
- The proposed model is not able to distinguish between Grade-II and Grade-III which is considered a complicated problem in cancer pathology as well

7.5 Directions towards Future Work

The following suggestions are made to carry this work forward.

- For the FTIR spectra, we have used basic pre-processing methods. It will be interesting to use various other pre-processing techniques found in the literature and to develop an automated method that can identify the best pre-processing technique for a particular raw spectral data set.
- For clustering algorithms, we have used Euclidean distance. Other type of distance measures can be used and results can be compared. Examples of other distance measures are, squared Mahalanobis distance, mutual neighbour distance and the Chebychev distance
- For FIS, we only used 3 PCs to keep initial rules manageable. Different number of PCs could be used and also there is scope for developing a method that can help in rule reduction or rule optimisation

- We have used HC and SA for membership functions and rule optimisation, other methods like GAs could also be used and the results compared to find the optimal method

- We have not used any automated method to extract features from spectral data. There is scope for developing an automated method that can look at various peak heights and troughs involved in spectra and find an optimal number of features

- A FIS can be developed that uses membership functions of zGT-II fuzzy sets instead of the commonly used T-I fuzzy sets. A comparison of FIS, with T-I and zGT-II fuzzy sets could also be made

- We have used only one new similarity measure for finding the similarity between zSlices. It is important to use other similarity measures as well and compare their results to select a particular similarity measure best suited to the data set

- We have mainly investigated breast cancer spectral data sets for this research. In future, various cancer spectral data sets of different cancer types or any other classification problem with data with similar characteristics can be used and their model prototypes can be used to develop an expert system. Such an expert system could be used to make different categorisations for unlabelled spectra

- A large volume of spectra can be used for model creation. Although it will be computationally very expensive to handle such a high volume. As technology is growing very rapidly, it is very likely that it will be possible easily in the near future

- It is proposed to develop a dedicated software sub system built into Spectrometer that can identify key features out of the spectral data and report them. Although it will require more time and effort to develop such a system, it will make the process of analysing features and extracting key information from them significantly easier

7.6 Publications

The following is a list of publications that are due to be submitted or have been published coming from this research work with a reference to the relevant chapter.

1. Shabbar Naqvi, Simon Miller and Jonathan M. Garibaldi: *A Type-II Fuzzy Logic based Model for the Classification of Breast Cancer Grading from FTIR Spectral Data Sets*, (in preparation) for submission in *IEEE Transactions on Fuzzy Systems* [Chapters 5 & 6]

2. Shabbar Naqvi, Simon Miller and Jonathan M. Garibaldi: *A Fuzzy Inferencing System for Breast Cancer Grade Classification with Membership Functions and Rules Tuning with a Spectral Data Set*, (in preparation) for submission in *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* [Chapter 4]

3. Shabbar Naqvi, Simon Miller and Jonathan M. Garibaldi: *A General Type-II Similarity Based Model for Breast Cancer Grading with FTIR Spectral Data* (Accepted) in *FUZZ IEEE 2014 (IEEE International Conference on Fuzzy Systems), July 6-11, 2014, Beijing, China* [Chapter 5]

4. Shabbar Naqvi, Simon Miller and Jonathan M. Garibaldi: *Towards Development of a Fuzzy Inferencing System for the Automation of Breast Cancer Grading with Spectral Data Sets* (Abstract publication) in *The 2012 Mini EURO Conference on Computational Biology, Bioinformatics and Medicine, University of Nottingham, September 2012* [Chapter 4]

5. Shabbar Naqvi and Jonathan M. Garibaldi: *The complexities involved in the analysis of Fourier Transform Infrared Spectroscopy of breast cancer data with clustering algorithms* in *3rd Computer Science and Electronic Engineering Conference (CEEC 2011), Colchester, 2011, U.K, pages:80-85* [Chapter 3]

6. Shabbar Naqvi and Jonathan M. Garibaldi: *An Investigation into the use of Fuzzy C-Means Clustering of Fourier Transform Infrared Microscopic Data for the Au-*

tomation of Breast Cancer Grading in *The 9th Annual Workshop on Computational Intelligence (UKCI 2009), Nottingham, 2009, U.K* [Chapter 3]

Bibliography

[1] Figures on health topics (http://www.about.com). Retrieved on 13th January, 2013.

[2] San diego mirror college lectures (http://www.sdmiramar.edu). Retrieved on 28th August, 2013.

[3] J. Anastassopoulou, P. Arapantoni, E. Boukaki, S. Konstadoudakis, T. Theophanides, C. Valavanis, C. Conti, P. Ferraris, G. Giorgini, S. Sabbatini, and G. Tosi. Micro-ftir spectroscopic studies of breast tissues. In Vasili Tsakanov and Helmut Wiedemann, editors, *Brilliant Light in Life and Material Sciences*, NATO Security through Science Series, pages 273–278. Springer Netherlands, 2007.

[4] J. Anastassopoulou, E. Boukaki, C. Conti, P. Ferraris, E. Giorgini, C. Rubini, S. Sabbatini, T. Theophanides, and G. Tosi. Microimaging ft-ir spectroscopy on pathological breast tissues. *Vibrational Spectroscopy*, 51(2):270 – 275, 2009.

[5] S. Auephanwiriyakul, S. Attrapadung, S. Thovutikul, and N. Theera-Umpon. Breast abnormality detection in mammograms using fuzzy inference system. In *The 14th IEEE International Conference on Fuzzy Systems, 2005 (FUZZ '05)*, pages 155 –160, may 2005.

[6] J. Backhaus, R. Mueller, N. Formanski, N. Szlama, H.G. Meerpohl, M. Eidt, and P. Bugert. Diagnosis of breast cancer with infrared spectroscopy from serum samples. *Vibrational Spectroscopy*, 52(2):173 – 177, 2010.

[7] Wyllis Bandler and Ladislav Kohout. Fuzzy power sets and fuzzy implication operators. *Fuzzy Sets and Systems*, 4(1):13 – 30, 1980.

[8] W. Banzhaf, P. Nordin, R. Keller, and F. Francone. *Genetic Programming: An Introduction*. Morgan Kaufmann Publishers, San Francisco, CA, 1999.

[9] A. Benard, C. Desmedt, V. Durbecq, G. Rouas, D. Larsimont, C. Sotiriou, and E. Goormaghtigh. Discrimination between healthy and tumor tissues on formalin-fixed paraffin-embedded breast cancer samples using ir imaging. *Spectroscopy*, 24:67–72, 2010.

[10] J. Bezdek. *Pattern Recognition with Fuzzy Objective Function Algorithms*. Kluwer Academic Publishers Norwell, MA, USA, 1981.

[11] R. Bhargava. Towards a practical fourier transform infrared chemical imaging protocol for cancer histopathology. *Analytical and Bioanalytical Chemistry*, 389(4):1155–1169, 2007.

[12] X.Y Bin, L. Qian, H.E. Fei, G. Chun-guang, W. Cheng-feng, and Z. Ping. Diagnosis of colon cancer with fourier transform infrared spectroscopy on the malignant colon tissue samples. *Chinese Medical Journal*, 124(16):2517–2521, 2011.

[13] Biomax. Extract from biomax website (http://www.biomax.us). Retrieved on 28th February 2013.

[14] B. Bird, M. Miljkovic, M. Romeo, J. Smith, N. Stone, M. George, and M. Diem. Infrared micro-spectral imaging: distinction of tissue types in axillary lymph node histology. *BMC Clinical Pathology*, 8(1):8, 2008.

[15] N.J. Bundred. Prognostic and predictive factors in breast cancer. cancer treatment reviews. *Cancer Treatment Reviews*, 27:137, 2001.

[16] H. Bustince. Indicator of inclusion grade for interval-valued fuzzy sets. application to approximate reasoning based on interval-valued fuzzy sets. *International Journal of Approximate Reasoning*, 23(3):137 – 209, 2000.

[17] M. Castanys, R. Perez-Pueyo, M. J. Soneira, E. Golobardes, and A. Fornells. Identification of raman spectra through a case-based reasoning system: application to artistic pigments. *Journal of Raman Spectroscopy*, 42(7):1553–1561, 2011.

[18] C. Cernuda, E. Lughofer, W. Mrzinger, and J. Kasberger. Nir-based quantification of process parameters in polyetheracrylat (pea) production using flexible non-linear fuzzy systems. *Chemometrics and Intelligent Laboratory Systems*, 109(1):22 – 33, 2011.

[19] X. Chen, D. Wu, Y. He, and S. Liu. Detecting the quality of glycerol monolaurate: A method for using fourier transform infrared spectroscopy with wavelet transform and modified uninformative variable elimination. *Analytica Chimica Acta*, 638(1):16 – 22, 2009.

[20] L.F. Chiu, P.Y. Huang, W.F. Chiang, T.Y. Wong, S.H. Lin, Y.C. Lee, and D.B. Shieh. Oral cancer diagnostics based on infrared spectral markers and wax physisorption kinetics. *Analytical and Bioanalytical Chemistry*, 405(6):1995–2007, 2013.

[21] S. Chumklin, S. Auephanwiriyakul, and N. Theera-Umpon. Microcalcification detection in mammograms using interval type-2 fuzzy logic system with automatic membership function generation. In *IEEE International Conference on Fuzzy Systems (FUZZ), 2010*, pages 1–7, 2010.

[22] S. Coupland and R. John. Geometric type-2 fuzzy sets. In Alireza Sadeghian, Jerry M. Mendel, and Hooman Tahayori, editors, *Advances in Type-2 Fuzzy Sets and Systems*, volume 301 of *Studies in Fuzziness and Soft Computing*, pages 81–96. Springer New York, 2013.

[23] E. Cox. *The Fuzzy Systems Handbook: A Practitioners Guide to Building, Using and Maintaining Fuzzy Systems*. San Diego: CA:AP Professionals, 1999.

[24] I. David and R.G. Ellis. Metabolic fingerprinting in disease diagnosis: biomedical applications of infrared and raman spectroscopy. *The Analyst*, 131:875, 2006.

[25] M. J. de Paula Castanho, C. Lacio de Barros, A. Yamakami, and L. L. Vendite. Fuzzy expert system: An example in prostate cancer. *Applied Mathematics and Computation*, 202(1):78 – 85, 2008.

[26] G. D'Eredita, C. Giardina, M. Martellotta, T. Natale, and F. Ferrarese. Prognostic factors in breast cancer: the predictive value of the nottingham prognostic index in patients with a long-term follow-up that were treated in a single institution. *European Journal of Cancer*, 37:591, 2001.

[27] J. C. Dunn. A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters. *Journal of Cybernetics*, 3(3):32–57, 1973.

[28] I.O. Ellis, M. Galea, N. Broughton, A. Locker, R.W. Blamey, and C.W. Elston. Pathological prognostic factors in breast cancer. ii. histological type relationship with survival in a large study with long-term follow-up. *Histopathology*, 20:479, 1992.

[29] A. G. Evsukoff, A. C.S. Branco, and S. Galichet. Intelligent data analysis and model interpretation with spectral analysis fuzzy symbolic modeling. *International Journal of Approximate Reasoning*, 52(6):728 – 750, 2011.

[30] H. Fabian, N. A. N. Thi, M. Eiden, P. Lasch, J. Schmitt, and D. Naumann. Diagnosing benign and malignant lesions in breast tissue sections by using ir-microspectroscopy. *Biochimica et Biophysica Acta (BBA) - Biomembranes*, 1758(7):874 – 882, 2006.

[31] M. Filippone, F. Masulli, and S. Rovetta. Simulated annealing for supervised gene selection. *Soft Computing*, 15:1471–1482, 2011.

[32] G.D. Francis, S. R. Stein, and G.D. Francis. Prediction of histologic grade in breast cancer using an artificial neural network. In *The 2012 International Joint Conference on Neural Networks (IJCNN)*, pages 1–5, 2012.

[33] J. Fuentes-Uriarte, M. Garcia, and O. Castillo. Comparative study of fuzzy methods in breast cancer diagnosis. In *Annual Meeting of the North American Fuzzy Information Processing Society, 2008. NAFIPS 2008*, pages 1 –5, may 2008.

[34] R. Giles. ?ukasiewicz logic and fuzzy set theory. *International Journal of Man-Machine Studies*, 8(3):313 – 327, 1976.

[35] A.R. Gonalves, E. Esposito, and P. Benar. Evaluation of panus tigrinus in the delignification of sugarcane bagasse by ftir-pca and pulp properties. *Journal of Biotechnology*, 66(23):177 – 185, 1998.

[36] M.M. Gupta and J. Qi. Theory of t-norms and fuzzy inference methods. *Fuzzy Sets and Systems*, 40(3):431 – 450, 1991. Fuzzy Logic and Uncertainty Modelling.

[37] H. Hagras and C. Wagner. Towards the wide spread use of type-2 fuzzy logic systems in real world applications. *IEEE Computational Intelligence Magazine*, 7(3):14–24, 2012.

[38] H.A. Hagras. A hierarchical type-2 fuzzy logic control architecture for autonomous mobile robots. *IEEE Transactions on Fuzzy Systems*, 12(4):524–539, 2004.

[39] G. Hamerly and C. Elkan. Alternatives to the k-means algorithm that find better clusterings. In *Proceedings of the eleventh international conference on Information and knowledge management*, CIKM '02, pages 600–607, New York, NY, USA, 2002. ACM.

[40] H. Hamrawi, S. Coupland, and R. John. A novel alpha-cut representation for type-2 fuzzy sets. In *IEEE International Conference on Fuzzy Systems (FUZZ), 2010*, pages 1–8, 2010.

[41] J.L. Haybittle, R.W. Blamey, and C.W. Elston. A prognostic index in primary breast cancer. *British Journal of Cancer*, 45:361, 1982.

[42] R. Hosseini, T. Ellis, M. Mazinani, and J. Dehmeshki. A genetic fuzzy approach for rule extraction for rule-based classification with application to medical diagnosis. In *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD); 05 - 09 Sep 2011, Athens, Greece*, 2011.

[43] P. R. Innocent, R. I. John, and J. Garibaldi. Fuzzy methods for medical diagnosis. *Applied Artificial Intelligence*, 19(1):69–98, 2004.

[44] A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: a review. *ACM Computing Surveys*, 31(3):264–323, September 1999.

[45] A.K. Jain and R.C. Dubes. *Algorithms for Clustering Data*. Prentice-Hall advanced reference series, Prentice-Hall. Englewood Cliffs, NJ, USA, 1988.

[46] R. Jain and A. Abraham. A comparative study of fuzzy classification methods on breast cancer data. In *7th International Work Conference on Artificial and Natural Neural Networks, IWANN03*, 2003.

[47] K. K. Jha, A. Samad, Y. Kumar, M. Shaharyar, R.L. Khosa, J. Jain, V. Kumar, and P. Singh. Design, synthesis and biological evaluation of 1,3,4-oxadiazole derivatives. *European Journal of Medicinal Chemistry*, 45(11):4963 – 4967, 2010.

[48] R. John and S. Coupland. Type-2 fuzzy logic: Challenges and misconceptions [discussion forum]. *IEEE Computational Intelligence Magazine*, 7(3):48–52, 2012.

[49] I.T. Jolliffe. *Principal Component Analysis*. Aberdeen,U.K, 2002.

[50] Y. Jusman, S.N. Sulaiman, N.A.M. Isa, I.A. Yusoff, R. Adnan, N.H. Othman, and A. Zaki. Capability of new features from ftir spectral of cervical cells for cervical precancerous diagnostic system using mlp networks. In *IEEE Region 10 Conference :TENCON 2009 - 2009*, pages 1–6, 2009.

[51] S. Kim, S. Min, J. Kim, S. Park, T. Kim, and J. Liu. Rapid discrimination of commercial strawberry cultivars using fourier transform infrared spectroscopy data combined by multivariate analysis. Technical report, Plant Biotechnology Reports, 2009.

[52] S.W. Kim, S.H. Ban, H. Chung, S. Cho, H.J. Chung, P.S. Choi, O.J. Yoo, and J.R. Liu. Taxonomic discrimination of flowering plants by multivariate analysis of fourier transform infrared spectroscopy data. Technical report, Plant Cell Reports, 2004.

[53] S. Kirkpatrick. Optimization by simulated annealing: Quantitative studies. *Journal of Statistical Physics*, 34(5-6):975–986, 1984.

[54] S.P.A. Kirubha and M. Anburajan. Spectrometric techniques for diagnosis of breast cancer. In *International Conference on Computer Communication and Informatics (ICCCI)*, pages 1–4, 2012.

[55] J. R. Kloss, T. H. Pedrozo, H. D. M. Follmann, P. Peralta-Zamora, J. A. Dionsio, L. Akcelrud, S.F. Zawadzki, and L.P. Ramos. Application of the principal component analysis method in the biodegradation polyurethanes evaluation. *Materials Science and Engineering*, 29(2):470 – 473, 2009.

[56] S.G. Kong, Y.R. Chen, I. Kim, and M.S. Kim. Analysis of hyperspectral fluorescence images for poultry skin tumor inspection. *Applied Optics*, 43(4):824–833, Feb 2004.

[57] C. Krafft, D. Codrich, G. Pelizzo, and V. Sergo. Raman mapping and ftir imaging of lung tissue: congenital cystic adenomatoid malformation. *The Analyst*, 133:361, 2008.

[58] S. Kumar, C. Desmedt, D. Larsimont, C. Sotiriou, and E. Goormaghtigh. Change in the microenvironment of breast cancer studied by ftir imaging. *Analyst*, 138:4058–4065, 2013.

[59] P. Lasch, W. Haensch, D. Naumann, and M. Diem. Imaging of colorectal adenocarcinoma using ft-ir microspectroscopy and cluster analysis. *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease*, 1688(2):176 – 186, 2004.

[60] S.H. Lee, W. Pedrycz, and S. Gyoyong. Design of similarity and dissimilarity measures for fuzzy sets on the basis of distance measure. *International Journal of Fuzzy Systems*, 11(2):67–72, 2011.

[61] F. Liu. An efficient centroid type-reduction strategy for general type-2 fuzzy logic system. *Information Sciences*, 178(9):2224 – 2236, 2008.

[62] H.W. Liu. New similarity measures between intuitionistic fuzzy sets and between elements. *Mathematical and Computer Modelling*, 42(12):61 – 70, 2005.

[63] J. Luts, J.B. Poullet, J.M. Garcia-Gomez, A. Heerschap, M. Robles, J. A. K. Suykens, and S.V. Huffel. Effect of feature extraction for brain tumor classification based on short echo time 1h mr spectra. *Magnetic Resonance in Medicine*, 60(2):288–298, 2008.

[64] E. Ly, O. Piot, A. Durlach, P. Bernard, and M. Manfait. Differential diagnosis of cutaneous carcinomas by infrared spectral micro-imaging combined with pattern recognition. *Analyst*, 134:1208, 2009.

[65] E. Ly, O. Piot, R. Wolthuis, A. Durlach, P. Bernard, and M. Manfait. Combination of ftir spectral imaging and chemometrics for tumour detection from paraffinembedded biopsies. *The Analyst*, 133:197, 2008.

[66] S. Z. Mahmoodabadi, J. Alirezaie, P. Babyn, A. Kassner, and E. Widjaja. Wavelets and fuzzy relational classifiers: A novel spectroscopy analysis system for pediatric metabolic brain diseases. *Fuzzy Sets and Systems*, 161(1):75 – 95, 2010.

[67] E. Manzano, N. Navas, R. Checa-Moreno, L.R. Simn, and L.F. Capitn-Vallvey. Preliminary study of uv ageing process of proteinaceous paint binder by ft-ir and principal component analysis. *Talanta*, 77:1724, 2009.

[68] J.B. Mbede, A. Melingui, B. Essimbi Zobo, R. Merzouki, and B.O. Bouamama. zslices based type-2 fuzzy motion control for autonomous robotino mobile robot. In *IEEE/ASME International Conference on Mechatronics and Embedded Systems and Applications (MESA), 2012*, pages 63–68, 2012.

[69] J. McCulloch, C. Wagner, and U. Aickelin. Expanding similarity measures of interval type-2 fuzzy sets to general type-2 fuzzy sets. In *IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, 2013.

[70] R. Mehrotra, D. K. Jangir, A. Gupta, and H. C. Kandpal. Differentiation of normal and malignant breast tissues using infrared spectroscopy. *AIP Conference Proceedings*, 1075(1):141–143, 2008.

[71] P. Melin and O. Castillo. A review on the applications of type-2 fuzzy logic in classification and pattern recognition. *Expert Systems with Applications*, 40(13):5413 – 5423, 2013.

[72] J. Mendel. *Uncertain Rule-Based Fuzzy Logic Systems: Introduction and New Directions*. Upper-Saddle River,NJ:Prentice Hall, 2001.

[73] J.M. Mendel, R.I. John, and Feilong Liu. Interval type-2 fuzzy logic systems made simple. *IEEE Transactions on Fuzzy Systems*, 14(6):808–821, Dec 2006.

[74] J.M. Mendel and R.I.B. John. Type-2 fuzzy sets made simple. *IEEE Transactions on Fuzzy Systems*, 10(2):117–127, 2002.

[75] S. Miller, C. Wagner, and J. Garibaldi. Modelling survey data with type-2 fuzzy sets. Intelligent Modelling and Analysis (IMA) Group Seminar, School of Computer Science, University of Nottingham, March 2012.

[76] S. Miller, C. Wagner, J.M. Garibaldi, and S. Appleby. Constructing general type-2 fuzzy sets from interval-valued data. In *IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), 2012*, pages 1–8, 2012.

[77] S.M. Miller. *Stochastic Search and Fuzzy Modelling for Real-World Complex Systems*. PhD thesis, De Montfort University, Leicester, 2011.

[78] H.B. Mitchell. Pattern recognition using type-ii fuzzy sets. *Information Sciences*, 170(24):409 – 418, 2005.

[79] T. Nagata, R. Schmelzeisen, D. Mattern, G. Schwarzer, and M. Ohishi. Application of fuzzy inference to european patients to predict cervical lymph node metastasis in carcinoma of the tongue. *International Journal of Oral and Maxillofacial Surgery*, 34(2):138 – 142, 2005.

[80] S. Naik, S. Doyle, S. Agner, A. Madabhushi, M. Feldman, and J. Tomaszewski. Automated gland and nuclei segmentation for grading of prostate and breast cancer histopathology. In *5th IEEE International Symposium on Biomedical Imaging: From Nano to Macro, 2008. ISBI 2008.*, pages 284 –287, may 2008.

[81] M. Negnevitsky. *Artificial intelligence: a guide to intelligent systems*. Pearson Education, 2005.

[82] L. Ohno-Machado. Modeling medical prognosis: Survival analysis techniques. *Journal of Biomedical Informatics*, 34(6):428 – 439, 2001.

[83] G. J. Ooi, J. Fox, K. Siu, R. Lewis, K. R. Bambery, D. McNaughton, and B. R. Wood. Fourier transform infrared imaging and small angle x-ray scattering as a combined biomolecular approach to diagnosis of breast cancer. *Medical Physics*, 35(5):2151–2161, 2008.

[84] T. Ozen and J.M. Garibaldi. Effect of type-2 fuzzy membership function shape on modelling variation in human decision making. In *IEEE International Conference on Fuzzy Systems (Fuzz IEEE 2004)*, volume 2, pages 971–976, 2004.

[85] J. D. Pallua, C. Pezzei, B. Zelger, G. Schaefer, L. K. Bittner, V. A. Huck-Pezzei, S. A. Schoenbichler, H. Hahn, A. Kloss-Brandstaetter, F. Kloss, G. K. Bonn, and C. W. Huck. Fourier transform infrared imaging analysis in discrimination studies of squamous cell carcinoma. *Analyst*, 137:3965–3974, 2012.

[86] C.A. Pea-Reyes and M. Sipper. A fuzzy-genetic approach to breast cancer diagnosis. *Artificial Intelligence in Medicine*, 17(2):131 – 155, 1999.

[87] C. Pena-Reyes. Evolutionary fuzzy modeling human diagnostic decisions. *Annals of the New York Academy of Sciences*, 1020(1):190–211, 2004.

[88] R. Perez-Pueyo, M. J. Soneira, and S. Ruiz-Moreno. A fuzzy logic system for band detection in raman spectroscopy. *Journal of Raman Spectroscopy*, 35(8-9):808–812, 2004.

[89] S. Petushi, F. Garcia, M. Haber, C. Katsinis, and A. Tozeren. Large-scale computations on histology images reveal grade-differentiating parameters for breast cancer. *BMC Medical Imaging*, 6(1):14, 2006.

[90] P.A. Phong and K. Q. Thien. Classification of cardiac arrhythmias using interval type-2 tsk fuzzy system. In *International Conference on Knowledge and Systems Engineering, 2009. KSE '09.*, pages 1–6, 2009.

[91] R. Radhakrishnan, M. Solomon, K. Satyamoorthy, L. E. Martin, and M.W. Lingen. Tissue microarray a high-throughput molecular analysis in head and neck cancer. *Journal of Oral Pathology & Medicine*, 37(3):166–176, 2008.

[92] E.A. Rakha, M.E. El-Sayed, A.H.S. Lee, C.W. Elston, M.J. Grainge, Z. Hodi, R.W. Blamey, and I.O. Ellis. Prognostic significance of nottingham histologic grade in invasive breast carcinoma. *Journal of Clincal Oncology*, 26:3153, 2008.

[93] R.S. Rampaul, S.E. Pinder, C.W. Elston, and I.O. Ellis. Prognostic and predictive factors in primary breast cancer and their role in patient management: The nottingham breast team. *European Journal of Surgical Oncology*, 27:229, 2001.

[94] S. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach*. Prentice Hall, 2012.

[95] F.M. Schleif, T. Riemer, U. Brner, L. Schnapka-Hille, and M. Cross. Genetic algorithm for shift-uncertainty correction in 1-d nmr-based metabolite identifications and quantifications. *Bioinformatics*, 27(4):524–533, 2011.

[96] W. Schumacher, M. Khnert, P. Rsch, and J. Popp. Identification and classification of organic and inorganic components of particulate matter via raman spectroscopy and chemometric approaches. *Journal of Raman Spectroscopy*, 42(3):383–392, 2011.

[97] S. Slobodan. Determining the coating thickness of tablets by chiseling and image analysis. *International Journal of Pharmaceutics*, 397(12):109 – 115, 2010.

[98] American Cancer Society. Cancer facts and figures. Technical report, 2012.

[99] I. Soerjomataram, M.W.J. Louwman, J.G. Ribot, J. Roukema, and J.W. Coebergh. An overview of prognostic factors for long-term survivors of breast cancer. *Breast Cancer Research and Treatment*, 107:309–330, 2008.

[100] G. Steiner, S. Tunc, M. Maitz, and R. Salzer. Conformational changes during protein adsorption. ft-ir spectroscopic imaging of adsorbed fibrinogen layers. *Analytical Chemistry*, 79(4):1311–1316, 2007.

[101] W. Steller, J. Einenkel, L.C. Horn, U.D. Braumann, H. Binder, R. Salzer, and C. Krafft. Delimitation of squamous cell cervical carcinoma using infrared microspectroscopic imaging. *Analytical and Bioanalytical Chemistry*, 384:145–154, 2006.

[102] K. Thumanu, W. Tanthanuch, C. Lorthongpanich, P. Heraud, and R. Parnpai. {FTIR} microspectroscopic imaging as a new tool to distinguish chemical composition of mouse blastocyst. *Journal of Molecular Structure*, 933(13):104 – 111, 2009.

[103] J. Trevisan, P. P. Angelov, P. L. Carmichael, A. D. Scott, and F. L. Martin. Extracting biological information with computational analysis of fourier-transform infrared (ftir) biospectroscopy datasets: current practices to future perspectives. *Analyst*, 137:3202–3215, 2012.

[104] A.E. Tutac, D. Racoceanu, T. Putti, Wei Xiong, Wee-Kheng Leow, and V. Cretu. Knowledge-guided semantic indexing of breast cancer histopathology images. In *International Conference on BioMedical Engineering and Informatics, 2008. BMEI 2008*, volume 2, pages 107 –112, may 2008.

[105] Cancer Research UK. Cancer research report. Technical report, 2012.

[106] P. Venkatachalam, L. L. Rao, N. K. Kumar, A. Jose, and S. S. Nazeer. Diagnosis of breast cancer based on ft-ir spectroscopy. *AIP Conference Proceedings*, 1075(1):144–148, 2008.

[107] C. Wagner and H. Hagras. Toward general type-2 fuzzy logic systems based on zslices. *IEEE Transactions on Fuzzy Systems*, 18(4):637–660, 2010.

[108] C. Wagner, S. Miller, and J.M. Garibaldi. Similarity based applications for data-driven concept and word models based on type-1 and type-2 fuzzy sets. In *Proceedings of IEEE Conference on Fuzzy Systems (FUZZ IEEE)*, 2013.

[109] X.Y. Wang. *Fuzzy Clustering in the Analysis of Fourier Transform Infrared Spectra for Cancer Diagnosis*. PhD thesis, School of Computer Science, University of Nottingham, 2006.

[110] X.Y. Wang and J.M. Garibaldi. Comparison of fuzzy and non-fuzzy clustering techniques in cancer diagnosis. In *Second international conference in Computational Intelligence in Medicine and Healthcare (The Biopattern Conference). Lisbon,Portugal*, 2005.

[111] X.Y. Wang and J.M. Garibaldi. Simulated annealing fuzzy clustering in cancer diagnosis. *Informatica*, 29:61–70, 2005.

[112] X.Y. Wang, J.M. Garibaldi, B. Bird, and M.W. George. Anovel fuzzy clustering algorithm for the analysis of axillary lymph node tissue sections. *Applied Intelligence*, 27:237–248, 2007.

[113] Y. Wang and Z. Yu. A type-2 fuzzy method for identification of disease-related genes on microarrays. *International Journal of Bioscience, Biochemistry and Bioinformatics (IJBBB)*, 1(1):73–78, 2011.

[114] S. Wartewing. *IR and Raman Spectroscopy: Fundamental Processing*. 2003.

[115] Siegfried Weber. A general concept of fuzzy connectives, negations and implications based on t-norms and t-conorms. *Fuzzy Sets and Systems*, 11(13):103 – 113, 1983.

[116] B.R. Wood, L. Chiriboga, H. Yee, M.A. Quinn, D. McNaughton, and M. Diem. Fourier transform infrared (ftir) spectral mapping of the cervical transformation zone, and dysplastic squamous epithelium. *Gynecologic Oncology*, 93(1):59 – 68, 2004.

[117] D. Wu. On the fundamental differences between interval type-2 and type-1 fuzzy logic controllers. *IEEE Transactions on Fuzzy Systems*, 20(5):832–848, 2012.

[118] D. Wu and J. M. Mendel. A comparative study of ranking methods, similarity measures and uncertainty measures for interval type-2 fuzzy sets. *Information Sciences*, 179(8):1169 – 1192, 2009.

[119] Ronald R. Yager. On a general class of fuzzy connectives. *Fuzzy Sets and Systems*, 4(3):235 – 242, 1980.

[120] M.S. Yang and D.C. Lin. On similarity and inclusion measures between type-2 fuzzy sets with an application to clustering. *Computers & Mathematics with Applications*, 57(6):896 – 907, 2009.

[121] Z. Ye. Artificial-intelligence approach for biomedical sample characterization using raman spectroscopy. *IEEE Transactions on Automation Science and Engineering*, 2(1):67 – 73, jan. 2005.

[122] L. A. Zadeh. Fuzzy sets. *Information and control*, 8(3):338–353, 1965.

[123] L.A. Zadeh. The concept of a linguistic variable and its application to approximate reasoningi. *Information Sciences*, 8(3):199 – 249, 1975.

[124] M.H. F. Zarandi, M. Zarinbal, and M. Izadi. Systematic image processing for diagnosing brain tumors: A type-ii fuzzy expert system approach. *Applied Soft Computing*, 11(1):285 – 294, 2011.

[125] W. Zeng and H. Li. Relationship between similarity measure and entropy of interval valued fuzzy sets. *Fuzzy Sets and Systems*, 157(11):1477 – 1484, 2006.

[126] Y. Zhao and G. Karypis. Soft clustering criterion functions for partitional document clustering: a summary of results. In *Proceedings of the thirteenth ACM international conference on Information and knowledge management*, CIKM '04, pages 246–247, New York, NY, USA, 2004. ACM.

[127] K. Zühtüoğullari, I. Saritaş, and N. Arikan. Diagnosis modelling of urethral obstructions using fuzzy expert system. In *Proceedings of the 9th International Conference on Computer Systems and Technologies and Workshop for PhD Students in Computing*, CompSysTech '08, pages 34:IIIA.14–34:1, New York, NY, USA, 2008. ACM.

[128] R. Zwick, E. Carlstein, and D.V. Budescu. Measures of similarity among fuzzy concepts: A comparative analysis. *International Journal of Approximate Reasoning*, 1(2):221 – 242, 1987.

[129] A. Zwielly, J. Gopas, G. Brkic, and S. Mordechai. Discrimination between drugresistant and non-resistant human melanoma cell lines by ftir spectroscopy. *The Analyst*, 134:294, 2009.