

University of Warwick institutional repository: <http://go.warwick.ac.uk/wrap>

A Thesis Submitted for the Degree of PhD at the University of Warwick

<http://go.warwick.ac.uk/wrap/66677>

This thesis is made available online and is protected by original copyright.

Please scroll down to view the document itself.

Please refer to the repository record for this item for information to help you to cite it. Our policy information is available from the repository home page.



Intelligent Data Analysis for
Pattern Recognition and Medical
Diagnosis of Ageing Spine

By

Atif Ali Khan

A dissertation submitted in fulfilment of the requirements for
the degree of Doctor of Philosophy

University of Warwick, School of Engineering

June 2014

TABLE OF CONTENTS

List of Figures	V
List of Tables.....	VIII
Acknowledgements.....	IX
Declaration	X
List of Publications	XI
Abstract	XIII
Abbreviations	XV
1 Introduction	1
1.1 Introduction.....	2
1.1.1 Human Spine and Natural Ageing	2
1.1.2 Research Motivation	4
1.2 Data Mining and Knowledge Discovery.....	4
1.2.1 Data Mining Tasks and Procedures.....	6
1.2.2 Machine Learning and Pattern Recognition	8
1.2.3 Common Data Mining Techniques.....	9
1.3 Data Mining in Biomedicine	10
1.3.1 Challenges and Scope.....	12
1.3.2 Intelligent Data Analysis.....	12
1.4 Research Objectives and Contribution to the Knowledge	14
1.5 Thesis Outline	16
References.....	18
2 Intelligent Systems Techniques	23
2.1 Introduction to Intelligent Systems.....	24
2.2 Artificial Neural Networks	25
2.2.1 Types of Artificial Neural Networks	27
2.2.2 Supervised, Unsupervised, and Reinforced Learning	29
2.3 Feed-Forward Neural Networks.....	30
2.4 Self-Organising Maps.....	32
2.4.1 Structure of Self-Organising Maps.....	33

2.4.2	SOM Training and Learning.....	34
2.5	Cluster Analysis.....	36
2.6	Fuzzy Logic.....	38
2.6.1	Mamdani Type Fuzzy Inference.....	39
2.6.2	Sugeno Type Fuzzy Inference.....	40
2.7	Hybrid Intelligent Systems.....	40
2.7.1	Adaptive Neuro-Fuzzy Inference System.....	41
2.8	Summary.....	43
	References.....	44
3	The Human Spine Anatomy, Problems and Diagnosis.....	47
3.1	Human Spine.....	48
3.1.1	Anatomy of Human Spine.....	50
3.1.2	Age Related Degenerative Changes.....	52
3.1.3	Some Common Pathologies of the Spine.....	53
3.2	Back Pain Problem.....	55
3.2.1	Ageing Population and Back Pain.....	56
3.2.2	Key Facts and Causes of Back Pain.....	56
3.2.3	Economic and Social Cost.....	59
3.3	Diagnostic Imaging.....	60
3.3.1	Lumbar Spine Magnetic Resonance Imaging.....	61
3.3.2	T1 and T2 Weighted Images.....	62
3.3.3	Studies on Age Related Changes in Human Spine with the Help of MRI Findings	63
3.4	Data Set.....	64
3.4.1	Data Acquisition and Pre-Processing.....	65
3.4.2	Feature Extraction and Measurement.....	66
3.4.3	Scoring of Lumbar Spine MRI.....	66
3.5	Summary.....	68
	References.....	69
4	Feature Extraction and Visualization of Multivariate Lumbar Spine Data.....	78
4.1	Principal Component Analysis.....	79
4.2	PCA Modelling.....	81

4.2.1	Statistical Analysis and Data Standardization	84
4.2.2	Data Visualization.....	87
4.2.3	Correlation of Features with Age	91
4.2.4	Non Correlated Features.....	94
4.3	Factor Analysis.....	95
4.4	Hierarchical Clustering	98
4.4.1	Agglomerative Clustering.....	99
4.4.2	Cluster Analysis	100
4.5	PCA Based Anomaly Detection.....	107
4.6	Summary.....	110
	References.....	111
5	Estimation of Age and Gender from Lumbar Spine Features.....	114
5.1	Machine Learning in Medicine	115
5.2	Artificial Neural Networks (ANNs).....	116
5.2.1	Neural Network Topologies	120
5.2.2	Training Methods for a Neural Network.....	120
5.2.3	Back-Propagation	121
5.2.4	Applications of Neural Networks	124
5.2.5	Back-Propagation with Levenberg-Marquardt Algorithm	125
5.3	Results of Neural Network Modelling	129
5.3.1	Spinal Age Estimation.....	132
5.3.2	Gender Estimation from Spinal Features.....	138
5.3.3	Both Age and Gender Estimation.....	141
5.4	Cross-Validation.....	143
5.4.1	Cross-Validation Techniques.....	144
5.4.2	Repeated K-fold Cross-Validation	145
5.5	Principal Component Neural Network Model.....	152
5.6	Fuzzy Inference System	155
5.6.1	Mamdani Fuzzy Inference System	156
5.6.2	Sugeno Fuzzy Inference System.....	157
5.7	Hybrid Model.....	160
5.7.1	Adaptive Neuro-Fuzzy Inference System (ANFIS)	160

5.7.2	ANFIS with Subtractive Clustering	164
5.8	Summary.....	167
	References.....	168
6	Unsupervised Pattern Recognition in Ageing Spine	172
6.1	Unsupervised Learning.....	173
6.2	Self-Organizing Maps (SOMs).....	174
6.3	Self-Organizing Maps (SOM) Modelling and Analysis	177
6.3.1	SOM Modelling.....	177
6.3.2	Visual Cluster Analysis.....	187
6.4	Clustering Analysis for Ageing Pattern of Spine	188
6.4.1	Ward's Clustering Methods	189
6.4.2	Modified SOM-Ward Clustering.....	190
6.4.3	Characteristics of the Clusters	192
6.5	Summary.....	203
	References.....	204
7	Conclusion and future work	206
7.1	Overview.....	207
7.2	Results from Principal Component and Factor Analysis	208
7.3	Comparison of Results with Statistical Analysis	211
7.4	Results from Neural Network Modelling	213
7.5	Results from SOM and Clustering	214
7.6	Final Prototype of the System	217
7.7	Future Research Directions	220

LIST OF FIGURES

Figure 1.1: Illustration of knowledge discovery steps	5
Figure 2.1: A simple artificial neuron design.....	26
Figure 2.2: Architecture of common ANNs (a) Multilayer feedforward neural network; (b) Self-organizing map; (c) Multilayer recurrent neural network; (d) cellular neural network	28
Figure 2.3: Multi-layer feedforward artificial neural network.....	30
Figure 2.4: Self-organizing map (SOM) structure	33
Figure 2.5: Training of self-organizing map (SOM)	34
Figure 2.6: Architecture of Mamdani fuzzy Inference systems	40
Figure 2.7: Adaptive neuro-fuzzy inference system architecture.....	42
Figure 3.1: Human vertebral column	50
Figure 3.2: Anatomy of lumbar spine.....	51
Figure 3.3: Number of back pain patients in Scotland consulting a GP or practice nurse at least once in the financial year 2012-13 per 1,000 (Statistics by NHS Scotland)	58
Figure 3.4: Sagittal and an axial view of the lumbar spine MRI.....	61
Figure 3.5: T1 and T2 weighted lumbar MRI.....	62
Figure 3.6: Distribution of data samples	65
Figure 4.1: Step by step demonstration of PCA modelling.....	81
Figure 4.2: (a) Variance shown by first five components of the data, (b) Plot of 1st principal component vs. 2nd principal component	87
Figure 4.3: PCA plot of first two components with age labelled	89
Figure 4.4: PCA plot of first two components with age and gender labelled	90
Figure 4.5: Plot of PC 1 vs. PC3.....	90
Figure 4.6: Plot of PC2 vs. PC3	91
Figure 4.7: Plot of samples and the variables	92
Figure 4.8: PCA plot of samples with variables.....	94
Figure 4.9: PC1 vs PC2 scatter plot	100
Figure 4.10: PC1 vs. PC 2 with 10 clusters.....	101
Figure 4.11: Clustering with (a) 5 clusters (b) 6 clusters.....	102
Figure 4.12: Clustering with (a) 7 clusters (b) 8 clusters.....	104
Figure 4.13 (a): Dendrogram with sample number on x-axis and dissimilarity measure on y-axis.....	106
Figure 4.13 (b): Clustering with 10 clusters	107
Figure 4.14: PC1 vs. PC2 with 2 clusters	108
Figure 4.15: Clustering with 9 clusters.....	110
Figure 5.1: A simple neural network.....	117

Figure 5.2: Some nonlinear neuron activation functions	118
Figure 5.3: Multilayer perceptron model.....	119
Figure 5.4: Multilayer perceptron and weights free parameters to be adapted	122
Figure 5.5: Random distribution of samples for neural network modelling	132
Figure 5.6: (a) Mean square error and (b) Neural network training state.....	133
Figure 5.7: Regression of first neural network model	134
Figure 5.8: Error histogram of neural network estimation.....	135
Figure 5.9: (a) Training state, (b) Performanceof neural network.....	135
Figure 5.10: Regression of second neural network model.	136
Figure 5.11: (a) Error histogram, (b) Actual vs. NN estimated spinal age.	137
Figure 5.12: (a) Performance (b) Training state of gender estimation model....	138
Figure 5.13: Regression of gender predicting neural network	140
Figure 5.14: Error histogram for gender prediction neural network.....	141
Figure 5.15: (a) Performance (b) Training state of age-gender neural network	141
Figure 5.16: Regression of age-gender predicting neural network	142
Figure 5.17: Error histogram for age-gender predicting neural network.....	143
Figure 5.18: K-fold cross-validation	144
Figure 5.19: 1st iteration of 10-fold cross validation.....	147
Figure 5.20: (a) Training state and (b) Performance of PCNN	153
Figure 5.21: Regression of principal component neural network	154
Figure 5.22: Error histogram for principal component neural network.....	155
Figure 5.23: Mamdani type fuzzy inference system for spinal age estimation..	157
Figure 5.24: Sugeno type fuzzy Inference system	158
Figure 5.25: Membership functions for Sugeno type FIS.	158
Figure 5.26: Surface view of Sugeno type FIS inputs	159
Figure 5.27: ANFIS model structure	161
Figure 5.28: ANFIS training with 41 Samples	162
Figure 5.29: (a) Testing of ANFIS (b) Checking of ANFIS	163
Figure 5.30: (a) Training of new FIS.	165
Figure 5.30: (b) Validation, (c) Testing of new FIS.	165
Figure 5.31: Surface view of inputs-outputs for new FIS.....	166
Figure 6.1: SOM neighbourhood weight distance	179
Figure 6.2: SOM (a) neighbour connections and (b) sample hits	180
Figure 6.3: SOM input weight plane for non-standardized data with map size 100	182
Figure 6.4: Weight planes for standardized data with map size 100	183
Figure 6.5: SOM U-matrix representations with non-standardized data	185
Figure 6.6: SOM U-matrix representations with standardized.....	186
Figure 6.7: SOM based on 24 input features along with age and gender	187
Figure 6.8: Ward'sclustering with standardized data	193
Figure 6.9: SOM-Ward clustering with standardized data.....	194

Figure 6.10: Clusters formed by using (a) Ward (b) SOM Ward method	195
Figure 6.11: SOM-Ward clustering on the basis of gender.....	197
Figure 6.12: Clustering on the basis of gender (a) SOM-Ward Method (b) Ward's Method.....	198
Figure 6.13: (a) Group profile of cluster C1 with SOM-Ward gender clustering.	199
Figure 6.13: (b) Group profile of cluster C2 with SOM-Ward gender clustering	199
Figure 6.13: (c) Group profile of cluster C3 with SOM-Ward gender clustering.	200
Figure 6.14: (a) Clustering (single color) on the basis of age by considering all attributes.....	201
Figure 6.14: (b) Clustering (multi-color) on the basis of age by considering all attributes.....	201
Figure 7.1: Results of principal component analysis.....	209
Figure 7.2: Anomaly detection with PCA	209
Figure 7.3: Results from factor analysis	211
Figure 7.4: Results from SOM-Ward clustering with standardized data	215
Figure 7.5: Changes in vertebral and disc heights among different age groups.	216
Figure 7.6: Changes in lumbar spine features among different groups.	216
Figure 7.7: Final prototype of the system.....	219

LIST OF TABLES

Table 2.1: List of intelligent techniques used in this thesis	25
Table 2.2: Different activation functions with their formulas	31
Table 3.1 Five randomly selected samples with their lumbar spine MRI scores..	67
Table 4.1: Illustration of extracted features of lumbar spine.....	82
Table 4.2: Statistics of original data set	84
Table 4.3: Statistics of standardized data set	86
Table 4.4: Loadings of first three factors	96
Table 4.5: Comparison of anomaly and average values of its respective age decade	109
Table 5.1: Results from first iterations of 10-fold cross validation.....	146
Table 5.2: Results from second iterations of 10-fold cross validation.....	147
Table 5.3: Results from third iterations of 10-fold cross validation.....	148
Table 5.4: Results from fourth iterations of 10-fold cross validation.....	149
Table 5.5: Results from fifth iterations of 10-fold cross validation.....	150
Table 5.6: Results from all 5 iterations of 10-fold cross validation.....	151
Table 6.1: Characteristics of Ward and SOM-Ward Clusters	195
Table 7.1: Correlation of spinal features with age.....	212

ACKNOWLEDGEMENTS

First and foremost I want to thank my supervisors Dr. Akeel Shah and Dr. Daciana Iliescu, for their valuable guidance and excellent support during the entire period of my research.

My sincere gratitude is reserved for Prof. Charles Hutchinson (Warwick Medical School), Prof. Evor Hines (School of Engineering), and Dr. Robert Sneath (University Hospital Coventry and Warwickshire) for their invaluable insights, kind assistance and extremely helpful suggestions.

I gratefully acknowledge the funding sources that made my PhD work possible. The work was partially funded by the Research Development Fund (RDF), University of Warwick, United Kingdom.

Thanks also go to my colleagues at Intelligent Systems Lab, School of Engineering, Dr. Adnan el Berjaoui Yakzan, Dr. Oumair Naseer, and Dr. Khair ul zaman Kadir for their fruitful discussions and support.

Last but not least, my deepest gratitude goes to my beloved parents, my wife and also to my brother and sister for their endless love and encouragement.

Atif Ali Khan

University of Warwick

June 2014

DECLARATION

This thesis is presented in accordance with the regulations for the degree of doctor of philosophy. The work described in this thesis is entirely original and my own, except where otherwise indicated.

Atif Ali Khan

June 2014

LIST OF PUBLICATIONS

Peer Reviewed Conference Papers

Khan, A., Iliescu, D., Hines, E., Hutchinson, C., Sneath, R., "Neural Network Based Spinal Age Estimation Using Lumbar Spine Magnetic Resonance Images (MRI)," Proceedings of the 4th IEEE International Conference on Intelligent Systems Modelling & Simulation (ISMS 2013), Bangkok, Thailand, 29-31 January 2013, pp. 88-93, DOI:10.1109/ISMS.2013.101.

Khan, A.A., Hines, E.L., Iliescu, D.D., "Using Self Organizing Maps to Visualize Age Related Changes in Lumbar Vertebrae and Intervertebral Discs", In the proceedings of 24th Midwest Artificial Intelligence and Cognitive Science Conference (MAICS 2013), April 13-14, 2013, Indiana University Southeast, New Albany, IN, USA.

Khan, A. A., Iliescu, D.D., Hines, E.L., Hutchinson, C.E., Sneath, R.J., "Classification of Age-Related Changes in Lumbar Spine with the help of MRI Scores," Proceedings of the 29th IEEE-EMB Southern Biomedical Engineering Conference (SBEC 2013), Miami, Florida, USA, 3-5 May, 2013, pp. 121-122, DOI:10.1109/SBEC.2013.69.

Peer Reviewed Journal Papers

Khan, A. A., Iliescu, D.D., Shah, A., Hutchinson, C.E., Sneath, R.J., "Principal Component and Factor Analysis to Study Age Related Changes in Lumbar Spine," in IEEE Journal of Biomedical and Health Informatics (J-BHI), Special Issue: Bioinformatics in Clinical Environments, June 2014.

Khan, A. A., Iliescu, D.D., Shah, A., Hutchinson, C.E., Sneath, R.J., "Spinal Age Estimation from Lumbar Spine MRI Scores by Using Principal Component Neural Network," submitted to International Journal of Intelligent Systems Technologies and Applications, Special Issue on Advances in Data Mining and Machine Learning. (Under Review).

Khan, A. A., Iliescu, D.D., Shah, A., Hutchinson, C.E., Sneath, R.J., "Study on Age Related Changes in Lumbar Spine with the Help of MRI Scores," submitted to IEEE Transactions on Medical Imaging (T-MI), Special Issue on Spine Imaging, Image-based Modelling, and Image Guided Intervention. (Under revision).

Poster Publication and Presentations:

Research Presentation on “Artificial Intelligence in Medicine with Special Focus on Lumbar Spine Aging”, in Integrated Clinical Academic Training (ICAT) Program, 13th Feb. 2013 at Clinical Science Research Labs (CSRL), University Hospital Coventry and Warwickshire, Coventry, UK.

Poster Presentation: “Analysis of Human Spine Using Intelligent System Techniques”, in Health Research at Warwick (Demonstrating Impact and Vitality) Event, 29th January 2013.

Research Presentation: “Intelligent Data Analysis for Age Related Changes in Human Lumbar Spine,” at Annual Postgraduate Symposium, School of Engineering, University of Warwick, 18th March 2013.

Poster Presentation: “Intelligent Data Analysis for Aging spine,” at Warwick Engineering in Biomedicine Open Day Poster Competition, 31st October 2012.

Poster Paper: “Visualization of Multivariate Lumbar Spine MRI Data Using Principal Component Analysis (PCA),” in 35th Annual International IEEE-EMBS Conference of the IEEE Engineering in Medicine and Biology, Osaka, Japan, July 3-7, 2013.

ABSTRACT

Every year, the healthcare industry collects a huge volume of data that is not mined properly and not put to optimal use. Discovery of the hidden patterns and relationships in data often goes unexploited. Data mining in the medical domain is more rigorous and complex to handle as most available raw medical data are voluminous and heterogeneous in nature. This research mines medical data related to human spine by learning patterns through the collected data and develops medical decision support systems based on intelligent system techniques. This study will help medical specialists in clinical decision making and disease diagnosis related to the spine.

The human spine is a multifunctional complicated structure of bones, joints, ligaments and muscles which all undergo change as we age. For most people, these changes occur in a gradual and painless manner. However, a sudden change caused naturally or through injury, can lead to serious medical conditions, which usually result in back pain. Due to the wide diversity of spine functions, any disorder in the spine triggers various severe problems, which negatively affect quality of life and place huge financial and health burdens on the society. While ageing is inevitable, the rate at which the spine shows the effects of ageing is of clinical significance. This research reveals the growth and degenerative pattern of the human spine using intelligent system techniques. The information extracted from lumbar spine MRIs is used to classify age related changes.

In this research, principal component analysis was used to detect anomalies in data and to transform the complex multivariate feature space to a smaller meaningful

representation. PCA transformation reduced the complexity and dimension of the data, hence permitting a 2D visualization and knowledge of spine growth and degeneration with age. Factor analysis (FA) was used to understand the significance and correlation of spinal features with the normal ageing. Spines were ranked on the basis of their features and clusters were made to group similar samples. Studying the characteristics of the clusters helped in developing an understanding of the variations in spinal features among different age groups. An artificial neural network (ANN) was used in the estimation of age from the extracted lumbar spine features. ANNs have several benefits, including their ability to process complex data, reduced likelihood of overlooking relevant information, and a reduction in the cost and diagnosis time. The ANN model worked well for the spinal age estimation but due to its black box nature, it failed to provide valuable information about the correlations among the spinal features. A hybrid intelligent model consisting of a fuzzy inference system (FIS) and ANN, called an adaptive neuro-fuzzy inference system (ANFIS) was used to extract meaningful information from the data set in terms of fuzzy rules. Self-organizing maps (SOM) were used to visualize variations in lumbar spine features with the natural ageing. Useful information was acquired through SOM exploratory data analysis. Ward and modified Ward clustering methods were employed on SOM to group similar samples and study the characteristics of the clusters. The results from this research are helpful in setting the standards for spinal growth and degeneration with age and for understanding of the spinal disease prevalence. This research will help spine specialists in diagnosing disease from scans. It can be considered as a stepping-stone towards developing a tool for the classification of normal and problematic spines.

ABBREVIATIONS

AAA	Abdominal Aortic Aneurysm
AI	Artificial Intelligence
ANFIS	Adaptive Neuro-Fuzzy Inference System
ANN	Artificial Neural Network
ARC	Age Related Changes
ART	Adaptive Resonance Theory
BP	Back-Propagation
CAD	Computer Aided Diagnosis
CSF	Cerebrospinal Fluid
DDD	Degenerative Disc Disease
DH	Disc Height
DICOM	Digital Imaging and Communications in Medicine
DM	Data Mining
DSI	Disc Signal Intensity
FA	Factor Analysis
FIS	Fuzzy Inference System
FL	Fuzzy Logic
FFNN	Feed-Forward Neural Network
FUN	Fuzzy Net
IS	Intelligent System
IST	Intelligent System Techniques
KNN	K-Nearest Neighbour
L1	Lumbar Vertebra Number 1

L2	Lumbar Vertebra Number 2
L3	Lumbar Vertebra Number 3
L4	Lumbar Vertebra Number 4
L5	Lumbar Vertebra Number 5
L1-L2	Intervertebral Disc between 1 st and 2 nd Lumbar Vertebrae
L2-L3	Intervertebral Disc between 2 nd and 3 rd Lumbar Vertebrae
L3-L4	Intervertebral Disc between 3 rd and 4 th Lumbar Vertebrae
L4-L5	Intervertebral Disc between 4 th and 5 th Lumbar Vertebrae
L5-S1	Intervertebral Disc between 5 th Lumbar and 1 st Sacrum Vertebrae
LMA	Levenberg–Marquardt Algorithm
LVQ	Learning Vector Quantization
MF	Membership Function
MLP	Multi-Layer Perceptron
MRI	Magnetic Resonance Imaging
MSE	Mean Square Error
NN	Neural Network
PACS	Picture Archiving and Communications System
PC1	Principal Component 1
PC2	Principal Component 2
PC3	Principal Component 3
PCA	Principal Components Analysis
PCNN	Principal Component Neural Network
PR	Pattern Recognition

PSM	Para-Spinal Muscle
R	Regression
RBF	Radial Basis Function
RL	Reinforcement Learning
RMSE	Root Mean Square Error
RNN	Recurrent Natural Networks
SA	Simulated Annealing
SI	Signal Intensity
SOM	Self-Organizing Map
SOM-WARD	Self-Organizing Map Based Ward Clustering
SVM	Support Vector Machine
T12-L1	Intervertebral Disc between 12 th Thoracic Vertebra and 1 st Lumbar Vertebra
U-Matrix	Unified Distance Matrix
VH	Vertebral Height

1

INTRODUCTION

1.1 INTRODUCTION

1.1.1 HUMAN SPINE AND NATURAL AGEING

1.1.2 RESEARCH MOTIVATION

1.2 DATA MINING AND KNOWLEDGE DISCOVERY

1.2.1 DATA MINING TASKS AND PROCEDURES

1.2.2 MACHINE LEARNING AND PATTERN RECOGNITION

1.2.3 COMMON DATA MINING TECHNIQUES

1.3 DATA MINING IN BIOMEDICINE

1.3.1 CHALLENGES AND SCOPE

1.3.2 INTELLIGENT DATA ANALYSIS

1.4 RESEARCH OBJECTIVES AND CONTRIBUTION TO THE

KNOWLEDGE

1.5 THESIS OUTLINE

REFERENCES

1.1 INTRODUCTION

Ageing in humans is a complex process of physical, psychological and social change over time. Some features grow and expand, while others decline with ageing. The ageing of the population in developed countries appears to be a non-reversible phenomenon. In developed countries, the decreased birth rates and increased life expectancies have led to a rising median age. This increasing median age not only has significant social and economic implications but also places a tremendous load on healthcare services. Today, the major challenge faced by the healthcare industry is the provision of quality services at affordable costs. A quality service implies accurate diagnosis and effective treatment as poor clinical decisions can lead to devastating outcomes. It is estimated that 1 out of 20 adults in USA could be misdiagnosed during outpatient visits, and about half of those errors could prove to be harmful for the patient. Misdiagnosis could affect 12 million US adults annually [1]. Misdiagnosis accounts for an estimated 40,000 to 80,000 hospital deaths per year in US [2], [3]. A recent study examined malpractice claims of over 25 years in US, identifying more than 100,000 cases that involved diagnostic error. Malpractice claim pay-outs for diagnostic errors amounted to \$38.8 billion USD between 1986 and 2010 in USA with an average price of \$386,849 (USD) per claim [4].

1.1.1 HUMAN SPINE AND NATURAL AGEING

With an ageing population, an increased number of musculoskeletal problems are seen. Back and neck pain are among the most frequently encountered complaints from older people. Back pain is usually associated with spine problems. The complex nature of the spine makes those problems difficult to diagnose and to

treat. The natural changes associated with ageing gradually affect almost all structures of the spinal units. Degeneration of the spinal structures makes alterations at several levels such as the bone, disc, joints, muscles, and ligaments.

These alterations in structures of the spine could be due to natural ageing or a spinal disease. To date, there are no standard reference values available that can be compared to conclude whether the changes are induced by natural ageing or by disease. The previous studies conducted to set the reference values for ageing spine produced contradicting results. This is why medical community was not able to define the lumbar spine characteristics with ageing. Some studies have attempted to investigate the effect of ageing on specific areas of the spine but the conclusions reached vary from study to study. For example, some studies suggest a vertebral height increase with the age [5], while others report a decrease in vertebral height with age [6]. Interestingly, another study suggests that vertebral height increases with age in males but does not change for females [7]. The variations among the studies are probably due to the methods used to analyse the data. Most of the studies have tried to correlate one feature at a time. Very few have tried to correlate two or more spinal features with the ageing process. The spine is a complex structure having several bones, ligaments, joints and, muscles, which often affect one another and move together. This research is based on the evaluation of 24 spinal features altogether with the natural ageing process. This work not only portrays the behaviour of an individual feature of a spine but also explores the correlations among the features.

1.1.2 RESEARCH MOTIVATION

The use of magnetic resonance images (MRI) for diagnosing back pain and other spine problem has become a standard practice. With the help of MRI, the anatomical reason for back pain and spinal pathologies can be identified. In the absence of a clear pathology, back pain is generally thought to originate from muscles. However, patients remain frustrated about the cause of their back pain when there is no pathological indication from the MRI. This pain may originate from the degenerative changes in the spine with natural ageing. Spine specialists do not, however, have any reference values with which to compare and, therefore, are often unable to reassure patients that the results of their MRI scans are normal and that the ageing effects are within normal limits.

Compared to other physical illness or injuries, the Social Security services receive a large number of applications for disability benefits based on back problems, which places a huge burden on society. Inevitably, some of these claims are bogus. If the normal limits of spinal features with natural ageing are known, such bogus application can be detected by comparing the outcome of MRI scans.

1.2 DATA MINING AND KNOWLEDGE

DISCOVERY

The process of scrutinising and analysing the dataset from different perspectives so that conclusive information can be extracted and summarized is commonly known as 'Data Mining'. The procedure is normally used for mining previously unknown, valid, potentially useful and ultimately understandable information from a dataset

[8]. Data can be in the form of facts, numbers, or text that can be processed by a computer. It allows users to analyse data from different perspectives, to categorize it, and to extract relationships between features in the data. The existing patterns, associations, and relationships among data can provide useful information that can be turned into knowledge. Technically speaking, data mining is the process of finding correlations or patterns in datasets where the overall goal is to extract information from a data set and transform it into an understandable format for further use. The process of knowledge discovery when dealing with datasets is illustrated in figure 1.1 below. These processes are based on a model which was initially presented by Fayyad in 1996 [9]. As demonstrated, the generic steps are: a) selection of the data from a database; b) pre-processing the target data; c) transformation of the data; d) mining the transformed data to identify patterns; e) interpretation and evaluation of the patterns to take appropriate actions.

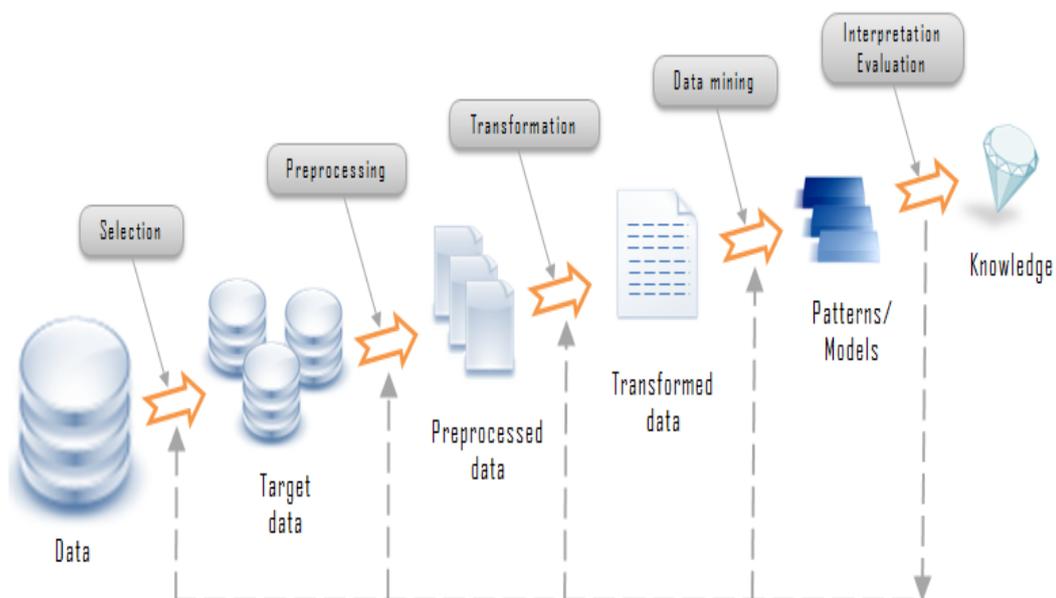


Figure 1.1: Illustration of knowledge discovery steps

1.2.1 DATA MINING TASKS AND PROCEDURES

Data mining tasks can be divided into five main categories: a) Anomaly detection; b) Rule association learning; c) Clustering; d) Classification; d) Regression analysis and e) Summarization. Data mining is an iterative process that typically involves the following phases:

Anomaly Detection:

Anomaly detection refers to the tasks conducted to solve the problem of finding patterns in data that do not conform to expected behaviour [10]. These nonconforming patterns are often referred to as anomalies, contaminant outliers, discordant observations, exceptions, aberrations or peculiarities in different application domains [11]. The main objective of anomaly detection is to identify cases that are unusual or rare within data that is seemingly homogeneous. Anomaly detection is increasingly used in medical science such as in the disease diagnosis from digital images [12], [13].

Association Rule Learning:

Association rules are used to express associations, connections and relationships between the variables in large datasets [14]. The general aim in association rule learning is to detect frequent or interesting patterns, associations, correlations among the data. Such algorithms are widely used in the financial [15], [16], marketing [17] and medical [18] domains.

Clustering:

Cluster analysis or simply clustering is the process of grouping a set of objects (according to some criteria) in a way that the objects that are more 'similar' to one another are put in the same group (called a cluster). There are several algorithms for clustering, which differ in their approach depending on the nature of the data and the area of application. Some common types of clustering algorithms are connectivity based clustering, centroid based clustering, distribution based clustering, and density based clustering. Clustering is used in several exploratory data mining situations, including medical diagnosis [19], [20], image segmentation [21], and in exploration of medical databases [22].

Classification:

Classification is the task of generalizing known structures to apply to new data. The process involves the mapping of values into classes using a derived function that provides a class representation of past trends enabling prediction based on classification dependencies [23].

Regression:

Regression attempts to find a function that models the data with the least error. Regression analysis is a statistical process for estimating the relationships among variables. It is widely used for prediction and forecasting.

Summarization:

The final task in data mining is summarization. It provides a more compact representation of the data set that includes description of main findings and report generation. In other words, it concludes or gives the abstract view.

1.2.2 MACHINE LEARNING AND PATTERN RECOGNITION

Machine learning is a branch of artificial intelligence that focuses on the construction and study of systems that can learn from data. The goal of machine learning is to make computers learn from example data or past experience to solve a given problem [24]. According to Arthur Samuel; machine learning can be defined as a "Field of study that gives computers the ability to learn without being explicitly programmed" [25]. Machine learning systems are representative and generalized. They represent data instances and have functions that are evaluated on these instances. They are generalized having capability of performing well on data instances other than the training data instances.

There is significant difference between machine learning and data mining. However, two terms are often confused with one another. Machine learning makes prediction on the basis of known properties learned from the training data whereas data mining is about discovering previously unknown properties in the data. These two areas overlap with each other as data mining could employ several machine learning techniques and for machine learning, data mining techniques can be used. Depending on the nature of the dataset, machine learning could range from supervised learning to unsupervised, semi-supervised or reinforced learning.

In machine learning, the term pattern recognition refers to as a process of assigning a label to a given input value. Classification is a most common type of pattern recognition technique. In classification each input value is assigned to one of a given set of classes [26]. Some other examples of pattern recognition are regression (assigning a real valued output to each input), sequence labelling (assigning a class to each input sequence), and parsing (assigning a parse tree to an input sentence). Medical community is using computer-aided diagnosis (CAD) systems. CAD is usually based on pattern recognition techniques that facilitate doctors in diagnosis and clinical decision making. Pattern recognition is widely used in medical informatics, medical imaging [27] and medical decision support systems [28].

Depending on the type of output label, there are several algorithms for pattern recognition. These algorithms differ in nature (statistical or non-statistical) and learning technique (supervised or unsupervised). Some prominent pattern recognition algorithms include classification algorithm (linear discriminant analysis, decision trees, and support vector machines), clustering algorithms (K-means clustering, hierarchical clustering), and regression algorithms (neural networks and principal components analysis) [29].

1.2.3 COMMON DATA MINING TECHNIQUES

A number of statistical and intelligent data mining techniques have been developed to date. The selection of a specific technique depends on the nature of the data set, the end goal and the area of application. Some of the common data mining techniques include classification, clustering, decision trees, rule induction, regression analysis, neural networks, and factor analysis [30]. The application

domain of data mining ranges from bioinformatics to drug discovery, business intelligence, risk analysis, decision support systems, market analysis, forecasting, and many other areas of science and engineering. Selection of appropriate data mining technique for a given problem is extremely important.

1.3 DATA MINING IN BIOMEDICINE

The healthcare industry collects a huge volume of data every year in the form of patient's history, clinical symptoms, pathology reports and outputs of medical imaging devices such as X-ray, MRI, CT-scan. Most of this data remains unmined and unexplored. Discovery of hidden patterns and correlations in data often goes untapped. This huge amount of untapped medical data can be turned into knowledge. For example, machine learning can be used to automatically extract diagnostic rules from past descriptions, and help medical specialists to make the accurate and reliable diagnosis. The decision support systems that have been developed in past to assist physicians in the diagnostic process are often based on static data that may not reveal the full patterns among the data. A decision support system that can learn the relationships between patient history, diseases prevalence in the population, symptoms, pathology of a disease, family history and test results is very much in demand in modern clinical practise [31]. This can be better and reliably achieved using intelligent system techniques. Whether it is genomics, study of human brain dynamics, analysis of biological relationships, image informatics, infectious disease study, or disease diagnosis, data mining finds its use in almost all areas of biomedicine [32].

With the advent of medical technology, clinicians now have an extensive amount of information available ranging from details of clinical symptoms to various types of biochemical data and outputs of medical imaging devices (such as X-ray, MRI, Ct-scan). Each type of data provides information that must be evaluated and assigned to a particular pathology during the diagnostic process. To modernize the conventional diagnostic process and avoid misdiagnosis, artificial intelligence methods (computer aided diagnosis) have been employed. These artificial intelligence based techniques have capacity to handle the complex and diverse types of medical data and process them to achieve desired goals [33]. This thesis focuses on such aspects of medical diagnosis by learning patterns through the collected data and developing medical decision support systems based on intelligent system techniques to support doctors in clinical decision making.

Computer-aided diagnosis (CAD) is an emerging field that has reshaped the field of medicine. CAD is a relatively new interdisciplinary technology that integrates powerful features of artificial intelligence and radiological image processing thus allowing medical specialists to conclude about the condition of the pathology. The role of CAD is to assist the medical specialists who are ultimately responsible for decision making. The emergence of computer-aided diagnosis has enabled radiologists to look very closely into the lumbar spine. They are now able to visualize the real-time changes in the anatomy of the spine, triggered naturally or by some disease. In this research, the characteristics of many spinal features were mined. The data was analysed with help of different data mining techniques to discover correlations between spinal features with age.

1.3.1 CHALLENGES AND SCOPE

Data mining has great potential for exploring the hidden patterns and unveiling the relationships in the medical data sets that can be utilized for clinical diagnosis. However, most of the raw medical data are voluminous, widely distributed, heterogeneous, and complex in nature. These data need to be collected in an organized form and require some pre-processing. The data used in this research is in the form of magnetic resonance images of the lumbar spine area. The MRI data in its raw form require noise removal and image pre-processing. Since the data is multivariate, it is very difficult to visualise, e.g., in 2D or 3D. Principal component analysis (PCA) is a commonly employed technique for dimensionality reduction. It can reduce the dimension of the input or output space, aid visualisation of the data, and uncover patterns that exist in the data. Data mining techniques provide a user friendly approach to uncover novel and hidden patterns in the data. A major challenge in this lumbar spine data analysis was that there was no reference bias available to compare with the results. Very few studies have been conducted on spinal variations with natural ageing and the conclusions are widely conflicting. A detailed literature review of the previous research in this area is given in chapter three. The validation of the results from this study therefore was a challenge. The results were continually discussed with medical specialists to confirm they are in line with medical sciences.

1.3.2 INTELLIGENT DATA ANALYSIS

Intelligent systems have proven suitable for the satisfactory diagnosis of various diseases. In addition, their use makes the diagnosis more reliable and therefore

increases patient satisfaction. However, despite their wide application of intelligent systems in modern diagnostics, they must be considered only as a facilitation tool for the clinicians. Ultimately clinicians will be responsible for critical evaluation of the intelligent system's output. Methods of summarizing and elaborating on informative and intelligent data analysis are continuously improving and hence can contribute greatly to effective, precise and swift medical diagnosis.

In the modern world, artificial intelligence is widely used in medicine for diagnosis, pattern recognition and clinical decision making. Neural networks are known for their strength when it comes to prediction. In this research a model was built for "spinal age prediction" using artificial neural networks. Using this model, medics can identify the early degeneration of lumbar spine. It was important to study the correlation between changing anatomy of spinal features and the age; and also the feature that move together (correlation among the features). A special type of unsupervised neural networks, called "Self Organising Maps" has the capacity to reflect the similarities among the features and was therefore used in this research to study the relationship among different spinal features. Self-organizing maps provided a good visual representation of multivariate lumbar spine data that helped in understanding the age related variations seen on the lumbar spine. SOM with colour coding particularly helped in formulating the relationships among the age, gender and 24 lumbar spine characteristics by comparing the maps of different spinal features. In addition to the SOM analysis, the results of Ward and modified Ward clustering techniques were presented in order to quantify the norms of different lumbar spine features among different groups.

Principal component analysis (PCA) is a powerful technique to visualise multivariate data. This technique is used to visualize 24 dimensional lumbar spine data and explain the variances among the scored data. Clustering is used in addition to PCA to group similar samples that help in understanding the behaviour of different groups. Factor analysis is used to assess the relationship between different spinal features. Fuzzy logic was used for generic rule extraction and designing an expert system that can predict the spinal age based on spinal features.

1.4 RESEARCH OBJECTIVES AND CONTRIBUTION TO THE KNOWLEDGE

In the developed world, the use of MRI for diagnosing lumbar spine problems has become standard practice. According to the OECD health report published in 2011, Japan and United States lead with the availability of MRI units per million population with 43.1 and 25.9 units respectively. Greece and United States has the highest number of MRI scans performed with 97.9 and 91.2 scans per 1000 population [34]. These numbers show that an enormous volume of data is gathered worldwide in the form of MRI. There is a need to design a sophisticated method that can mine this huge volume of data to explore hidden patterns and extract useful information. This work proposes, models, and validates individual and hybrid data analysis techniques so that useful knowledge can be extracted from raw MRI data. This work uses a Lumbar Spine MRI dataset to classify the age related changes in the lumbar spine. Classification of lumbar spine features will be helpful in setting the standards for spinal growth and degeneration with age (which are not available to date). If the ageing pattern of the lumbar spine is known, medical specialists can

easily classify a healthy spine. The data set used in this research was genuine and unique; it has not been used anywhere else for similar research. Studying the ageing pattern of the spine is an area that has received little attention. There is no significant research describing the ageing patterns of the lumbar spine in depth. Researchers from the medical and engineering communities have attempted to study the ageing effects on different areas of the lumbar spine either focusing on one feature or a small group of 4-5 features at a time. Considering many lumbar spinal features simultaneously makes it very complex to model and difficult to investigate. In this thesis twenty four spinal features are studied together.

Artificial intelligence is widely used in medicine for diagnosis, pattern recognition, and clinical decision making. This research classifies the spine behaviour and ranks spines on the basis of their physical features. The growth and degeneration patterns of the spine not only helped in understanding the changing anatomy of the spine with the age but also in better understanding spine related diseases. The data mining tools and algorithms, independent of their nature and characteristics, are designed to solve problems by producing the most effective and efficient result. These tools are constantly redefined, optimised and improved in order to enhance their accuracy and boost their performance, which ultimately results in the generation of valuable knowledge. The research reported in this thesis suggests the selection of appropriate data analysis techniques for complex medical data and validates the use of these data mining techniques against the given dataset. The unique contribution of this thesis comes from three directions: a) the application and testing of several individual and hybrid data mining techniques in the context

of a complex medical application; b) setting the standards for spinal growth and degeneration with age that can be adopted as a reference for the medical community; c) providing an aid to the medical specialists for disease diagnosis and clinical decision making.

1.5 THESIS OUTLINE

This thesis comprises seven chapters. The first three introduce the problem, present the basics of the intelligent systems techniques used and provide a detailed literature review. Chapters four, five and six contain the novel contribution of the author and explain the significant results. The last chapter summarises the main findings of this research, along with discussing future research directions.

The first chapter gives a brief overview of the thesis. Chapter 2 provides a background of some intelligent systems techniques (IST) that are widely used in the fields of classification, data mining, clustering, and pattern recognition. It covers both theoretical and practical aspects of intelligent system techniques that are relevant to this thesis, including Artificial Neural Networks (ANN), Self-Organising Maps (SOM), Fuzzy Logic (FL), and Clustering Analysis.

Chapter 3 gives a detailed insight of the characteristics, features and anatomy of a human spine. It explains various problems associated with the human spine and the growth and degenerative process in the light of previous studies in this area. It gives a comprehensive description of data set, feature extraction, feature measurement and the scoring criteria for spinal MRIs (selection of the most noticeable features from raw data).

Chapter 4 describes the analysis of extracted MRIs data using principal component analysis (PCA). PCA for visualization and Clustering for anomaly detection are presented in this chapter. It also demonstrates the use of Factor Analysis (FA) for estimating the factors that influence the responses of observed variables. It allows us to investigate the concepts that are not measured directly, by collapsing a large number of variables into a few interpretable underlying factors.

In Chapter 5, a neural network model is presented. This model is capable of predicting the spinal age from the extracted features of the spine. A new term “Spinal Age” is introduced as an identifier to explain the overall behaviour of the spine. In this chapter, a principal component neural network model is presented along with the comparison of both neural network models. Furthermore, a fuzzy and a hybrid intelligent model (adaptive neuro-fuzzy inference system) are presented in this chapter to extract simple linguistic rules from the data.

Chapter 6 demonstrates the results produced by using self-organizing maps and highlights the age related degenerative changes in the human spine. In this chapter, two clustering techniques are used in addition to SOM and a comparison of the results is provided along with the other findings.

The final chapter summarises the main findings of the research presented throughout this thesis. It also provides a brief comparison of all the techniques used in this thesis. It highlights the significance of work done and its potential value in medical decision making. This chapter also provides the final prototype of the spinal diagnostic model. The thesis closes by outlining a framework for further research and recommending potential enhancements.

REFERENCES

- [1] Singh, Hardeep, Ashley ND Meyer, and Eric J. Thomas. "The frequency of diagnostic errors in outpatient care: estimations from three large observational studies involving US adult populations." *BMJ quality & safety* (2014): bmjqs-2013.
- [2] Newman-Toker, David E., and Peter J. Pronovost. "Diagnostic errors—the next frontier for patient safety." *JAMA* 301, no. 10 (2009): 1060-1062.
- [3] Winters, Bradford, Jason Custer, Samuel M. Galvagno, Elizabeth Colantuoni, Shruti G. Kapoor, HeeWon Lee, Victoria Goode et al. "Diagnostic errors in the intensive care unit: a systematic review of autopsy studies." *BMJ quality & safety* 21, no. 11 (2012): 894-902.
- [4] Tehrani, Ali S. Saber, HeeWon Lee, Simon C. Mathews, Andrew Shore, Martin A. Makary, Peter J. Pronovost, and David E. Newman-Toker. "25-Year summary of US malpractice claims for diagnostic errors 1986–2010: an analysis from the National Practitioner Data Bank." *BMJ quality & safety* 22, no. 8 (2013): 672-680.
- [5] Videman, Tapio, Michele C. Battié, Laura E. Gibbons, and Kevin Gill. "Aging Changes in Lumbar Discs and Vertebrae and Their Interaction A 15-year Follow-up Study." *The Spine Journal* 14, no. 3 (2013): 469-478.
- [6] Diacinti, D., M. Acca, E. D'Erasmo, E. Tomei, and G. F. Mazzuoli. "Aging changes in vertebral morphometry." *Calcified tissue international* 57, no. 6 (1995): 426-429.

- [7] Mosekilde, Lis, and Leif Mosekilde. "Sex differences in age-related changes in vertebral body size, density and biomechanical competence in normal individuals." *Bone* 11, no. 2 (1990): 67-73.
- [8] Shirgaonkar, Saurabh, T. Rajkumar, and V. Singh. "Application of improved apriori in university library." In *Proceedings of the International Conference and Workshop on Emerging Trends in Technology*, ACM, (2010): 535-540.
- [9] Fayyad, U. M., G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy. "Advances in Knowledge Discovery and Data Mining." Association for the Advancement of Artificial Intelligence (AAAI), (1996).
- [10] Bertino, Elisa. "Data Protection from Insider Threats." *Synthesis Lectures on Data Management* 4, no. 4 (2012): 1-91.
- [11] Niu, Zhixian, Shuping Shi, Jingyu Sun, and Xiu He. "A survey of outlier detection methodologies and their applications." In *Artificial Intelligence and Computational Intelligence*, Springer Berlin Heidelberg, (2011): 380-387.
- [12] Beutel, Jacob, and Richard G. Stafford. "Application of neural networks as an aid in medical diagnosis and general anomaly detection." U.S. Patent 5,331,550, issued July 19, (1994).
- [13] Chen, Li-Fei. "An improved negative selection approach for anomaly detection: with applications in medical diagnosis and quality inspection." *Neural Computing and Applications* 22, no. 5 (2013): 901-910.

- [14] Seel, Norbert M., ed. "Encyclopedia of the Sciences of Learning." Springer Verlag, (2012).
- [15] Giles, C. Lee, Steve Lawrence, and Ah Chung Tsoi. "Rule inference for financial prediction using recurrent neural networks." In Computational Intelligence for Financial Engineering (CIFEr), 1997., Proceedings of the IEEE/IAFE (1997): 253-259.
- [16] Martens, David, Bart Baesens, Tony Van Gestel, and Jan Vanthienen. "Comprehensible credit scoring models using rule extraction from support vector machines." European journal of operational research 183, no. 3 (2007): 1466-1476.
- [17] Johansson, Ulf, C. Sonstrod, R. Konig, and Lars Niklasson. "Neural networks and rule extraction for prediction and explanation in the marketing domain." In Neural Networks proceedings, IEEE, 4, (2003): 2866-2871.
- [18] Kahramanli, Humar, and Novruz Allahverdi. "Rule extraction from trained adaptive neural networks using artificial immune systems." Expert Systems with Applications 36, no. 2 (2009): 1513-1522.
- [19] Greene, Derek, Alexey Tsymbal, Nadia Bolshakova, and Pdraig Cunningham. "Ensemble clustering in medical diagnostics." In Computer-Based Medical Systems, CBMS 2004 proceedings, IEEE, (2004): 579-581.
- [20] Masulli, Francesco, and Andrea Schenone. "A fuzzy clustering based segmentation system as support to diagnosis in medical imaging." Artificial Intelligence in Medicine 16, no. 2 (1999): 129-147.

- [21] Haralick, Robert M., and Linda G. Shapiro. "Image segmentation techniques." *Computer vision, graphics, and image processing* 29, no. 1 (1985): 100-132.
- [22] Berks, Georg, Diedrich Graf V. Keyserlingk, Jan Jantzen, Mariagrazia Dotoli, and Hubertus Axer. "Fuzzy clustering-a versatile mean to explore medical database." *ESIT 2000, Aachen, Germany* (2000).
- [23] Nemat, Hamid R., and Christopher D. Barko, eds. "Organizational data mining: leveraging enterprise data resources for optimal performance." IGI Global, (2004).
- [24] Alpaydin, Ethem. "Introduction to machine learning." MIT press, (2004).
- [25] Simon, Phil. "Too Big to Ignore: The Business Case for Big Data." John Wiley & Sons, (2013), ISBN 978-1118638170.
- [26] Bishop, Christopher M. "Pattern recognition and machine learning." Vol. 1. New York: Springer, (2006).
- [27] Meyer-Baese, Anke. "Pattern recognition for medical imaging." Academic Press, (2004).
- [28] Coomans, D., and I. Broeckaert. "Potential pattern recognition in chemical and medical decision making." John Wiley & Sons, Inc., (1986).
- [29] Pal, Sankar K., and Pabitra Mitra. "Pattern recognition algorithms for data mining." CRC press, (2004).

- [30] Han, Jiawei, Micheline Kamber, and Jian Pei. "Data mining: concepts and techniques." Morgan kaufmann, (2006).
- [31] Jaspers, Monique WM, Marian Smeulers, Hester Vermeulen, and Linda W. Peute. "Effects of clinical decision-support systems on practitioner performance and patient outcomes: a synthesis of high-quality systematic review findings." *Journal of the American Medical Informatics Association* 18, no. 3 (2011): 327-334.
- [32] Chen, Hsinchun, Sherrilynne S. Fuller, Carol Friedman, and William Hersh. "Knowledge management, data mining, and text mining in medical informatics." In *Medical Informatics*, pp. 3-33. Springer US, (2005).
- [33] Amato, Filippo, Alberto López, Eladia María Peña-Méndez, Petr Vaňhara, Aleš Hampl, and Josef Havel. "Artificial neural networks in medical diagnosis." *Journal of Applied Biomedicine* 11, no. 2 (2013): 47-58.
- [34] Organization for Economic Cooperation and Development OECD. "Health at a Glance 2011: OECD Indicators." OECD publishing, (2011).

2

INTELLIGENT SYSTEMS TECHNIQUES

2.1 INTRODUCTION TO INTELLIGENT SYSTEMS

2.2 ARTIFICIAL NEURAL NETWORKS

2.2.1 TYPES OF NEURAL NETWORKS

2.2.2 SUPERVISED, UNSUPERVISED AND REINFORCED
LEARNING

2.3 FEED-FORWARD NEURAL NETWORKS

2.4 SELF-ORGANISING MAPS (SOMS)

2.4.1 STRUCTURE OF SOM

2.4.2 SOM TRAINING AND LEARNING

2.5 CLUSTERING ANALYSIS

2.6 FUZZY LOGIC

2.6.1 MAMDANI TYPE FUZZY INFERENCE

2.6.2 SUGENO TYPE FUZZY INFERENCE

2.7 HYBRID INTELLIGENT TECHNIQUES

2.7.1 ADAPTIVE NEURO FUZZY INFERENCE SYSTEM (ANFIS)

2.8 SUMMARY

REFERENCES

2.1 INTRODUCTION TO INTELLIGENT SYSTEMS

Artificial intelligence (AI) is a branch of computer science that studies and develops machines and software with intelligence. Intelligent Systems are based on artificial intelligence methods, which have ability to perceive, reason, learn, and act intelligently. The term Intelligent Systems is used interchangeably with 'Soft Computing'. It is a collection of techniques that one way or another provides flexible information processing capability for handling real-life situations. Compared to conventional data analysis techniques (e.g. statistical methods); intelligent techniques are tolerant to imprecision, uncertainty, partial truth, and approximation [1]. Intelligent systems provide tractability, robustness, and low-cost solutions for applications in many domains.

An intelligent system has ability of learning how to act so that its objectives can be reached. The definition of intelligent systems differ in many contexts and is subject to a great deal of debate. From the perspective of computation, the intelligence of a system can be characterized by its flexibility, adaptability, memory, learning, temporal dynamics, reasoning, and the ability to manage uncertain and imprecise information [2]. An intelligent system emulates some aspects of intelligence exhibited by nature. Many artificial intelligent techniques can support/interact with each other. However, each discipline has its own distinct attributes that make it particularly useful for certain types of problems and applications. Statistical techniques alone may not be sufficient to address some of the more challenging issues in data mining, especially those arising from very large and complex datasets

[3]. However, the power of intelligent systems techniques renders them very useful tools for data mining and other practical applications.

This chapter provides a background of some intelligent systems techniques that are widely used in the field of classification, data mining, feature extraction, clustering, and pattern recognition. It begins with a thorough review of both theoretical and practical aspects of intelligent system techniques used in this thesis. The techniques used in this are listed below.

Table 2.1: List of intelligent techniques used in this thesis

Artificial Neural Network (ANN)
Self-Organising Map (SOM)
Clustering Analysis
Fuzzy Inference System
Adaptive Neuro-Fuzzy Inference System (ANFIS)

2.2 ARTIFICIAL NEURAL NETWORKS

Artificial neural networks (ANNs) refer to a mathematical model inspired by biological neural networks found in the human brain [4]. The brain is the most important and complex parts of the human anatomy. It is well known that the biological neural systems can perform extraordinarily complex tasks and are capable of learning over time [5]. ANNs in nutshell are a computationally primitive approximation of the human brain.

The smallest unit of an ANN is a simple arithmetic processing unit, namely an Artificial Neuron (AN), which can be considered as a mathematical model of a biological neuron. Each artificial neuron is responsible for acquiring information from one or more input signals and generating a single output signal. Similarities in the design and functionalities between biological and artificial neurons can be seen in figure 2.1 where an artificial neuron is represented in its simplest form with its inputs, weights, transfer function, bias and outputs. The model is closely associated to its biological root with its soma, dendrites and axon. Multiple input signals can be injected to an artificial neuron simultaneously. In response, similar to the biological model, a neuron either triggers (fires) or not; based on a threshold level. A simple ANN will allow a single output. Additionally, ANNs may provide a mechanism for other factors such as a forcing term also known as a bias [6]. Figure 2.1 demonstrates the relationship between the inputs/bias and the output variable.

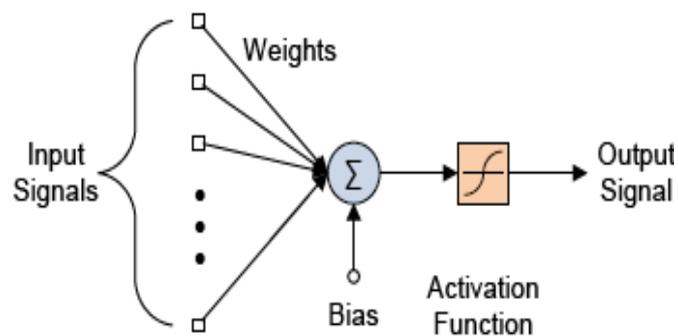


Figure 2.1: A simple artificial neuron design

In mathematical terms, the function of an artificial neuron can be expressed as follows:

$$y = f\left(\sum_i x_i w_i + b\right) \quad (2.1)$$

Where x_i represents the i^{th} input signal from the connected neurons, w_i is the corresponding weight assigned to that neuron, b is the bias, f is the activation function and y is the output of the artificial neuron.

2.2.1 TYPES OF ARTIFICIAL NEURAL NETWORKS

Artificial neural networks have several types depending on the number of layers, direction of logic, and feedback loops. There are many types of artificial neural networks. Some common types of artificial neural networks are discussed below:

Feed-Forward Neural Network (FFNN): One of the simplest types of ANN is feedforward neural network in which the information moves only in forward direction. The data from the input nodes flows through the hidden nodes (if there is any) to reach the output nodes. This network is free from any loops or cycles. Different types of units, e.g. binary McCulloch-Pitts neurons, and perceptron can be used to build feedforward networks.

Radial Basis Function (RBF) Network: Radial basis functions are useful for interpolation in multidimensional space. A RBF is a function that incorporates a distance criterion with respect to a centre. Radial basis functions have been applied on ANN where they may be used in place for the sigmoidal hidden layer transfer characteristic in multilayer perceptrons (MLP).

Kohonen Self-Organizing Network: The self-organizing map (SOM) was introduced by Teuvo Kohonen. SOM is based on unsupervised learning in which a set of neurons learn to map points in an input space to coordinates in an output space.

The input and output space can differ in dimensions and topologies however SOM tries to preserve these topologies.

Learning Vector Quantization: Learning Vector Quantization (LVQ) was also suggested by Teuvo Kohonen. LVQ is often considered as neural network architecture that combines competitive learning with supervision.

Recurrent Neural Network: Unlike feedforward neural networks, recurrent neural networks (RNN) have bi-directional data flow. In a feedforward neural network data propagates linearly from input to output layer but in RNN data also propagates from later processing stages to earlier ones. Recurrent neural networks can be used as general sequence processors.

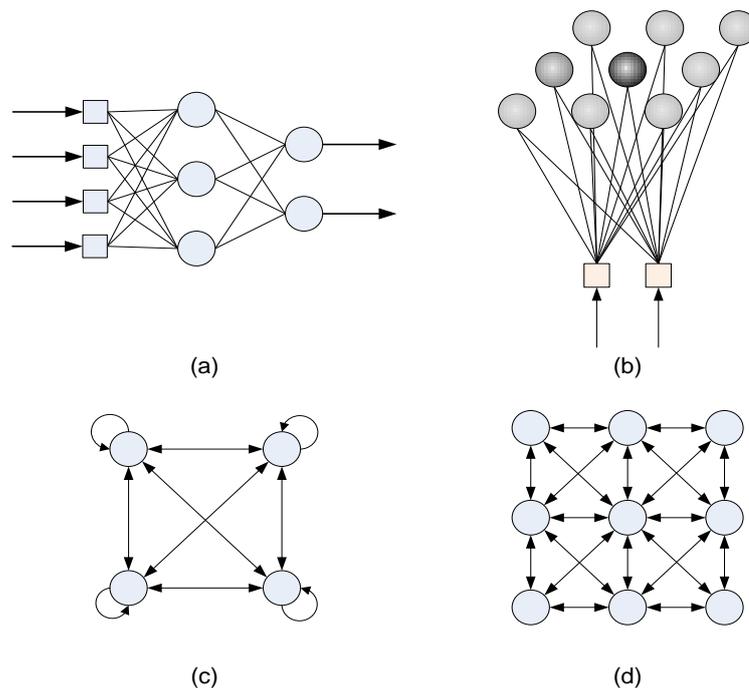


Figure 2.2: Architecture of common ANNs. (a) Multilayer feedforward neural network; (b) Self-organizing map; (c) Multilayer recurrent neural network; (d)

Cellular neural network

2.2.2 *SUPERVISED, UNSUPERVISED, AND REINFORCED LEARNING*

On the basis of learning techniques, neural networks can be categorised into three main groups as supervised learning, unsupervised learning and reinforcement learning. In supervised learning the model is provided with the correct output during the training process. Data is fed to the network in a pair consisting of an input sample and a desired output value. A supervised learning algorithm examines the training data and produces an inferred function, which can be used for mapping new samples. Some commonly used supervised learning methods include backpropagation, Bayesian network, decision trees, and support vector machines.

In the context of machine learning, an unsupervised learning is the method of finding the hidden patterns in unlabelled data. The examples given to the learner are unlabelled so there is no error or reward signal to evaluate a potential solution [7]. This differentiates unsupervised learning from supervised and reinforcement learnings. Some commonly used unsupervised learning approaches include clustering (k-mean, hierarchical, fuzzy, mixture models etc.), adaptive resonance theory (ART), and self-organizing maps (SOM).

Reinforcement learning is learning what to do, and how to map situations to actions, so as to maximize a numerical reward signal. As in most of the forms of machine learning, the learner is not told which actions to take but instead must discover which actions yield the most reward by trying them [8]. Common methods used for reinforced learning are temporal difference (TD) learning, Q-learning, and State-Action-Reward-State-Action (SARSA). In this thesis, both supervised and

unsupervised learning methods are used along with conventional statistical techniques.

2.3 FEED-FORWARD NEURAL NETWORKS

Different types of neural networks were discussed in the previous section. In this thesis a feed-forward neural network is used due to its ability to handle high dimensional inputs. Also, such networks possess a relatively simple structure that is easy to implement. A feedforward neural network is an ANN in which information flows in only forward direction. The connections between the units do not form a cycle or loop. In this network, the information propagates from the input nodes to the output nodes through the hidden nodes (if any). Figure 2.3 below shows multilayer feedforward neural network. This network has one input layer (having 3 input neurons), one hidden layer (having 3 hidden neurons), and one output layer (having two output neurons). It can be seen that there is no connection between the neuron in a layer. However, each neuron is connected to the neurons of next layer. The input is connected to the first layer of the network. Each subsequent layer in the network is connected from the previous layer and the final layer produces the network's output.

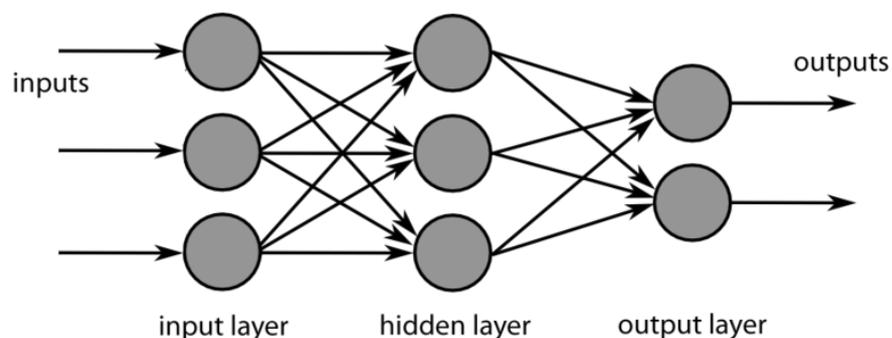


Figure 2.3: Multi-layer feedforward artificial neural network

Activation Function:

The activation function is a key element in the artificial neurons, as it characterises the behaviours (linear, non-linear, discrete, etc.) of artificial neurons. The activation function defines the output of a neuron in terms of the induced local field. There are several types of activation functions. Some basic well-known types of activation function are shown in table 2.2 below.

Table 2.2: Different activation functions with their formulas

Name	Formula
Identity Function	$f(n) = n$
Threshold Function	$f(n) = \begin{cases} 1 & \text{if } n \geq 0 \\ 0 & \text{if } n < 0 \end{cases}$
Piecewise-Linear Function	$f(n) = \begin{cases} 1, & n \geq +\frac{1}{2} \\ n, & +\frac{1}{2} > n > -\frac{1}{2} \\ 0, & n \leq -\frac{1}{2} \end{cases}$
Sigmoid Function	$\varphi(n) = \frac{1}{1 + \exp(-n)}$
Hyperbolic Tangent (tanh) Function	$f(n) = \tanh(n) = \frac{\exp(n) - \exp(-n)}{\exp(n) + \exp(-n)}$

Training and Learning Algorithm:

In artificial neural networks, weights and biases are adjusted through training and learning functions. The role of training function is to command a global algorithm that affects all the weights and biases of the network. The learning function can be applied to individual weights and biases within a network. There are number of training algorithms which includes gradient descent methods, conjugate gradient

methods, the Levenberg-Marquardt algorithm (LM), and the backpropagation algorithm. Learning methods include gradient descent, Hebbian learning, Widrow-Hoff.

2.4 SELF-ORGANISING MAPS

The self-organising map (SOM) algorithm was developed by Kohonen to transform a data-set of arbitrary dimension into a one- or two- dimensional discrete map [9]. Kohonen's Self-organising Map (SOM) with unsupervised learning is amongst the most common ANN models. Like other ANNs, SOM has its origin in biology too (i.e. the visual cortex). According to Teuvo Kohonen [9]; "The SOM is a new, effective software tool for the visualization of high-dimensional data. It converts complex, nonlinear statistical relationships between high-dimensional data items into simple geometric relationships on a low-dimensional display. As it thereby compresses information while preserving the most important topological and metric relationships of the primary data items on the display, it may also be thought to produce some kind of abstractions".

Neurons in SOM are often arranged in a two dimensional rectangular grid, forming a discrete topological mapping of the input space and then self-organise based on the presented observations with no external influence. The map will eventually reach convergence depending on the size and dimensionality of the input vector. Self-organizing maps use a neighbourhood function to preserve the topological properties of the input space which differentiate them from other artificial neural networks. SOM is a useful tool for visualizing high-dimensional data on a low-dimensional map, similar to multidimensional scaling. SOMs when trained

successfully can cluster the data presented to them by providing topologically preserved mapping from input to output space which may be used to discover clusters and relationships among input dataset. This makes SOM resilient with respect to faulty, noisy or unknown observations. These characteristics make SOM an attractive method to be applied in the fields of vector quantisation, dimensionality reduction, data visualisation, clustering, and classification [10].

2.4.1 *STRUCTURE OF SELF-ORGANISING MAPS*

A self-organizing map consists of small components called nodes or neurons. Each node has a weight vector of the same dimension as the input data vectors and a specific position in the map space. Usually the nodes are arranged in a two-dimensional hexagonal or rectangular grid. During the self-organising process, competition and cooperation is introduced between neurons, so that only one or few units in the network respond to each input pattern. Each neuron will learn to respond best to a cluster of similar patterns. Based on similarity, different neurons will respond to different clusters in the input space. In this way, the network will learn in a map-like representation of the inputs. The model typically consists of two sheets (layers) of neural units: input layer and computational layer. Figure 2.4 demonstrates the interaction between input and computation layers.

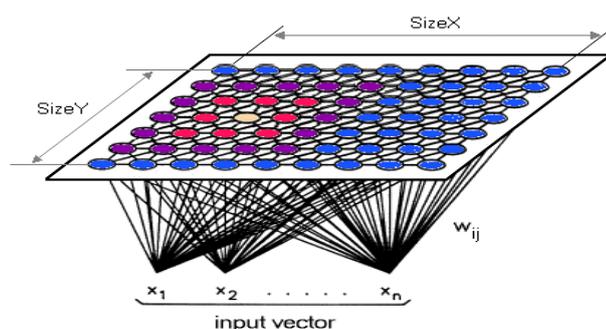


Figure 2.4: Self-organizing map (SOM) structure

If the input space is N dimensional, the input pattern can be written as $x = [x_i : i = 1, \dots, N]$. The relation between connection weights, the total number of neurons N , the input units i and the neurons j in the computation layer can be written as $w_j = \{w_{ji} : j = 1, \dots, N; i = 1, \dots, N\}$. The discriminant function can be defined as squared Euclidean distance between the input vector x and the weight vector w_j for each neuron j .

$$d_j(x) = \sum_{i=1}^N (x_i - w_{ji})^2 \quad (2.2)$$

The neuron with weight vector closest to the input vector will be declared as a winning neuron.

2.4.2 SOM TRAINING AND LEARNING

In SOM, several units compete for the current object. With incoming data, the network of artificial neurons is trained by giving information about the inputs. The weight vector of the unit closest to the current object becomes the winning unit. In training stage, the values for the input variables are gradually adjusted to preserve the existing neighbourhood relationships within the input data set. SOM training is shown below in figure 2.5 below.

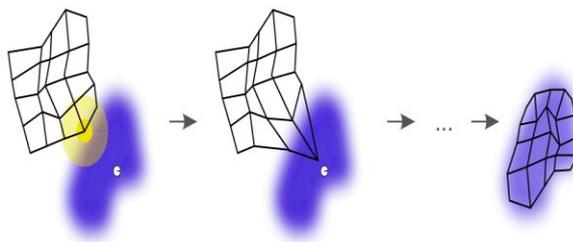


Figure 2.5: Training of self-organizing map (SOM)

Every sample from the dataset recognizes them by competing for representation. Mapping in SOM begins with the initialization of the weight vectors. Then a sample vector is picked randomly and the weight vectors in the map are searched to find best representation of that sample. Each weight vector has information about neighbouring weights. The chosen weight as well as its neighbours is rewarded by being able to become more like that randomly selected sample vector. This whole process is repeated several times. The self-organization training process involves four major steps:

- Initialization: The training process starts with the initialization of connection weights with small random values.
- Competition: For every input vector, the neurons compute their respective values of a discriminant function which gives the basis for competition. The value of the discriminant function decides the winning neuron.
- Cooperation: The winning neuron determines the spatial location of a topological neighbourhood of excited neurons, thereby providing the basis for cooperation among neighbouring neurons.
- Adaptation: The excited neurons adjust their individual values of the discriminant function in relation to the input vector. These adjustments are made through associated connection weights in such a way that the response of the winning neuron to the successive application of a similar input vector is enhanced.

Unlike other learning techniques in neural networks, training of SOM is unsupervised meaning that no target vector is required. A Unified distance matrix

(U-Matrix) is often used for visualizing SOMs. The value of a particular node in unified distance matrix presents the mean distance between the node and its closest neighbours [11]. Once the SOM is trained using the input data, the distance between the codebook vectors of neighbouring neurons gives an approximation of the distance between different parts of the underlying data. When such distances are represented in a gray scale image, light colours depict closely spaced node codebook vectors and darker colours indicate more widely separated node codebook vectors [12]. Thus, it can be said that the group of light colours forms a cluster whereas the dark parts serves as a boundaries between the clusters. Such representation helps to visualize the clusters in the low-dimensional space. These clusters can be recognized in an automated way by using relatively simple image processing techniques.

2.5 CLUSTER ANALYSIS

Cluster analysis or simply clustering is the process of grouping a set of objects (according to some criteria) in a way that the objects that are more 'similar' to one another are put in the same group (called a cluster). Clustering is among one of the main task of exploratory data mining and statistical data analysis. Clustering found its application in number of fields, ranging from machine learning, pattern recognition, image analysis, information retrieval, and bioinformatics. Clustering is usually an unsupervised learning process that aims to learn the structural or spatial relationships and similarities among data samples. Clustering techniques can be applied indiscernibly to highly dimensional feature space in which a human would

find it difficult to identify clusters. Generally the process of clustering involves the following three steps [13]:

- Defining a dissimilarity measure between data samples, e.g. Euclidean distances.
- Defining the criterion for clustering to be optimised in terms of within or between cluster structures.
- Defining an algorithm to find a correct assignment of samples to clusters.

There are different kinds of clustering techniques. It depends on the nature of data and application area for selecting the appropriate clustering techniques. Some most commonly used clustering techniques are hierarchical based clustering, density based clustering and partitioning-based clustering (K-means clustering) [14]. K-means is one of the widely used clustering techniques in practical applications. In K-mean clustering, n observations are partitioned into k clusters in which each observation belongs to the cluster with the closest mean. Another common type of clustering is density-based clustering, in which clusters are defined as areas of higher density than the remainder of the data set [15]. Hierarchical clustering is a method of cluster analysis, which seeks to build a hierarchy of clusters. Generally there are two types of hierarchical clustering techniques as Agglomerative and Divisive. Agglomerative clustering is a "bottom up" approach where each observation is initially considered as a cluster and then pairs of clusters are merged as one move up the hierarchy. Whereas divisive clustering is a "top down" approach where all observations are initially considered to be a single cluster and splits are

performed recursively as one move down the hierarchy. Generally, these merges and splits are determined in a greedy manner [16]. Greedy algorithm follows the problem solving heuristic of making the locally optimal choice at each stage with the hope of finding a global optimum. Ward's clustering method is another type of hierarchical cluster analysis. Ward's clustering is used and explained in chapter six.

2.6 FUZZY LOGIC

The term "fuzzy logic" (FL) was introduced with the 1965 proposal of fuzzy set theory by Lotfi A. Zadeh [17]. Fuzzy logic is a form of multi-valued logic or probabilistic logic that deals with approximate reasoning rather than fixed or exact one. Compared to conventional binary logic (where variables can have either true or false values), fuzzy logic variables may have a truth value that ranges anything between 0 and 1. Fuzzy logic has been extended to handle the concept of partial truth, where the truth value may range between completely true and completely false [18]. In recent years, Fuzzy Logic has been employed extensively in industrial control, machine learning and image processing [19]. Variables in conventional logic usually take numerical values but in fuzzy logic, the non-numeric linguistic variables (e.g. high, low, medium) are often used for expression of rules and facts [20]. Fuzzy logic is used to deal with issues of uncertainty in classical logic. In this thesis fuzzy logic is used to develop a fuzzy inference system (FIS) capable of computing spinal age based on lumbar spine characteristics. This expert system also validates the previous knowledge (generic rules) extracted from PCA and SOM analysis.

2.6.1 MAMDANI TYPE FUZZY INFERENCE

The basic architecture of fuzzy logic (FL) is based on the concept of a 'crisp' input and 'crisp' output. Fuzzy inference systems (FIS) are one of the most famous applications of fuzzy logic and fuzzy sets theory [21]. They are widely used in classification, offline process simulation, diagnosis, decision support system, and process control. Fuzzy inference is a process of formulating the mapping from given input(s) to output(s) using fuzzy logic. The mapping them provides a basis from which decisions can be made, or patterns discerned. There are two main types of fuzzy inference systems as Mamdani and Sugeno. Mamdani type fuzzy inference is the most common fuzzy inference method. It was proposed in 1975 by Ebrahim Mamdani [22] as an attempt to control a steam engine and boiler combination by synthesizing a set of linguistic control rules obtained from experienced human operators. The process of Mamdani fuzzy inference involves four steps: fuzzification, defuzzification, implication and aggregation. Fuzzification means the mapping of crisp inputs to fuzzy membership function (degree of belongingness). Implication is the process of evaluating an individual rule. Aggregation is the process of computing the cumulative output from the output of all the rules. Finally, this cumulative fuzzy value is mapped back onto crisp value via defuzzification process. The architecture of Mamdani FIS is shown in Figure 2.6.

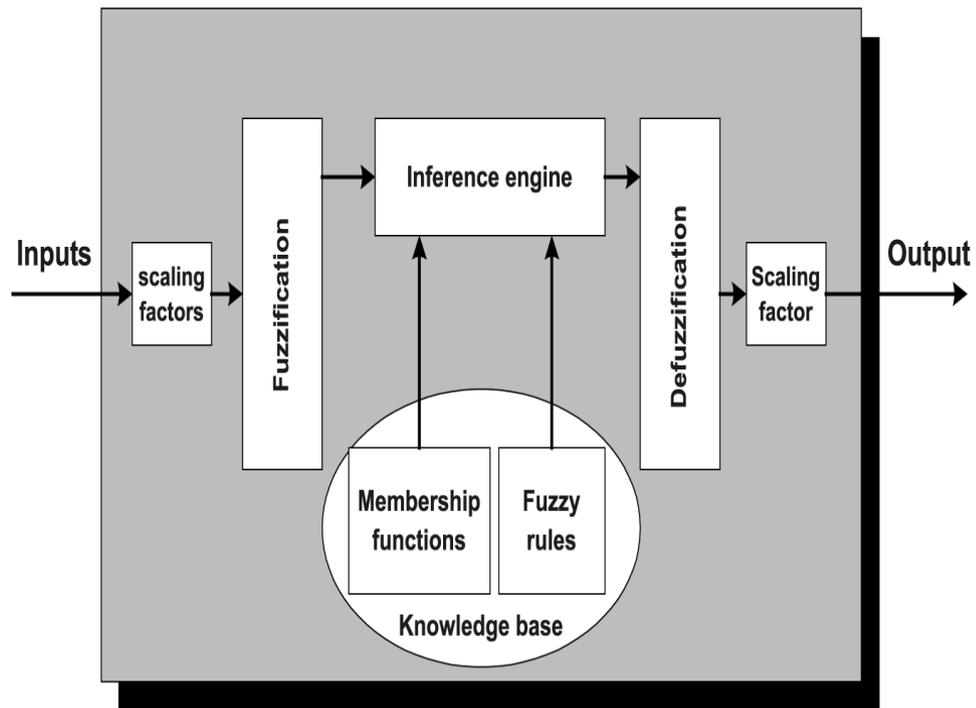


Figure 2.6: Architecture of Mamdani fuzzy inference systems

2.6.2 *SUGENO TYPE FUZZY INFERENCE*

Sugeno type fuzzy inference systems were introduced in 1985 [23]. The architecture of Sugeno FIS is very similar to the Mamdani FIS. The only difference between the two methods is the nature of output membership functions. In Sugeno type FIS; output membership functions are either linear or constant. The advantage of Mamdani method is that they are capable of capturing expert knowledge and is well suited to human input. On the other hand, Sugeno method is efficient in terms of computations and works well with linear, optimization and adaptive techniques.

2.7 HYBRID INTELLIGENT SYSTEMS

Hybrid Intelligent system refers to the combination of two or more intelligent techniques. Fusion of different intelligent techniques such as of artificial neural

network, fuzzy logic, rough set, genetic algorithm, evolutionary algorithms and swarm intelligence gives a computationally intelligent system capable of achieving desired results. Also, the integration of different learning and adaptation techniques helped to overcome the individual limitations of a certain techniques. All individual techniques have their constraints and limitations. Having the possibility to put two or more of them together in a hybrid system can improve the system's capabilities and performance. These hybrid intelligent systems can have combination, integration, fusion and association of two or more intelligent techniques [24], [25]. In this thesis, a hybrid intelligent technique consisting of fuzzy logic and a neural network is used. A fuzzy inference system is used in this thesis in order to predict the "spinal age" from 24 spinal features based on linguistic if-then rules. This fuzzy inference system however lacked the ability to learn from the data. Therefore all its parameters (shapes, ranges and overlapping of membership functions) needed to be tuned manually. To overcome the issue of manual tuning, a neural network was used to optimise the parameters of the fuzzy inference system. This makes the fuzzy inference system adaptive; hence the term adaptive neuro-fuzzy inference system. A brief explanation of this technique is given in this chapter.

2.7.1 ADAPTIVE NEURO-FUZZY INFERENCE SYSTEM

The term Neuro-fuzzy refers to the combination of artificial neural networks and fuzzy logic. Fusion of ANN and fuzzy logic results in a hybrid intelligent system that combines the human-like reasoning style of fuzzy systems with the learning and adaptation ability of artificial neural networks. The main strength of neuro-fuzzy

systems is that they are use easy to understand linguistic IF-THEN rules and they have ability to learn.

The idea behind an adaptive neuro-fuzzy inference system (ANFS) was proposed in [26], which suggested that by using a given input/output data set, a fuzzy inference system (FIS) can be constructed and its membership function parameters can be tuned (adjusted) using neural networks. This fine-tuning allows fuzzy systems to learn from the data they are modelling. Thus, a Fuzzy Inference system (FIS) has the ability to deal with linguistic expression and a neural network has the ability to self-learn and improve. Integrating the best features of these two techniques provide leads to a powerful method for analysing data. The architecture of ANFIS contains a five-layer feed-forward neural network as shown in Figure 2.7. Here, x_1, x_2 are inputs; A_1, A_2, B_1, B_2 are membership functions, and y is the output.

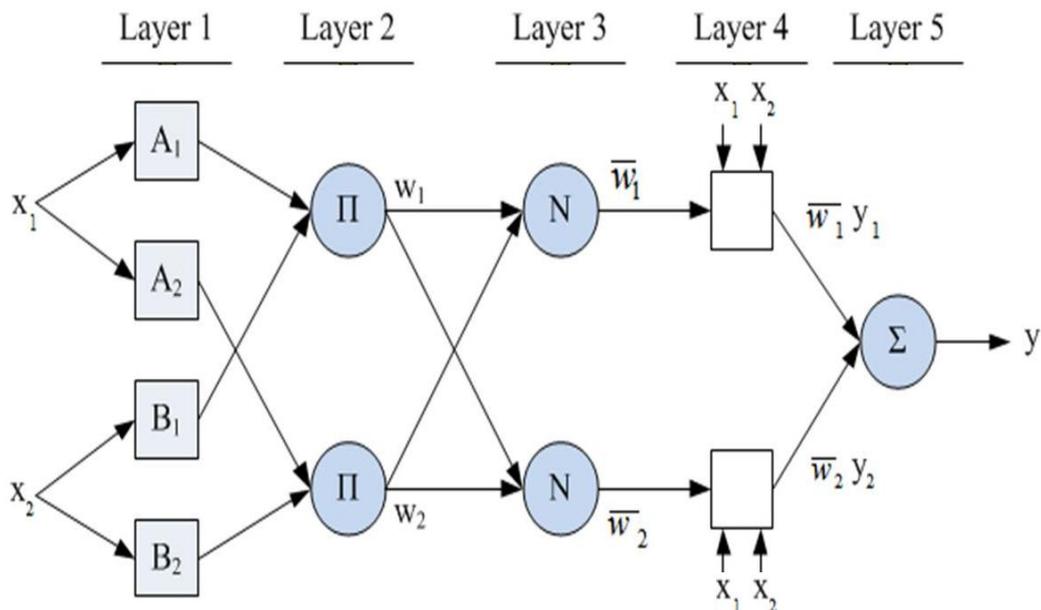


Figure 2.7: Adaptive neuro-fuzzy inference system architecture

The first layer is called a fuzzification layer, which maps the grade of membership for each input. The second layer evaluates the firing strength of each rule on the basis of membership grades. In the third layer a normalized firing strength is calculated. In the fourth layer, the output of each node is calculated. In the fifth layer, the overall output is calculated by summation of all incoming signals.

2.8 SUMMARY

This chapter provided a background of the intelligent systems techniques (IST) that were used in this thesis for data mining, clustering, and pattern recognition. It covered both theoretical and practical aspects of artificial neural networks (ANN), self-organising maps (SOM), fuzzy logic (FL), clustering analysis and adaptive neuro fuzzy inference systems (ANFIS).

REFERENCES

- [1] Zadeh, Lotfi A. "Roles of soft computing and fuzzy logic in the conception, design and deployment of information/intelligent systems." In Computational intelligence: soft computing and fuzzy-neuro integration with applications, pp. 1-9. Springer Berlin Heidelberg, (1998).
- [2] Engelbrecht, Andries P. Computational intelligence: an introduction. John Wiley & Sons, (2007).
- [3] Hand, David J., Heikki Mannila, and Padhraic Smyth. "Principles of data mining (adaptive computation and machine learning)." (2001).
- [4] Jain, Anil K., Jianchang Mao, and K. Moidin Mohiuddin. "Artificial neural networks: A tutorial." IEEE computer 29, no. 3 (1996): 31-44.
- [5] Berthold, Michael R., and David J. Hand, eds. Intelligent data analysis: an introduction. Springer, (2003).
- [6] McCord-Nelson, Marilyn, and William T. Illingworth. A practical guide to neural nets. Addison-Wesley Longman Publishing Co., Inc., (1991).
- [7] Barlow, Horace B. "Unsupervised learning." Neural computation 1, no. 3 (1989): 295-311.
- [8] Sutton, Richard S., and Andrew G. Barto. "Reinforcement learning: an introduction." Neural Networks, IEEE Transactions on 9, no. 5 (1998): 1054-1054.

- [9] T. Kohonen. "Self-organised formation of topologically correct feature maps." *Biological Cybernetics* 43, (1982): 59-69.
- [10] Smith, Andrew James. "Applications of the self-organising map to reinforcement learning." *Neural Networks* 15, no. 8 (2002): 1107-1124.
- [11] Ultsch, Alfred. *U*-matrix: a tool to visualize clusters in high dimensional data*. Fachbereich Mathematik und Informatik, (2003).
- [12] Ultsch, Alfred, and H. Peter Siemon. "Kohonen's Self Organizing Feature Maps for Exploratory Data Analysis." (1990): 305-308.
- [13] Gutierrez-Osuna, Ricardo. "Pattern analysis for machine olfaction: a review." *Sensors Journal, IEEE* 2, no. 3 (2002): 189-202.
- [14] Jain, Anil K., M. Narasimha Murty, and Patrick J. Flynn. "Data clustering: a review." *ACM computing surveys (CSUR)* 31, no. 3 (1999): 264-323.
- [15] Kriegel, Hans - Peter, Peer Kröger, Jörg Sander, and Arthur Zimek. "Density - based clustering." *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 1, no. 3 (2011): 231-240.
- [16] Black, P. E. "Greedy algorithm. *Dictionary of Algorithms and Data Structures* (online)," US National Institute of Standards and Technology, February (2005).
- [17] Zadeh, Lotfi A. "Fuzzy sets." *Information and control* 8, no. 3 (1965): 338-353.

- [18] Perfilieva, Irina, and Jiří Močkoř. "Mathematical principles of fuzzy logic." Springer, (1999).
- [19] Ross, Timothy J. Fuzzy logic with engineering applications. John Wiley & Sons, (2009).
- [20] Zadeh, Lotfi Asker. Fuzzy sets, fuzzy logic, and fuzzy systems: selected papers. Vol. 6. World Scientific, (1996).
- [21] Zadeh, Lotfi A. "Fuzzy sets." Information and control 8, no. 3 (1965): 338-353.
- [22] Mamdani, Ebrahim H., and Sedrak Assilian. "An experiment in linguistic synthesis with a fuzzy logic controller." International journal of man-machine studies 7, no. 1 (1975): 1-13.
- [23] Sugeno, "Michio. Industrial applications of fuzzy control." Elsevier Science Inc., (1985).
- [24] Medsker, Larry R., and Lotfi A. Zadeh. "Hybrid intelligent systems." Dordrecht: Kluwer Academic Publishers, (1995).
- [25] Abraham, Ajith, and Baikunth Nath. "Hybrid intelligent systems design: A review of a decade of research." IEEE Transactions on Systems, Man and Cybernetics (Part C) 3, no. 1 (2000): 1-37.
- [26] Jang, J-SR. "ANFIS: adaptive-network-based fuzzy inference system." Systems, Man and Cybernetics, IEEE Transactions on 23, no. 3 (1993): 665-685.

3

THE HUMAN SPINE ANATOMY, PROBLEMS AND DIAGNOSIS

3.1 HUMAN SPINE

3.1.1 ANATOMY OF HUMAN SPINE

3.1.2 AGE RELATED DEGENERATIVE CHANGES

3.1.3 SOME COMMON PATHOLOGIES OF SPINE

3.2 BACK PAIN PROBLEM

3.2.1 AGEING POPULATION AND BACK PAIN

3.2.2 KEY FACTS AND CAUSES OF BACK PAIN

3.2.3 ECONOMIC AND SOCIAL COST

3.3 DIAGNOSTIC IMAGING

3.3.1 LUMBAR SPINE MAGNETIC RESONANCE IMAGING (MRI)

3.3.2 T1 AND T2 WEIGHTED IMAGES

3.3.3 STUDIES ON AGEING SPINE WITH MRI FINDINGS

3.4 DATA SET

3.4.1 DATA ACQUISITION AND PRE-PROCESSING

3.4.2 FEATURE EXTRACTION AND MEASUREMENT

3.4.3 SCORING OF LUMBAR SPINE MRI

3.5 SUMMARY

3.1 HUMAN SPINE

We humans have a fairly amazing backbone structure called the vertebral column or spine. The spine is a strong, flexible column having several bones that runs from the skull to the pelvis. It is a supporting structure that keeps the body upright and helps in bending and twisting the body. The spine is curved in shape [1], which provides flexibility and helps in coping with different kinds of loads. The spine is so flexible that it can bend far enough to form two thirds of a circle. A human spine is a complex unit performing several functions. Some key functions of the spine are given below:

Support and Strength to the Body: The spine provides support and strength to the human body, particularly the heavy bones of the skull. The upper region of the spine provides strength and stability to the body. The lower region allows flexible movements like bending and twisting. The curved nature of spine coupled with the group of muscles, ligaments and tendons helps in distributing the body weight and adaptation to changes like weight gain or pregnancy. In such circumstances, the curves of the spine become more marked by finding the centre of gravity.

Movement: A complex integration of muscles, tendons, ligaments allows the movements such as bending, stretching, rotating and leaning. The top most part of spine allows the movement and rotation of the head and neck. The neck movements [2]-[4] include flexion (40° - 60°), extension (45° - 70°), bending (45°) and rotation (60° - 80°).

Protection of Nerves: Certain nerve impulses control the functions of major body organs without which humans could not function properly. The spinal column provides reliable protection to the nerves and the spinal cord. The structure and placement of the vertebrae and ligaments form a several layers of protection that prevents spinal cord from injury.

Blood Supply: The blood cells are produced in the bone marrow. The spine has 33 bones [5] called vertebrae, which provide bone marrow to produce blood. There are two types of bone marrow; red and yellow [6]. Red bone marrow is responsible for the production of red blood cells, platelets and white blood cells, while yellow bone marrow contains high levels of fat cells and produces some amounts of white blood cells.

Protection of Major Organs: The spine provides a base for the ribs to attach which surround and protect major body organs. 12 pairs of ribs and the 12 thoracic vertebrae join to make a cage called ribcage [7]. The ribcage forms a protection around the heart and lungs.

Impact Absorption: The spine provides a way of absorbing impact by containing intervertebral discs. These discs are located between each pair of vertebrae and they prevent the vertebrae from bumping into each other [8]. They contain a jelly like substance that absorbs forceful motion preventing the impact from being transferred to the next vertebra acting like a shock absorber.

Other Functions: The spine provides a means of connecting the upper and lower body via the sacrum, which connects the spine to the pelvis. New-born babies have

fairly straight spines. The spine develops its characteristic curves until they begin to hold the weight of their head independently [9]. This shows the changing and adaptive nature of human spine.

3.1.1 ANATOMY OF HUMAN SPINE

A human spine consists of bones, joints, ligaments and muscles. There are normally 33 vertebrae in the human spine. 7 of them are in the neck (cervical region), 12 are in the middle back (thoracic region), 5 are in the lower back (lumbar region) and 5 of them are fused to form the sacrum and the 4 coccygeal bones that form the tailbone as shown in figure 3.1 below. However, in very rare cases some people do have more or fewer than 33 vertebrae [10]. The spine contains over 120 muscles, 220 ligaments and over 100 joints. The human spine is shown in the figure 3.1 below.

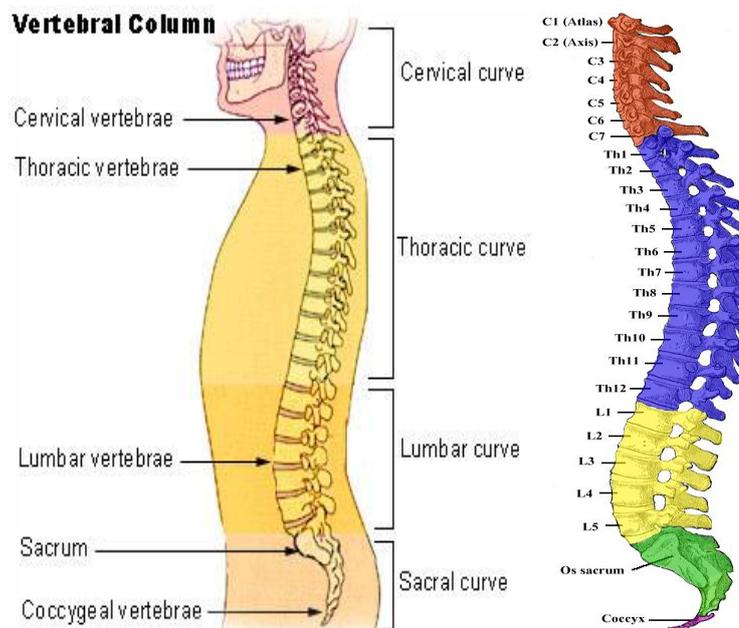


Figure 3.1: Human vertebral column

Lumbar Spine: The lumbar spine refers to the lower back consisting of five vertebrae between the rib cage and the pelvis. The lumbar area is where the spine curves inwards toward the abdomen. The lumbar spine begins roughly 5-6 inches below the shoulder blades. It is connected with the thoracic spine at the top and sacral spine at the bottom [11]. The lower back performs a lot more activity than the rest of the spine and also carries all the weight of the torso, which makes it vulnerable to injuries. In human anatomy, the lumbar spine consists of five vertebrae in the lower half of the spine. They are the largest segments of the vertebral column and are characterized by the absence of the foramen transversarium within the transverse process, and by the absence of facets on the sides of the body [12]. The lumbar vertebrae help to support the body weight by keeping it upright and permitting a wide range of movements. The lumbar spine has 5 intervertebral segments, termed lumbar segment 1 through 5 (as L1, L2, L3, L4 and L5). The anatomy of lumbar spine is shown in figure 3.2 below.

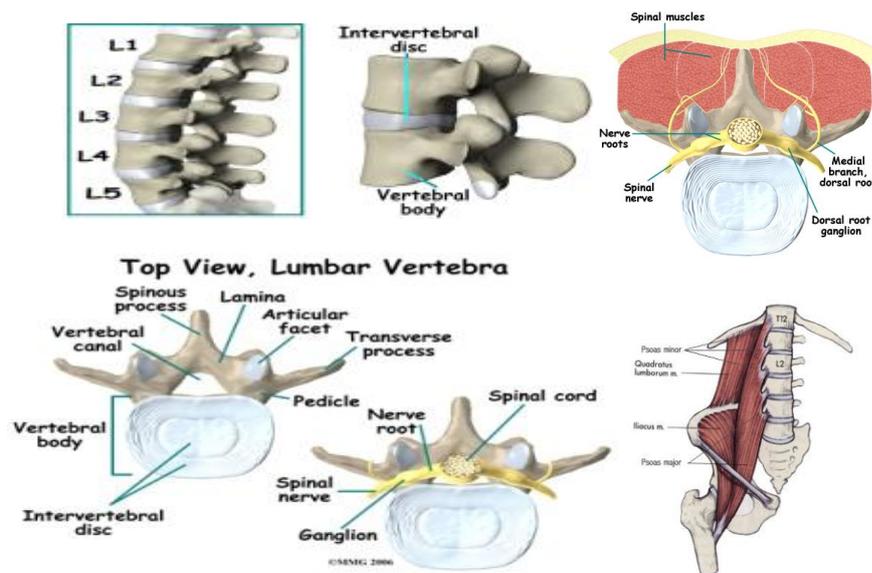


Figure 3.2: Anatomy of lumbar spine

The joints that connect two adjacent vertebrae in the back of the spine are called facet joints. The function of facet joints is to allow movements like bending and twisting in different directions. Between each lumbar vertebra, there is a soft gel-like cushion called an intervertebral disc. These intervertebral discs act like a shock absorbers. They help in absorbing pressure and prevent the vertebrae from rubbing against each other. The muscles attached to the spine are called the para-spinal muscles. They provide support to the spine and act as motors for movement of the spine [13]. A long fusiform muscle called the psoas muscle is located on the side of the lumbar region. It joins the iliacus muscle to form the iliopsoas [13]. Hence, the joints allow flexibility and muscles allow mobility.

3.1.2 AGE RELATED DEGENERATIVE CHANGES

During the normal ageing process, the spine undergoes progressive and regressive changes, which presumably follow some kind of pattern. There is a need to uncover the growth and degeneration pattern of the human spine with reference to the normal ageing process for better understanding of diseases prevalence. There has been no comprehensive in-depth analysis of age-related spinal changes in humans to date. There is no such morphological framework available for classifying age-related spinal changes that can be used as a reference base for finding the causative factors of spinal disorders. Spine specialists (including radiologists and orthopaedics) use their expert knowledge to identify spine disorders relative to the patient age and opinion may vary from one spine specialist to another. There is a need to set standards for growth and degeneration of spinal features in relation to the human ageing process.

In this research, the appearance and variation of several key spine structures are reported in detail. In general, when the appearances of these structures vary with age they are often described as degenerative or age related changes (ARC's). Several questions remain unanswered: what are the norms of the heights of the intervertebral discs with respect to the age? What is the effect of ageing on the heights of the discs? Similarly what are the effects of ageing on vertebrae, para spinal muscles and the cerebrospinal fluid (CSF)? Is there any difference between growth and degeneration of male and female spine? This research answers these questions with the help of different statistical and artificial intelligence techniques. For example, self-organizing map and clustering were used to visualize the natural grouping in the data set and quantify the norms of different spinal features among different age groups. Intelligent techniques helped in forming more refined groups (e.g. group with age 23 to 37 year) according to certain criteria than by simply grouping the subjects according to age decades (20s,30, etc.) and taking average values of all variables in those groups. Also, the features of lumbar spine are interconnected (i.e., most of them effect each other or move together), in this case intelligent techniques are appropriate choice of selection than conventional statistical techniques. Similarly, principal component analysis and factor analysis were used to study the age related variations in male and female spines and to group those variables that are somehow interconnected.

3.1.3 SOME COMMON PATHOLOGIES OF THE SPINE

An illustration of the pathology of an ageing spine is given in [14]. Some of the common pathologies of the spine are given below:

Spondylolisthesis: is the anterior or posterior displacement or slippage of a vertebra in relation to the vertebrae below [15]. It occurs mostly in the lumbar area. It may develop in childhood due to the weakened area of the vertebra or adolescent years due to some injury or later in life as a result of degeneration.

Spinal Stenosis: is an abnormal narrowing of the spinal canal that may occur in any of the regions of the spine [16]. Spinal stenosis may be congenital but most often is attributed to the normal ageing process. This narrowing causes a restriction to the spinal canal, resulting in a neurological deficit.

Sciatica: is a name given to pain radiating from the lower back which travels down the leg [17]. This pain is caused by irritation or compression of the sciatic nerve (the longest nerve in the body).

Disc Herniation: A herniated disc, also known as a prolapsed or slipped disc is a condition where one of the discs in the spine ruptures and the gel inside leaks out causing pain and pressure [18]. Disc herniation is usually due to age related degeneration of the annulus fibrosus, although trauma, lifting injuries, or straining can also cause disc herniation.

Radiculopathy: refers to a set of conditions in which one or more nerves is affected and does not work properly [19]. It is often referred to as a "pinched nerve" which may cause pain, weakness and numbness and controlling specific muscles.

Spinal Instability: is a term used to describe abnormal movement between two vertebrae. Spinal instability is inability of the spinal column to maintain its normal configuration under normal usage conditions [20]. It occurs as a disc degenerates

and flattens, vertebrae become unstable and slip back and forth, irritating facet joints and nerves.

Degenerative Disc Disease: With normal ageing, the discs naturally wear out. Over the decades, the repeated daily stresses on the spine and occasional minor unnoticed injuries as well as major ones begin to take their toll [21]. The damaged disc can cause pain.

Arthritis: is a form of joint disorder that involves inflammation of one or more joints. Arthritis in the spine is caused as vertebrae and discs age and wear out, allowing formation of bone spurs [22]. Arthritis can cause or worsen spinal stenosis and may irritate nerve roots.

Strains and Sprains: are a common type of injury that affects muscles and ligaments. Sprains and strains are caused by improper lifting, twisting, bending, falling, or other injuries, such as sports injuries [23.] Strains and sprains usually involve irritation of muscles and ligaments around the spine.

3.2 BACK PAIN PROBLEM

Back pain is usually associated with the lumbar spine in most cases. We can identify an anatomical reason for the pain, such as a lumbar disc herniation or stenosis (reduction in the circumference of the spinal canal) in only 20% of the patients. There might be reasons for back pain that cannot be identified on visual analysis magnetic resonance imaging (MRI), for example muscular pain.. The exact reasons for pain in other areas like the knee or hip pain are far easier to diagnose. The low percentage for back pain diagnosis might be due to the fact that the back is a much

more complex and complicated structure. Some very small changes in spinal features with the ageing process could lead to back pain, which can be easily overlooked by inspecting an MRI using the naked eye. Therefore, there was always a need to develop a systematic method that can automatically seek and evaluate the tiny changes in spinal features in order to find the reasons for back pain.

3.2.1 AGEING POPULATION AND BACK PAIN

The ageing of the population in developed countries appears to be a non-reversible phenomenon. In Europe, the proportion of subjects over 65 was roughly 11% in 1950, became 14% in 1970, 19% in 1995 and is projected by some sources at 30% in 2025 and 42% in 2050. The proportion of subjects over 75 was 2.7% in 1950 and is projected at about 9% by 2025 and 14.6% by 2050 [24]. When only Western Europe is considered, the individuals over 65 are estimated to be doubled by 2050. These numbers are almost same in the USA [25]. With the ageing population, an increased number of musculoskeletal problem are seen. Back and neck pain are among the most frequently encountered complaints of older people. The complex nature of the spine makes those problems difficult to investigate and treat. The age related changes gradually affect almost all structures of the spinal units. Degeneration of the spinal structures produces alterations at many levels such as bone, disc, joints, ligaments, and muscles.

3.2.2 KEY FACTS AND CAUSES OF BACK PAIN

The incidence of significant pain does not only affect the patient but often the family. A study on Pain in Europe [26] interviewed 46,000 people in 16 countries,

and represents the largest and most in-depth long term chronic pain survey ever conducted in Europe. According to this survey, nearly one in five Europeans suffers chronic pain for an average of seven years. One in six chronic pain sufferers feel the pain is sometimes so bad they wanted to die. Seventy percent are treated by their primary care doctor. The median length of time with their current doctor was 4.5 years. Chronic pain strikes one in five (19%) adults across Europe. Over one third of European households have at least one pain sufferer and the most common source of pain reported by chronic pain sufferers is the back (24%). Back pain is usually associated with the spine disorder. Spraining of ligaments, straining of muscles, rupture disks, irritated joints, accidents and sports injuries, all of these can lead to back pain. In addition to that, arthritis, poor posture, prolonged sitting, obesity and psychological stress can cause or complicate back pain. Some diseases of the internal organs like kidney stones, kidney infections, blood clots or bone loss can also originate back pain [27]. The big challenge for the clinicians is to properly pinpoint the area which originates pain. MRI of the spine provides a very good visualization of the spine but does not always pinpoint the area where the pain is coming from, especially when the pain is related to muscles.

Back pain is the second most common reason for visits to the doctor's clinic, outnumbered only by the upper-respiratory infections [28]-[30]. Back pain is one of the most common reasons for missed work. One-half of all working Americans admit to having back pain symptoms each year [29], [31]. The evidence from Britain shows that back pain is becoming a bigger problem over time [32]-[34]. The total number of days lost in Britain for back incapacity obtained through sickness and

invalidity benefit has risen dramatically in recent years [35], [36]. Britain along with other developed countries is seeing a steeper rise in back pain sufferers [37]-[39]. For the back pain patients; untreated back pain effects on the quality of life, impacting on their work and often causing depression. One in five chronic pain sufferers have lost a job as a result of their pain. One in five chronic pain sufferers have been diagnosed with depression as a result of their pain. Over two thirds of sufferers believe their medication is not sufficient to control their pain at times [26]. Figure 3.3 gives the number of back pain patients in Scotland consulting a GP or practice nurse at least once in the financial year 2012/13.

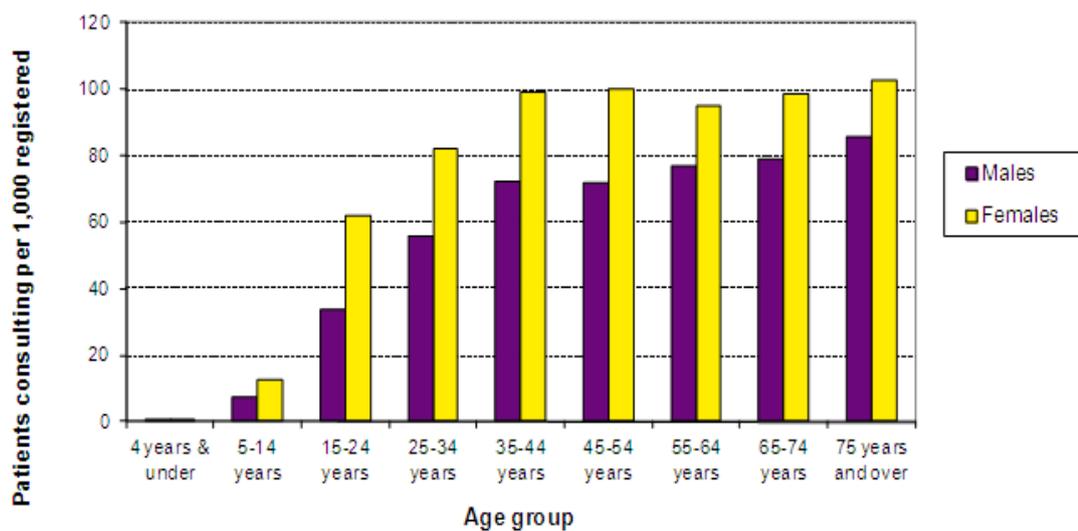


Figure 3.3: Number of back pain patients in Scotland consulting a GP or practice nurse at least once in the financial year 2012-13 per 1,000 (Statistics by NHS Scotland)

Lower back pain is an extremely common problem, which most people experience at some point in their life. The statistics from industrialized countries shows that up to four out of every five adults will experience back pain at some stage of their life. The incidence of low back pain is highest in the third decade, and the overall

prevalence increases with age until the 60–65 year age group and then gradually declines [40]. Another study shows that back pain is just as common in adolescents as in adults [41]. But with the progressing age, it is more likely to have a back pain. This could be a single or multiple episodes of back pain. For some people it becomes chronic and is intermittent. In most cases it is very difficult to identify a single cause for back pain. In about 85% of back pain sufferers no clear pathology can be identified [42]. The factors like having had back pain in the past, smoking and obesity could contribute to back pain [43]. Some physical factors such as heavy physical work, frequent bending, twisting, lifting, pulling and pushing, repetitive work, static posture and vibrations can lead to back pain [44]. Psychosocial factors such as work stress, anxiety, depression, job dissatisfaction and mental stress etc. can also lead to back pain [45].

3.2.3 ECONOMIC AND SOCIAL COST

In the United States, it is estimated that the annual direct cost of lower back pain is \$12.2 to \$90.6 billion and the total (direct and indirect) cost is estimated to be between \$84.1 and \$624.8 billion. Thus the total annual cost of lower back pain in the US is estimated to be over 100 billion dollars [46]. The National Health Service spends more than £1 billion per year on back pain related costs. This includes £512 million on hospital costs for back pain patients, £141 million on GP consultations for back pain and £150.6 million on physiotherapy treatments for back pain. In the private healthcare sector £565 million is spent on back pain every year. This brings the healthcare costs for back pain to a total of £1.6 billion per year [47]. In addition there are other indirect costs. The Health and Safety Executive estimates that

musculoskeletal disorders, which include back pain cost UK employers between £590 million and £624 million per year [48]. The total cost of back pain corresponds to between 1% and 2% of gross national product (GDP). The charity BackCare estimates that back pain costs the NHS, business and the economy easily over a billion pound per year. Other European countries report similar high costs; back pain related costs in The Netherlands were more than 4 billion euro. For Sweden, these were more than 2 billion euro [49].

Nearly 5 million working days were lost as a result of back pain in 2003-04. This means that on any one day 1% of the working population is on sickness leave due to a back problem. Each person suffering from such a condition took an estimated 17.4 days off work on average in this period [50]. Back pain is the number two reason for long term sickness in much of the UK. In manual labour jobs, back pain is the number one reason [51].

3.3 DIAGNOSTIC IMAGING

Spine entertains tremendous forces in our daily routine showing that it is very resilient to both injury and degeneration. But some sudden impact, accident, injury or prolonged diseases can give rise to early degeneration of spine. There are number of problems and diseases which are associated with the lumbar spine. Some of them are described in the previous section. Due to the advancement in medical imaging, radiologists and spine specialists are now able to gain a very good insight into spinal features and can pinpoint the problematic areas. Typically, X-rays and medical resonance imaging (MRI) are used to look at the patient's spine. Magnetic Resonance Imaging (MRI scan) was developed in the 1980's and since

then it has revolutionized the ability to see normal and abnormal spinal structures and helps in diagnosing the causes of back pain. The MRI scan is a diagnostic imaging test that allows physicians to assess a patient's spinal anatomy and investigate an anatomical reason of back pain. The physician correlates the findings on the MRI scan with the patient's signs and symptoms of back pain in order to arrive at a clinical diagnosis and decision making.

3.3.1 *LUMBAR SPINE MAGNETIC RESONANCE IMAGING*

In the developed world, the use of MR imaging of the lumbar spine for the diagnosis and management of lower back pain has become standard practice. With the advent of picture archiving and communication system (PACS) images, it has become easier to acquire more accurate data of ARC's. Recommendations have also been published for more accurate continuous data of these changes to get a better idea of their correlation to clinical symptoms and age [52], [53]. As clinical symptoms do not have a clear correlation to ARC's, so they were not taken into account in this research. Figure 3.4 shows sagittal and an axial view of the lumbar spine MRI.

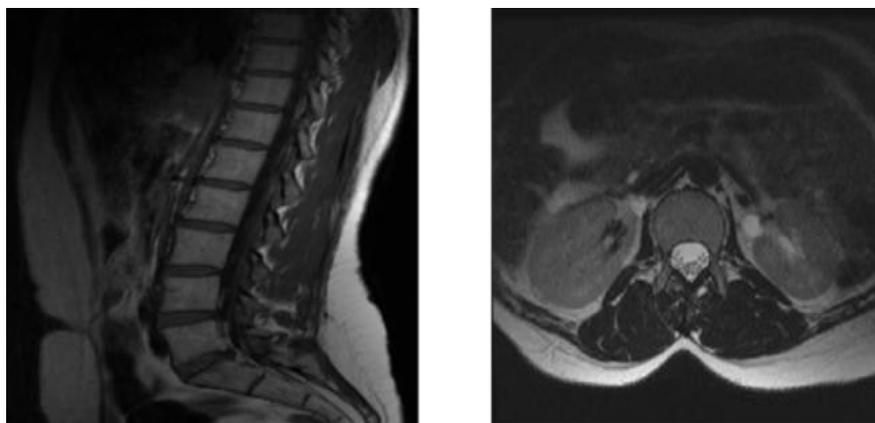


Fig 3.4: Sagittal (left) and an axial (right) view of the lumbar spine MRI

3.3.2 T1 AND T2 WEIGHTED IMAGES

When looking at an MRI scan, some areas appear to be brighter or darker than other ones. This variation is due to the density of protons in that area where an increased density being associated with a darker area. The relaxation times for protons can vary for which usually two terms as T1 and T2 are used for the measurement. T1 is the longitudinal relaxation time and T2 is the transverse relaxation time [54]. White matter is darker than grey matter in T1-weighted images and brighter than grey matter in T2-weighted images. T1 and T2 weighted MRIs are shown in figure 3.5.



Figure 3.5: T1 and T2 weighted lumbar MRI

The main difference between T1 and T2 weighted MRI images is the density of fluids. Water appears white on T2 weighted images whereas in T1 weighted images it appears dark. These two types of MRI scan help in differentiation of abnormalities in soft tissue injury and solid organ. For example, injury resulting in surrounding

edema looks white on T2 weighted images; the reverse is true in the case of T1 weighted images. Usually, T1 used to look for solid organ pathology e.g. liver, spleen whereas T2 is used to look for soft tissue injury e.g. in para vertebral muscles in back injury.

3.3.3 STUDIES ON AGE RELATED CHANGES IN HUMAN SPINE WITH THE HELP OF MRI FINDINGS

Mechanical lower back pain in the absence of pathology such as fracture, tumour or infection is commonly thought to come from the intervertebral discs, the facet joints, tendinous or ligamentous insertions of the lumbar spine. However a clear contradiction has been seen among the studies carried out on changing anatomy of ageing spine and its correlation with back pain and other spinal disease. Some studies have confirmed a correlation of spinal variations with age [55]-[57] whereas some others have failed to do so [58], [59]. In some studies, when looking at specific aspects of the scan, there seems to be an extremely strong correlation with age [60], [61]. A handful of studies are carried out to understand the ageing effect on the lumbar spine. Most of them study the behaviour of a specific feature independently. But due to the complexity of the spine, many of the spinal features are interconnected and often move together. So looking at the one feature may not always give an accurate picture. Most studies have tried to correlate these changes to clinical symptoms with few succeeding [62].

There is a general agreement that changes induced by ageing lead to alterations in the thickness of the disc and muscles [63], but there are differences in the accounts of the effect of ageing on the thickness of the lumbar discs. One study stressed that

reduction of the intervertebral disc height with age is inevitable [64]. In contrast, an increase in disc height with age has been reported by other study [65]. One study suggests that disc degeneration is highly correlated with age and educational level [66]. Similarly, many others have tried to link spinal ageing with environment and genetics. According to another study, the only degenerative feature associated with self-reported lower back pain was spinal stenosis [67]. Some studies suggest that there is a significant difference in pattern of vertebral growth in male and female [68], [69]. Hence, the variation in the incidence of ARC's varies widely between the studies. The method of assessing ARC's also varies widely from ordinal observer data to continuous, computer generated data. The number and type of ARC's varies but are often presented to mean the same in regards the amount of degeneration seen on the lumbar spine MRI.

3.4 DATA SET

The data set used in this pilot study was taken from University Hospital Coventry and Warwickshire NHS Trust, Coventry, United Kingdom in the form of patient MRIs. All the necessary ethical approvals were obtained from the concerned bodies against the use of data for research purposes. The ethical approval letter is attached at the end of the thesis. As each patient has number of scans for the same area, the best samples were selected from the series of scan. Images were obtained from GE Scanners (GE healthcare Milwaukee) at 1.5T using standard imaging protocol. All the subjects were scanned in supine position. Images were visualized and scored using the PACS web browser (GE). The format of data is Digital Imaging and Communications in Medicine (DICOM). An attempt was made to pick those MR

scans that do not exhibit any clearly visible lumbar spine deformity. It was also made sure that the samples do not have any spinal implants. However no assessment was made into the patients clinical symptoms.

3.4.1 DATA ACQUISITION AND PRE-PROCESSING

Lumbar spine MR scans of 61 patients were selected to develop an initial model. There were 35 female samples and 26 male samples in total. Age and gender distribution of samples are shown in the figure3.6 below. Ten groups were formed on the basis of age decades as: G0: new born to 10 years, G1: 11-20 years, G2: 21-30, years, G3: 31-40 years, G4: 41-50 years, G5: 51-60 years, G6: 61-70 years, G7: 71-80 years, G8: 81-90 years and G9: 91 and above years of age). The age of samples ranges from 2 to 93 years.

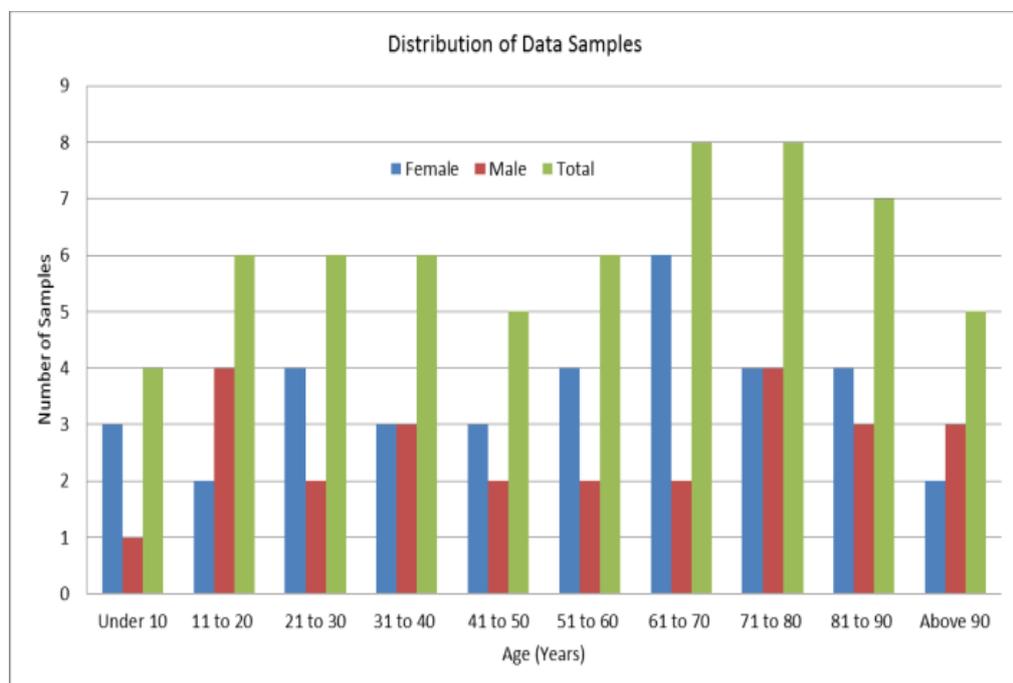


Figure 3.6: Distribution of data samples

3.4.2 FEATURE EXTRACTION AND MEASUREMENT

Although there are tools and procedures available for automated segmentation and feature extraction [70], [71] in this pilot study, features were measured manually by spine specialist to minimize or eliminate any chances of machine error. These features were measured from MR images using digital measuring tools and were recorded against lumbar spine MRIs of 61 patients. All the measurements were taken twice to get a good approximate. Roughly 5-6 patients were selected from each age decade for the pilot study. Only 4 lumbar spine MRIs were available for age under 10 years. Vertebral and disc heights were measured from the centre of vertebrae and discs.

3.4.3 SCORING OF LUMBAR SPINE MRI

There are many features that can be studied from a lumbar spine MR image. After a thorough literature study, a list of significant lumbar spine features was prepared by the orthopaedic spinal surgeon and radiologist involved in this research. The scoring criteria were set to look initially at the vertebral heights (L1, L2, L3, L4 and L5), disc heights (T12-L1, L1-L2, L2-L3, L3-L4, L4-L5 and L5-S1), disc signal intensities (T12-L1, L1-L2, L2-L3, L3-L4, L4-L5 and L5-S1), para-spinal muscle signal intensities measured at L3 (both left and right), subcutaneous fat on left and right, psoas muscle (both left and right) and cerebrospinal fluid (CSF). Vertebral and disc heights were measured in millimetres (mm) whereas all other features were measured in terms of their signal intensities. The signal intensities are the average value of the selected region of interest. All the models and experiments presented in the forthcoming chapters were built and conducted using these 24 lumbar spine

features. Table 3.1 below shows five random samples (of age 2, 29, 49, 68, and 90 years) scored on the basis of their spinal characteristics.

Table 3.1 Five randomly selected samples with their lumbar spine MRI scores

Gender	m/f	f	m	f	f	m
Age	Years	90	68	49	29	2
Vertebral Height (mm)	L1	21.44	25.41	22.35	21.87	12.6
	L2	20.39	26.14	22.2	21.37	12.93
	L3	20.16	27.95	22.62	21.65	13.55
	L4	21.41	25.58	23.08	20.75	13.46
	L5	21.59	24.96	23.05	18.91	13.72
Disc Height (mm)	T12 L1	7.03	8.87	6.95	6.91	4.33
	L1 L2	7.82	9.05	8.62	9.18	4.69
	L2 L3	8.8	9.68	9.55	9.05	5.8
	L3 L4	9.36	6.24	10.07	8.82	5.75
	L4 L5	6.89	11.4	5.34	10.38	6.45
	L5 S1	4.17	8.36	10.71	10.89	5.68
Disc Signal (average intensity)	T12 L1	54.6	59.7	47.7	145	358.3
	L1 L2	44.2	47.4	58	133.3	359.2
	L2 L3	39.9	56.7	70.3	136.2	318
	L3 L4	49.8	36.7	57.2	138.4	327.7
	L4 L5	54.7	50	48.8	149.4	368.6
	L5 S1	71.6	44.9	102.2	126.9	328.6
Para Spinal Muscles (average intensity)	Right	258.9	194.3	109.9	98.1	80.1
	Left	252.9	196.7	111	81.1	77
Psoas Muscle (average intensity)	Right	109.2	77.5	64.6	44.7	63.7
	Left	94.6	75.4	56.5	44	57.5
Subcutaneous Fat	Right	407	749.2	772.5	734.5	477.1

(average intensity)	Left	718.5	537.5	671.4	528.5	298
CSF (average intensity)	at L3	392.3	367.6	350.8	445.4	655.1

3.5 SUMMARY

This chapter provided an insight of the key features and anatomy of a human spine. It explained various problems associated with the human lumbar spine. It also discussed the age related degenerative changes seen on lumbar spine in the light of previous studies conducted in this area. It gave a comprehensive description of MRI data set, feature extraction, feature measurement and the scoring criteria for spinal MRIs (selection of the most noticeable features from raw data).

REFERENCES

- [1] Dorland, William Alexander Newman. "Dorland's Illustrated Medical Dictionary." Elsevier Health Sciences, (2011).
- [2] Youdas, James W., Tom R. Garrett, Vera J. Suman, Connie L. Bogard, Horace O. Hallman, and James R. Carey. "Normal range of motion of the cervical spine: an initial goniometric study." *Physical Therapy* 72, no. 11 (1992): 770-780.
- [3] Penning, L. "Normal movements of the cervical spine." *American Journal of Roentgenology* 130, no. 2 (1978): 317-326.
- [4] Range of Joint Motion Evaluation Chart. Washington State, USA: Department of Social and Health Services; June (2013); at online http://www.dshs.wa.gov/pdf/ms/forms/13_585a.pdf
- [5] Parke, Wesley W. "Applied anatomy of the spine." *The spine* 1 (1992): 35-87.
- [6] Wickramasinghe, Sunitha Nimal. *Human bone marrow*. Oxford London Edinburgh Melbourne: Blackwell Scientific Publications, (1975).
- [7] Oda, Itaru, Kuniyoshi Abumi, Duosai Lü, Yasuhiro Shono, and Kiyoshi Kaneda. "Biomechanical role of the posterior elements, costovertebral joints, and rib cage in the stability of the thoracic spine." *Spine* 21, no. 12 (1996): 1423-1429.
- [8] Roberts, Sally, Helena Evans, Jayesh Trivedi, and Janis Menage. "Histology and pathology of the human intervertebral disc." *The Journal of Bone & Joint Surgery* 88, no. suppl_2 (2006): 10-14.

- [9] Mac-Thiong, Jean-Marc, Éric Berthonnaud, John R. Dimar, Randal R. Betz, and Hubert Labelle. "Sagittal alignment of the spine and pelvis during growth." *Spine* 29, no. 15 (2004): 1642-1647.
- [10] Gray, Henry, "*Gray's Anatomy*," New York: Crown Publishers, Inc. p. 34, (1977).
- [11] McKenzie, Robin A., and Stephen May. "The lumbar spine." *Mechanical diagnosis & therapy* 2 (1981): 731.
- [12] Bogduk, Nikolai. *Clinical anatomy of the lumbar spine and sacrum*. Churchill Livingstone, (2005).
- [13] Bogduk, N., "*Clinical anatomy of the lumbar spine and sacrum*," Elsevier Health Sciences, (2005).
- [14] Prescher, Andreas. "Anatomy and pathology of the aging spine." *European journal of radiology* 27, no. 3 (1998): 181-195.
- [15] Fredrickson, Bruce E., Daniel Baker, W. J. McHolick, H. A. Yuan, and J. P. Lubicky. "The natural history of spondylolysis and spondylolisthesis." *The Journal of Bone & Joint Surgery* 66, no. 5 (1984): 699-707.
- [16] Postacchini, Franco, and Wolfgang Rauschnig. *Lumbar spinal stenosis*. Vienna: Springer-Verlag, (1989).
- [17] Frymoyer, John W. "Back pain and sciatica." *The New England journal of medicine* 318, no. 5 (1988): 291.
- [18] Cyriax, James Henry. *The slipped disc*. Gower P., (1970).
- [19] Eck, Jason C. "Radiculopathy" *MedicineNet.com*. Retrieved 10 May (2012).
- [20] Panjabi, Manohar M. "Clinical spinal instability and low back pain." *Journal of electromyography and kinesiology* 13, no. 4 (2003): 371-379.

- [21] Szpalski, M., Andersson, G. B. (Eds.), "Degenerative disc disease". Lippincott Williams & Wilkins, (2004).
- [22] Koopman, William J; Moreland, Larry W. "Arthritis and allied conditions: a textbook of rheumatology." Philadelphia: Lippincott Williams & Wilkins, (2005).
- [23] Buka, Alfred J. "Strains and sprains." The American Journal of Surgery 12, no. 2 (1931): 290-293.
- [24] World Population Program, Population Research at International Institute for Applied Systems Analysis (IIASA), Luxemburg, Austria.
<http://www.iiasa.ac.at/web/home/research/Population.en.html>
- [25] Vincent, Grayson K., and Victoria Averil Velkoff. The next four decades: The older population in the United States: 2010 to 2050. No. 1138. US Department of Commerce, Economics and Statistics Administration, US Census Bureau, (2010).
- [26] Fricker, J. "Pain in Europe- A 2003 report." Mundipharma International Ltd, Cambridge. Retrieved from <http://www.britishpainsociety.org>
- [27] Hestbaek, Lise, Charlotte Leboeuf-Yde, Marianne Engberg, Torsten Lauritzen, Niels Henrik Bruun, and Claus Manniche. "The course of low back pain in a general population. Results from a 5-year prospective study." Journal of manipulative and physiological therapeutics 26, no. 4 (2003): 213-219.
- [28] Vällfors, Birgitta. "Acute, subacute and chronic low back pain: clinical symptoms, absenteeism and working environment." Scandinavian journal of rehabilitation medicine. Supplement 11 (1984): 1-98.

- [29] Health and safety statistics - A National Statistics publication, "Back Health at Work." Health and Safety Commission (HSE) 2005.
- [30] Deyo, R. A. "Back Pain Patient Outcomes Assessment Team (BOAT)." US Department of Health & Human Services-Agency of Healthcare Research (1994).
- [31] Department of Health Statistics Division. The prevalence of back pain in Great Britain in 1998. London: Government Statistical Service, (1999).
- [32] Palmer, Keith T., Kevin Walsh, Holly Bendall, Cyrus Cooper, and David Coggon. "Back pain in Britain: comparison of two prevalence surveys at an interval of 10 years." *Bmj* 320, no. 7249 (2000): 1577-1578.
- [33] Burton, A. Kim, Federico Balagué, Greet Cardon, Hege R. Eriksen, Yves Henrotin, Amnon Lahad, Annette Leclerc, Gert Müller, and Allard J. Van Der Beek. "Chapter 2 European guidelines for prevention in low back pain." *European Spine Journal* 15 (2006): 136-168.
- [34] Nachemson, A.L., G. Waddell, A. L. Norlund, "Neck and back pain: The scientific evidence of causes, diagnosis and treatment," Lippincott Williams and Wilkins, (2000): 165-188.
- [35] Maniadakis, Nikolaos, and Alastair Gray. "The economic burden of back pain in the UK." *Pain* 84, no. 1 (2000): 95-103.
- [36] Andersson, G. B. J., "The epidemiology of spinal disorders," in *The adult spine: Principles and practice*, Philadelphia: Lippincott-Raven, (1997).
- [37] Hoogendoorn, Wilhelmina E., Mireille NM van Poppel, Paulien M. Bongers, Bart W. Koes, and Lex M. Bouter. "Systematic review of psychosocial factors

- at work and private life as risk factors for back pain." *Spine* 25, no. 16 (2000): 2114-2125.
- [38] Gordon, F., and D. Risley. "The costs to Britain of workplace accidents and work-related ill health in 1995/6." London: Health and Safety Executive (1999): 1-128.
- [39] Norlund, A.I., G. Waddell, "Cost of back pain in some OECD countries," *Neck and back pain: the scientific evidence of causes, diagnosis and treatment*, no. 1 (2000): 421-425.
- [40] Hoy, D., P. Brooks, F. Blyth, and R. Buchbinder. "The epidemiology of low back pain." *Best Practice & Research Clinical Rheumatology* 24, no. 6 (2010): 769-781.
- [41] Burton, A. Kim, Federico Balagué, Greet Cardon, Hege R. Eriksen, Yves Henrotin, Amnon Lahad, Annette Leclerc, Gert Müller, and Allard J. Van Der Beek. "Chapter 2 European guidelines for prevention in low back pain." *European Spine Journal* 15 (2006): 136-168.
- [42] Nachemson, All, G. Waddell, and A. L. Norlund. "Epidemiology of neck and low back pain." *Neck and Back Pain: The scientific evidence of causes, diagnosis and treatment* (2000): 165-188.
- [43] Burton, A. Kim, Federico Balagué, Greet Cardon, Hege R. Eriksen, Yves Henrotin, Amnon Lahad, Annette Leclerc, Gert Müller, and Allard J. Van Der Beek. "Chapter 2 European guidelines for prevention in low back pain." *European Spine Journal* 15 (2006): 136-168.
- [44] Andersson, G. B. J. "The epidemiology of spinal disorders." *The adult spine: Principles and practice* (1997).

- [45] Hoogendoorn, Wilhelmina E., Mireille NM van Poppel, Paulien M. Bongers, Bart W. Koes, and Lex M. Bouter. "Systematic review of psychosocial factors at work and private life as risk factors for back pain." *Spine* 25, no. 16 (2000): 2114-2125.
- [46] Dagenais, Simon, Jaime Caro, and Scott Haldeman. "A systematic review of low back pain cost of illness studies in the United States and internationally." *The Spine Journal* 8, no. 1 (2008): 8-20.
- [47] Maniadakis, Nikolaos, and Alastair Gray. "The economic burden of back pain in the UK." *Pain* 84, no. 1 (2000): 95-103.
- [48] The costs of accidents at work, Health and Safety Guidelines 1996, 2nd Edition, Her Majesty's Stationery Office (HMSO) (1997).
- [49] Norlund, A. I., and G. Waddell. "Cost of back pain in some OECD countries." *Neck and back pain: the scientific evidence of causes, diagnosis and treatment* 1 (2000): 421-425.
- [50] Waddell, Gordon, Mansel Aylward, and Philip Sawney. *Back pain, incapacity for work and social security benefits: an international literature review and analysis*. RSM Press, (2002).
- [51] Department of Health Statistics Division. "The prevalence of back pain in Great Britain in 1998." London: Government Statistical Service, department of Health, (1999).
- [52] Haughton, Victor. "Imaging intervertebral disc degeneration." *The Journal of Bone & Joint Surgery* 88, no. 2 (2006): 15-20.

- [53] An, Howard S., Paul A. Anderson, Victor M. Haughton, James C. Iatridis, James D. Kang, Jeffrey C. Lotz, Raghu N. Natarajan et al. "Introduction: disc degeneration: summary." *Spine* 29, no. 23 (2004): 2677-2678.
- [54] Dr. Gurvinder Rull, "Magnetic Resonance Imaging", Document ID: 631, Version: 24, Egton Medical Information Systems Limited, Available at: <http://medical.cdn.patient.co.uk/pdf/631.pdf>
- [55] Battié, Michele C., and Tapio Videman. "Lumbar disc degeneration: epidemiology and genetics." *The Journal of Bone & Joint Surgery* 88, no. suppl_2 (2006): 3-9.
- [56] Videman, Tapio, Michele Crites Battié, Kevin Gill, Hannu Manninen, Laura E. Gibbons, and Lloyd D. Fisher. "Magnetic resonance imaging findings and their relationships in the thoracic and lumbar spine: insights into the etiopathogenesis of spinal degeneration." *Spine* 20, no. 8 (1995): 928-935.
- [57] Boos, Norbert, Sabine Weissbach, Helmut Rohrbach, Christoph Weiler, Kevin F. Spratt, and Andreas G. Nerlich. "Classification of age-related changes in lumbar intervertebral discs: 2002 Volvo Award in basic science." *Spine* 27, no. 23 (2002): 2631-2644.
- [58] Sharma, Aseem, Matthew Parsons, and Thomas Pilgram. "Temporal interactions of degenerative changes in individual components of the lumbar intervertebral discs: A sequential magnetic resonance imaging study in patients less than 40 years of age." *Spine* 36, no. 21 (2011): 1794-1800.
- [59] Miller, J. A. A., C. Schmatz, and A. B. Schultz. "Lumbar disc degeneration: correlation with age, sex, and spine level in 600 autopsy specimens." *Spine* 13, no. 2 (1988): 173-178.

- [60] Videman, Tapio, Laura E. Gibbons, and Michele C. Battié. "Age-and pathology-specific measures of disc degeneration." *Spine* 33, no. 25 (2008): 2781-2788.
- [61] Niemeläinen, R., T. Videman, S. S. Dhillon, and M. C. Battié. "Quantitative measurement of intervertebral disc signal using MRI." *Clinical radiology* 63, no. 3 (2008): 252-255.
- [62] Cheung, Kenneth MC, Jaro Karppinen, Danny Chan, Daniel WH Ho, You-Qiang Song, Pak Sham, Kathryn SE Cheah, John CY Leong, and Keith DK Luk. "Prevalence and pattern of lumbar magnetic resonance imaging changes in a population study of one thousand forty-three individuals." *Spine* 34, no. 9 (2009): 934-940.
- [63] Parkkola, R., and M. Kormano. "Lumbar disc and back muscle degeneration on MRI: correlation to age and body mass." *Journal of Spinal Disorders & Techniques* 5, no. 1 (1992): 86-92.
- [64] Vernon-Roberts, Barrie. "Disc pathology and disease states." *The biology of the intervertebral disc* 2 (1988): 73-119.
- [65] Shao, Zengwu, Gerhard Rompe, and Marcus Schiltenswolf. "Radiographic changes in the lumbar intervertebral discs and lumbar vertebrae with age." *Spine* 27, no. 3 (2002): 263-268.
- [66] Igbinedion, BO E., and A. Akhigbe. "Correlations of radiographic findings in patients with low back pain." *Nigerian medical journal: journal of the Nigeria Medical Association* 52, no. 1 (2011): 28.
- [67] Kalichman, Leonid, David H. Kim, Ling Li, Ali Guermazi, and David J. Hunter. "Computed tomography–evaluated features of spinal degeneration:

prevalence, intercorrelation, and association with self-reported low back pain." *The spine journal* 10, no. 3 (2010): 200-208.

- [68] Geusens, Piet, Jan Dequeker, Anne Verstraeten, and Jos Nijs. "Age-, sex-, and menopause-related changes of vertebral and peripheral bone: population study using dual and single photon absorptiometry and radiogrammetry." *Journal of nuclear medicine: official publication, Society of Nuclear Medicine* 27, no. 10 (1986): 1540-1549.
- [69] Duan, Yunbo, Charles H. Turner, Bom - Taeck Kim, and Ego Seeman. "Sexual Dimorphism in Vertebral Fragility Is More the Result of Gender Differences in Age - Related Bone Gain Than Bone Loss." *Journal of Bone and Mineral Research* 16, no. 12 (2001): 2267-2275.
- [70] Carballido-Gamio, Julio, Serge J. Belongie, and Sharmila Majumdar. "Normalized cuts in 3-D for spinal MRI segmentation." *Medical Imaging, IEEE Transactions on* 23, no. 1 (2004): 36-44.
- [71] Zheng, Yalin, Mark S. Nixon, and Robert Allen. "Automated segmentation of lumbar vertebrae in digital videofluoroscopic images." *Medical Imaging, IEEE Transactions on* 23, no. 1 (2004): 45-52.

4

FEATURE EXTRACTION AND VISUALIZATION OF MULTIVARIATE LUMBAR SPINE DATA

4.1 PRINCIPAL COMPONENT ANALYSIS (PCA)

4.2 PCA MODELLING

4.2.1 STATISTICAL ANALYSIS AND DATA STANDARDIZATION

4.2.2 DATA VISUALIZATION

4.2.3 CORRELATION OF FEATURES WITH AGE

4.2.4 NON CORRELATED FEATURES

4.3 FACTOR ANALYSIS (FA)

4.4 HIERARCHICAL CLUSTERING

4.4.1 AGGLOMERATIVE CLUSTERING

4.4.2 CLUSTER ANALYSIS

4.5 PCA BASED ANOMALY DETECTION

4.6 SUMMARY

REFERENCES

4.1 PRINCIPAL COMPONENT ANALYSIS

One of the difficulties in multivariate statistics is the problem of visualizing data that has many variables. In datasets with several variables, groups of variables often vary together. One reason for this is that more than one variable might be measuring the same driving principle governing the behaviour of the system [1]. In many systems there are only a few such driving forces. When this happens, one can take advantage of any redundancy in information. This problem can be simplified by replacing a group of variables with a smaller set of new variables [2]. Medical data is usually collected with large number of variables. With a large number of variables, the dispersion matrix becomes too large to study and interpret properly. There would be a number of pair wise correlations between the variables to study. Graphical display of data may also not be of particular help in case the data set is very large. For example, with 12 variables, there will be 220 three-dimensional scatter plots to be studied. It is therefore necessary to reduce the number of variables to a few, interpretable linear combinations of the data for better interpretation and exploratory analysis of data. Each linear combination will correspond to a principal component [3], [4]. The number of principal components is equal to the number of the original set of variables. Often, however, the first few components account for most of the variance in the original data (meaning that the variance seen with changes in the other components can safely be neglected) [5]. This allows us to remove redundancy in the dataset. By examining plots of the new variables (principal components), researchers often unveil hidden patterns and relationships in the original data.

PCA is a well-known and widely used statistical procedure that transforms a set of multivariate data linearly into a new coordinate system of the same number of dimensionality. The dimensionality reduction using PCA is to replace the original data with a number of principal components usually less than the number of variables in the original data but carrying a significant amount of variation [6].

Mathematically, PCA is defined as an orthogonal linear transformation that transforms the data to a new coordinate system such that the greatest variance by any projection of the data comes to lie on the first coordinate (called the first principal component), the second greatest variance on the second coordinate, and so on [7]. In principal components analysis (PCA), a reduced set of m components are extracted from a set of p variables that accounts for most of the variance in the p variables. In other words, it is desired to reduce a set of p variables to a set of m underlying super ordinate dimensions. One of the benefits of PCA is that it is completely reversible and the original data can be recovered from the principal components. This makes PCA a powerful tool for data reduction, noise rejection, visualization and data compression.

In this thesis, first the dimension of the data was reduced than the anomalies and irregularities in the samples were detected. If the noise or irregular data is not removed, the results could be misleading. Relationships among the spinal features were understood from the analysis of principal components. PCA was used to transform the 24 input features to a 2D plane. Finally, clusters were formed to study the behaviour of the spinal features among different age groups. This multivariate analysis is helpful in uncovering the growth and degeneration pattern

of the human spine with the normal ageing process so that back pain and other spine problems can be better understood and dealt with.

PCA reduced the input dimensions by a factor of 8. The reduction in the dimension of input is helpful in using the compressed data in more advanced analysis, such as in neural network modelling for estimation of the spinal age. As the human mind can only visualise in 2D or 3D, by reducing the number of dimensions, this research provides an easily interpretable model for spine feature variations. The numerical experiments carried out in this section show that the principal component analysis is capable of reducing the dimensions whilst preserving the pattern and without losing vital information.

4.2 PCA MODELLING

For the proposed principal component analysis, 24 notable features of the lumbar spine magnetic resonance images were considered. These features were measured for 61 selected samples (MRI scans of the patients), belonging to different age groups and gender. The ages of the subjects ranged from 2 to 93 years. Figure 4.1 below shows the steps involved in PCA analysis.



Figure 4.1: Step by step demonstration of PCA modelling

A glimpse of the five random samples with extracted features and their scores is shown in table 4.1 below.

Table 4.1: Illustration of extracted features of lumbar spine

Gender	m/f	f	f	m	m	f
Age at time of scan	Years	8	23	40	68	89
Vertebral height mid (mm)	L1	16.94	22.82	27.16	23.95	21.7
	L2	17.34	22.98	27.16	23.57	22.06
	L3	16.8	24.57	26.08	23.53	21.94
	L4	17.34	24.65	27.85	23.53	21.33
	L5	17.22	25.94	27.25	23.95	19.11
Disc height (mm)	T12 L1	5.95	7.51	9.33	9.48	4.45
	L1 L2	7.43	9.92	11.41	12.13	6.3
	L2 L3	7.75	10.22	13.05	13.27	5.35
	L3 L4	8.34	10.84	12.67	15.15	4.69
	L4 L5	8.0	9.06	11.83	15.41	7.15
	L5 S1	6.84	11.13	7.71	10.74	5.35
Disc signal (average intensity)	T12 L1	272.4	132.5	189.4	138.8	61.9
	L1 L2	268.6	126.1	180.8	127.9	69.6
	L2 L3	255.1	123	185.2	120.2	43
	L3 L4	307.6	104.4	208.7	129.9	75.1
	L4 L5	263	95.3	138.4	137.6	67.2
	L5 S1	260	109.3	52.6	57.4	89.6
Para spinal Muscles (average intensity)	SI Right	94.6	96.4	81.4	60.4	260.1
	SI Left	97.9	80.5	74.9	54.3	155.7
Psoas Muscle (average intensity)	Right	64.2	46.9	88.6	48.1	60
	Left	65.5	42	94.9	31.7	55.5
Subcutaneous Fat (average intensity)	Right	646.5	756.8	454.5	414.9	325
	Left	785.4	624.1	462	299.6	447
CSF (average intensity)	at L3	561.8	359.4	534.8	519	639.9

In the proposed model, 61 samples ($\{y_i\}_{i=1}^{61}; y_i \in R^{24}$) were used each having 24 features or input variables. For the human perception, to visualize these 24 features all at the same time is hardly possible. To overcome this, the mean centered dataset $D = [y_1, y_2 \dots y_{61}] \in R^{24 \times 61}$ (24 features and 61 samples) is transformed to the new dataset based on the principal components.

Defining a new basis for the data (the columns \mathbf{v}_i of an orthogonal matrix V) by the transformation $\mathbf{z} = V^T \mathbf{y}$, where \mathbf{y} is any column of D , the direction of maximum variance corresponds to maximizing the variance of the first coefficient of \mathbf{y} in this new basis, i.e., z_1 , the first component of \mathbf{z} :

$$\begin{aligned} \max_{\mathbf{v}_1^T \mathbf{v}_1 = 1} \text{var}(z_1) &= \max \left\{ \text{var}(z_1) - \lambda_1 (\mathbf{v}_1^T \mathbf{v}_1 - 1) \right\} \\ &= \max \left\{ \mathbf{v}_1^T S \mathbf{v}_1 - \lambda_1 (\mathbf{v}_1^T \mathbf{v}_1 - 1) \right\} \end{aligned} \quad (4.1)$$

Where $S = (1/64)D^T D$ is the sample covariance matrix (bearing in mind that the data set has been centred) and λ_1 is a Lagrange multiplier. Taking derivatives w.r.t. \mathbf{v}_1 , it is straightforward to show that the optimal choice of \mathbf{v}_1 is an eigenvector of S , with corresponding eigenvalue λ_1 . Noting that (by the orthonormality of the \mathbf{v}_i) $\text{var}(z_1) = \mathbf{v}_1^T S \mathbf{v}_1 = \mathbf{v}_1^T \lambda_1 \mathbf{v}_1 = \lambda_1$, λ_1 should correspond to the largest eigenvalue in order to maximize $\text{var}(z_1)$. Assuming that the covariance matrix has a complete set of distinct eigenvalues, the entire PCA consists of the eigenvectors of S , arranged such that the corresponding eigenvalues satisfy. Furthermore, the principal components are uncorrelated and have variances equal to the corresponding eigenvalues: $\text{var}(z_j) = \mathbf{v}_j^T S \mathbf{v}_j = \mathbf{v}_j^T V \Lambda V^T \mathbf{v}_j = \lambda_j$, using the spectral decomposition $V \Lambda V^T$ of S , where $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_{24})$. The PCA transformation was performed using the Matlab Statistical Toolbox [8].

Although the number of components obtained (24) is the same as the number of features in the original data set, in the transformed data set, most of the information is captured by the first few components along the principal directions

v_i . These components, which explain most of the variance in the data, are called the principal components.

4.2.1 STATISTICAL ANALYSIS AND DATA STANDARDIZATION

The spinal characteristics change with age. The maximum value, minimum value, mean, and standard deviation of the variables are shown in table 4.2 below.

Table 4.2: Statistics of original data set

Variables	Minimum	Maximum	Mean	Standard Deviation
Vertebral Height L1 (mm)	12.6	27.57	23.5659	3.03906
Vertebral Height L2 (mm)	12.93	28.45	24.06328	2.855286
Vertebral Height L3 (mm)	13.55	29.02	23.92492	2.887122
Vertebral Height L4 (mm)	13.46	28.16	23.88836	2.850366
Vertebral Height L5 (mm)	13.72	27.25	23.4259	2.80882
Disc Height T12 L1 (mm)	3.27	11.53	7.874426	1.914816
Disc Height L1 L2 (mm)	2.55	12.58	8.717705	2.10848
Disc Height L2 L3 (mm)	4.19	13.88	9.537541	2.251526
Disc Height L3 L4 (mm)	3.1	15.46	9.967869	2.523685
Disc Height L4 L5 (mm)	2.18	16.88	9.493443	3.026889
Disc Height L5 S1 (mm)	1.63	14.37	9.057377	2.681584
Disc Signal T12 L1 (average intensity)	28	374.5	119.218	78.36515
Disc Signal L1 L2 (average intensity)	20.9	373.4	126.4164	79.50125
Disc Signal L2 L3 (average intensity)	26.7	436.8	118.1197	80.39867
Disc Signal L3 L4 (average intensity)	28.4	460	118.8803	85.02646
Disc Signal L4 L5 (average intensity)	28.6	473	115.677	84.86374
Disc Signal L5 S1 (average intensity)	21.9	374.2	109.9984	85.50005
PSM SI Right (average intensity)	14.1	405.5	135.382	85.61407

PSM SI Left (average intensity)	16.4	303.3	128.4443	71.88857
Psoas Right (average intensity)	25.3	150.2	66.63279	25.64504
Psoas Left (average intensity)	21.5	169.4	64.68197	28.50589
Subcutaneous Fat Right (average intensity)	167.4	836.3	510.6869	174.2971
Subcutaneous Fat Left (average intensity)	159.7	961.7	547.3131	167.2379
CSF at L3 (average intensity)	186.7	1316.3	561.1066	214.9814

Often it is convenient to compute principal components from raw data. This is appropriate when all the variables are in the same units. Standardizing the data is often preferred when the variables are in different units or when the variance of the different columns is substantial, as in this case. If the standard deviations of the variables are significantly different from one another, then one variable might dominate in the analysis [9]. Equation 4.2 below gives the standardization formula used in the PCA analysis, where x is the original observed data value, μ is the mean and σ is the sample standard deviation.

$$Z = \frac{x - \mu}{\sigma} \quad (4.2)$$

In other words, the original data set is standardized by subtracting the mean and dividing each column by its standard deviation. Then the principal components are computed using the technique described above [10], implemented in Matlab. The statistics of standardized data are shown in table 4.3 below. All the variables now have the same standard deviation, which is equal to one.

Table 4.3: Statistics of standardized data set

Variables	Observations	Minimum	Maximum	Mean	Standard Deviation
Vertebral Height L1 (mm)	61	-3.6083	1.31754	0	1
Vertebral Height L2 (mm)	61	-3.8992	1.53635	0	1
Vertebral Height L3 (mm)	61	-3.5935	1.76476	0	1
Vertebral Height L4 (mm)	61	-3.659	1.49863	0	1
Vertebral Height L5 (mm)	61	-3.4555	1.36146	0	1
Disc Height T12 L1 (mm)	61	-2.4046	1.90910	0	1
Disc Height L1 L2 (mm)	61	-2.9252	1.83179	0	1
Disc Height L2 L3 (mm)	61	-2.3751	1.92867	0	1
Disc Height L3 L4 (mm)	61	-2.7214	2.17623	0	1
Disc Height L4 L5 (mm)	61	-2.4162	2.44031	0	1
Disc Height L5 S1 (mm)	61	-2.7698	1.98115	0	1
Disc Signal T12L1 (average intensity)	61	-1.1640	3.25760	0	1
Disc Signal L1 L2 (average intensity)	61	-1.3272	3.106663	0	1
Disc Signal L2 L3 (average intensity)	61	-1.1371	3.96375	0	1
Disc Signal L3 L4 (average intensity)	61	-1.0641	4.01192	0	1
Disc Signal L4 L5 (average intensity)	61	-1.0261	4.21055	0	1
Disc Signal L5 S1 (average intensity)	61	-1.0304	3.09008	0	1
PSM SI Right (average intensity)	61	-1.4166	3.15507	0	1
PSM SI Left (average intensity)	61	-1.5586	2.43232	0	1
Psoas Right (average intensity)	61	-1.6118	3.25861	0	1
Psoas Left (average intensity)	61	-1.5149	3.67356	0	1
Fat Right (average intensity)	61	-1.9695	1.8681	0	1
Fat Left (average intensity)	61	-2.3177	2.47783	0	1
CSF at L3 (average intensity)	61	-1.7416	3.51283	0	1

4.2.2 DATA VISUALIZATION

Using the Matlab Statistical Toolbox, 24 components were computed from 24 original features. Figure 4.2(a) below shows the first five components where each component shows the variance in the data. It can be seen that much of the variance (about 88.5%) is captured by first three components. The first component accounts for 36.08 % of the variance, second component accounts for 31.64 % and the third component accounts for about 20.78 %. Since 88.5% of the variance is explained by the first three components, the remaining components were discarded. In other words, 24 features were replaced by 3 components, giving a new dataset in $R^{3 \times 61}$ (compression by a factor of 8).

Using the three components, the original data can now be easily visualized using 3D plots or a series of three 2D plots (1st comp vs. 2nd comp, 1st comp vs. 3rd comp and 2nd comp vs. 3rd comp).

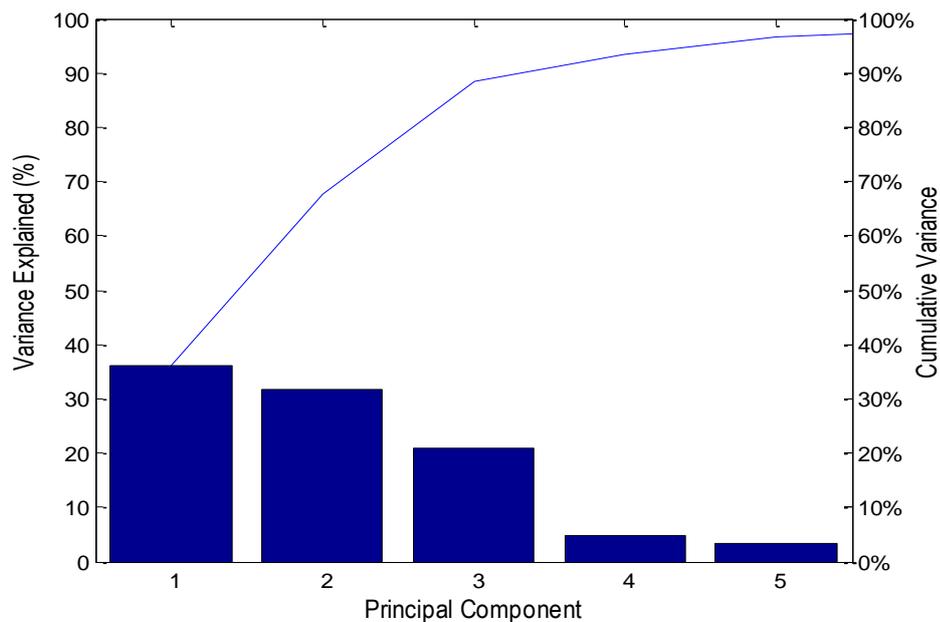


Figure 4.2: (a) Variance shown by first five components of the data

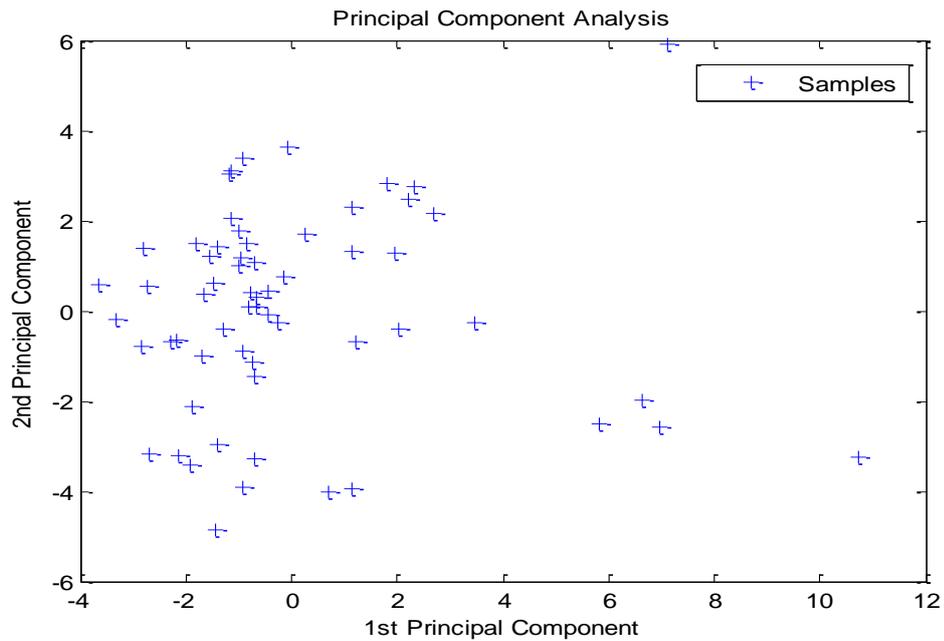


Figure 4.2: Plot of 1st principal component vs. 2nd principal component

A plot of first principal component vs. the second principal component, which account for about 68% of the variance in the data, is shown in figure 4.2 (b) above. It demonstrates the distribution of the 61 samples ranked on the basis of their spinal scores. It can be seen that most of the samples lie in the range $[-4, 4]$ of the 1st principal component whereas the samples are evenly distributed with respect to the 2nd principal component.

Each MRI sample was labelled with the corresponding age to study the patterns in the data. Figure 4.3 below shows the plot of the samples using the first two principal components labelled according to age. It can be seen from figure 4.3 that as we move along the 1st principal component, age tends to increase. Samples corresponding to age > 30 are almost entirely contained in the positive range of the 1st principal component. The first principal component can therefore be used as a descriptor for age. With respect to the second principal component, the samples ages exhibit no discernible pattern.

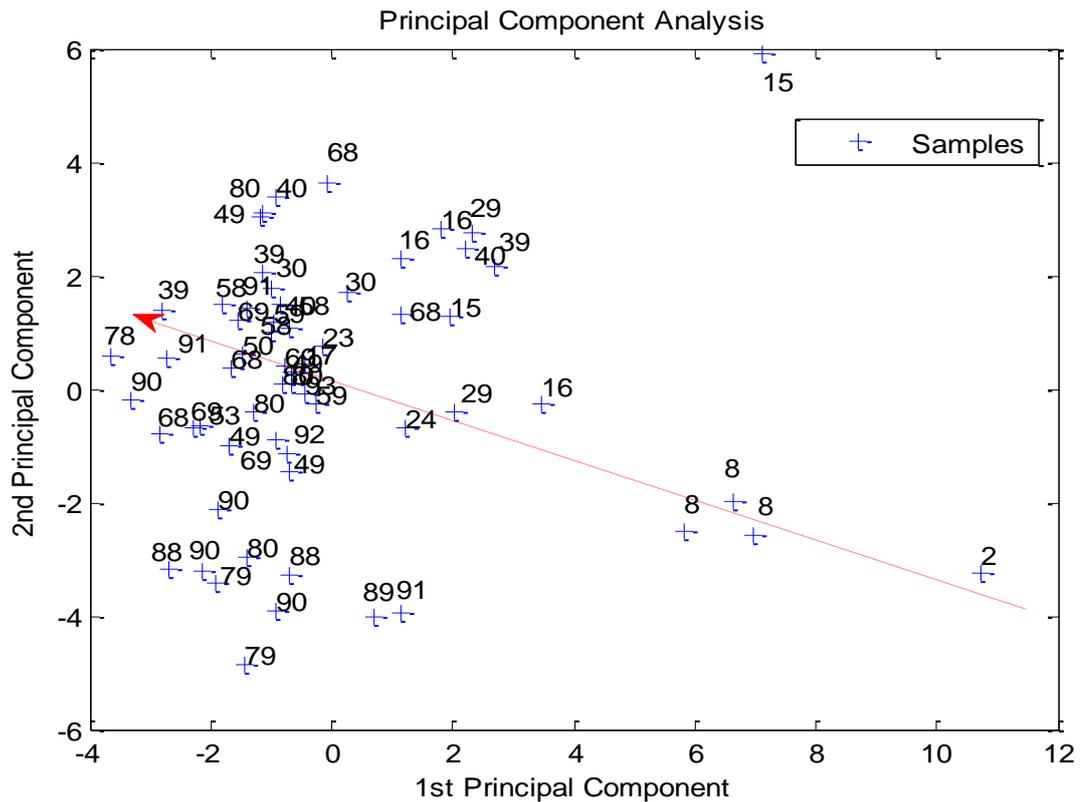


Figure 4.3: PCA plot of first two components with age labeled

A plot of first two components with samples labelled with age and the gender (m=male; f=female) is shown in figure 4.4. It can be seen that the majority of the samples in the positive (upper) half of the 2nd principal component axis are of females whereas most of the samples in negative (lower) half of the second principal axis are of males. This gender bias is shown with the red dotted line in figure 4.4. Although there are few exceptions, but it can be fairly said that the second principal component is a descriptor for gender.

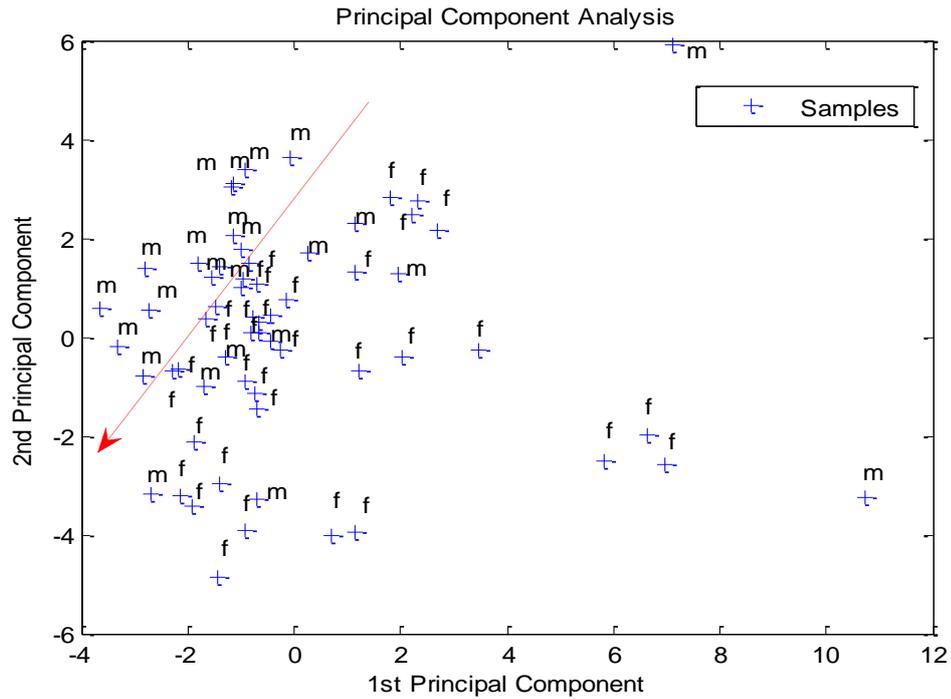


Figure 4.4: PCA plot of first two components with age and gender labeled

A plot of principal component1 vs. principal component3 is shown in figure 4.5 below. Again principal component 1 acts well as a descriptor for age. With respect to principal component 3, female samples are concentrated close to zero whereas male samples tend to lie away from zero, either between: 1 to 4 or -1 to -4.

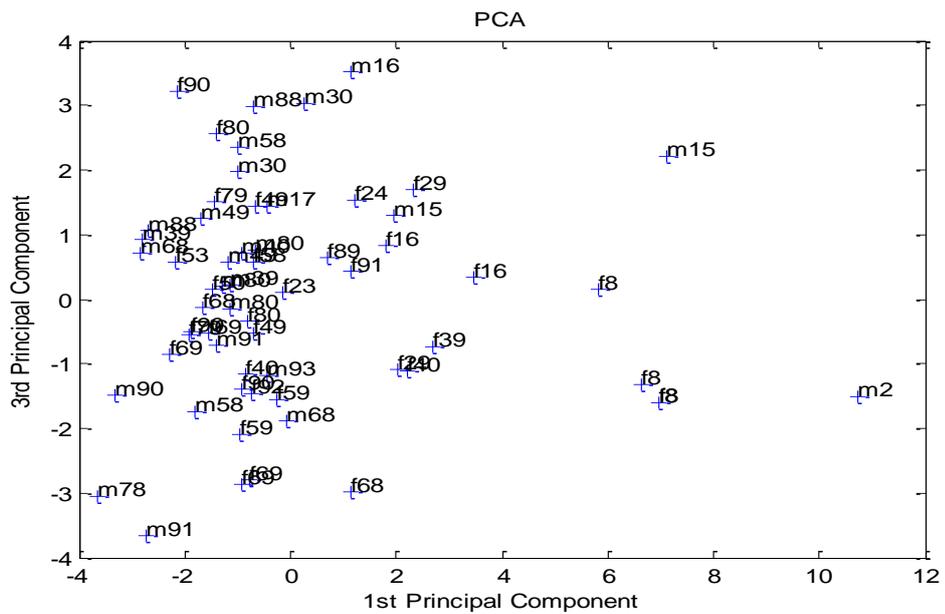


Figure 4.5: Plot of PC 1 vs. PC3

A plot of the second and third principal components is shown in figure 4.6 below. There is a gender bias with respect to principal component 2 (a majority of male samples lie on the positive half and majority of female samples lie in negative half of the component 2 axis).

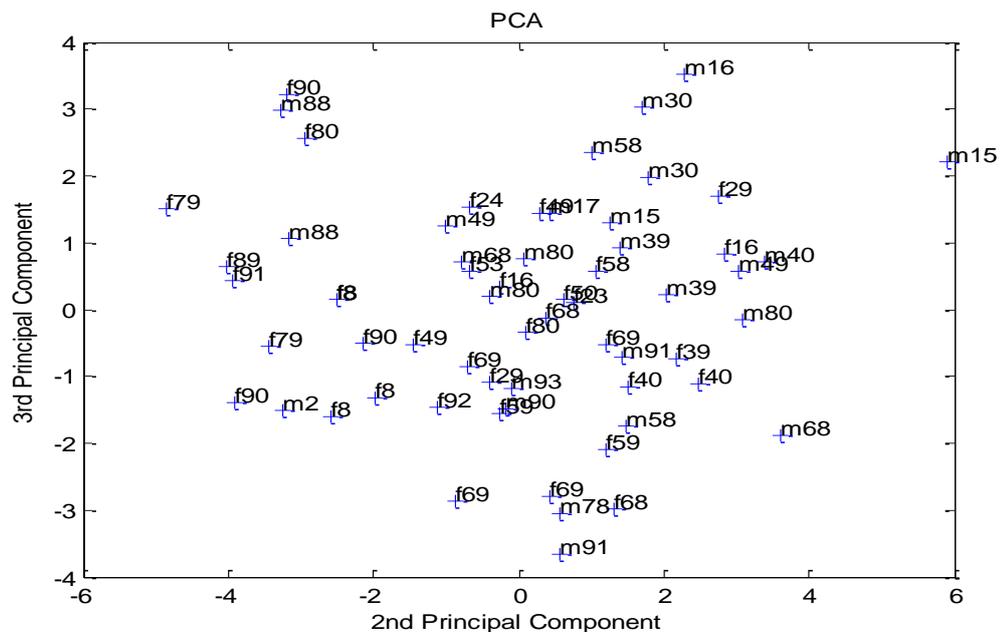


Figure 4.6: Plot of PC2 vs. PC3

4.2.3 CORRELATION OF FEATURES WITH AGE

Reducing the dimension of the data and plotting the principal components provides an interpretable visual representation of the data. This 2D representation uncovers several patterns in the data. More knowledge can be extracted by exploring the driving force (dominance of certain features) with allocates or ranks the samples in the 2D plane. Plotting the 1st vs. 2nd principal components and exploring the driving force for the allocation of samples provides the clinical significance of the input variables.

A plot of the samples and variables for principal components 1 and 2 (accounting for 68% variance) is shown in figure 4.7 below. Each blue line in figure 4.7 represents the corresponding variables as labelled. The magnitude of line indicates the significance of that variable in terms of its effect on ageing spine. The most significant here is disc signal intensities located in the upper right (first) quadrant. The other significant variables are vertebral heights and disc heights located in the bottom right (second) quadrant. In the bottom left (third) quadrant, the para spinal muscle signal intensity shows high significance whereas subcutaneous fat signals and psoas signals shows exhibit a lesser significance. The least significant is CSF, located in the top left (fourth) quadrant.

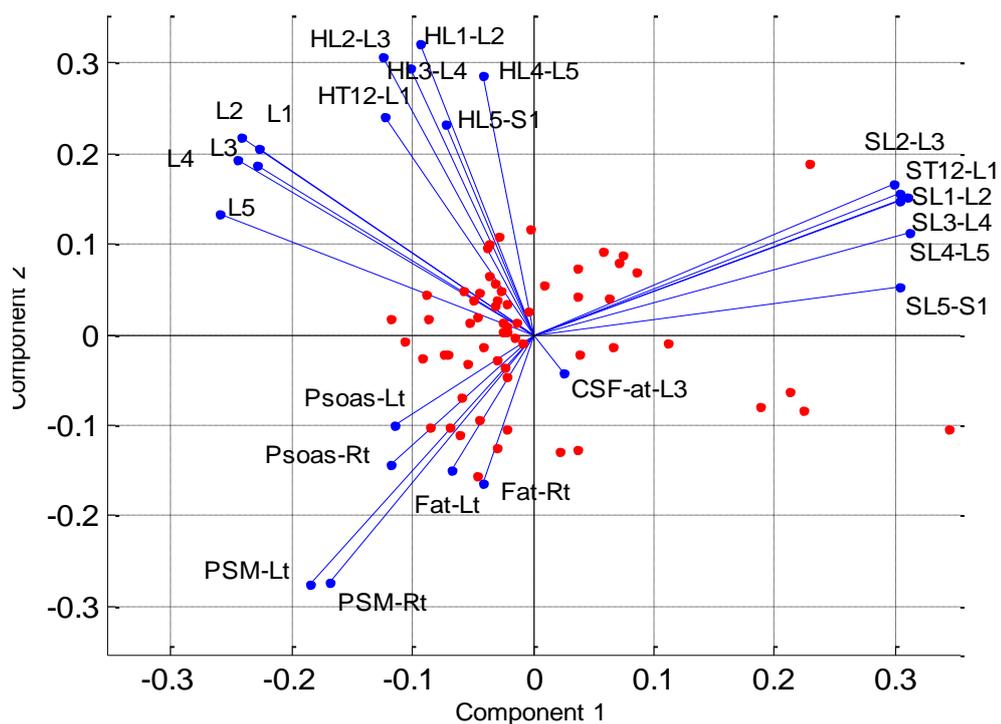


Figure 4.7: Plot of samples and the variables

Figure 4.7 above not only gives the significance of the input variables but also indicates the correlation among the input variables. Looking at the first quadrant, it

can be seen that disc signal intensities (T12-L1, L1-L2, L2-L3 and L3-L4) overlap considerably, showing that these variables are correlated and exhibit a similar pattern. The disc signal intensities L4-L5 and L5-S1 exhibit a slightly different behaviour from the rest. Similarly, by looking at the vertebral heights in the second quadrant, L1, L2, L3, and L4 are very close to one another whereas L5 is somewhat isolated. This shows that the pattern of vertebral height L5 is slightly different from the rest of lumbar vertebrae.

For the disc heights, it can be seen from figure 4 above that the pattern of disc height L1-L2 resembles more to disc height L3-L4 than the disc height L2-L3 (which is anatomically the next disc). Also, disc height L4-L5 is located slightly away from the rest. Disc height L5-S1 is the least significant in lumbar disc heights. In the third quadrant, the para spinal muscle signal intensities are close to one another. It can also be noticed that pattern of the left psoas and left fat signals is slightly different from the right psoas and right fat signals.

Here we have seven sets of input variables as vertebral heights, disc heights, disc signal intensities, psoas signals, fat signals, para-spinal muscle signal intensities, and CSF. These features are thought to be interconnected and to affect each other. From the PCA representation given in figure 4.7 we can draw conclusions regarding the existing correlation among these set of variables. The variables located in the same quadrant are correlated and move together. The variables that are located in adjacent quadrants (or are at right angles) appear to be uncorrelated. The variables in opposite quadrants (1st vs. 3rd and 2nd vs. 4th) or at roughly 180° to each other are negatively correlated. In other words, an increase in one is associated with a

decrease in the other. From figure 4, vertebral heights and disc heights lie in the same quadrant and are therefore correlated. Similarly, the psoas signal, fat signal and para spinal muscle signal intensity are correlated.

4.2.4 *NON CORRELATED FEATURES*

The magnitude of the blue lines in figure 4.7 gives the significance of the respective variable. These results are based on PC1 and PC2 only, since they capture more than 2/3 of the variance in the data. It can be seen from figure 4.8 that CSF is the least significant variable. Disc signal, psoas signals, fat signals, and para-spinal muscle signals are located adjacent to CSF meaning that they have no correlation with CSF. However, CSF has a negative correlation with the vertebral heights and disc heights. Similarly, the disc signal intensities have a negative correlation with the psoas signal, fat signal, and para-spinal muscle signal. Vertebral and disc heights have no correlation with the disc signal, psoas signal, fat signal, and para-spinal muscle.

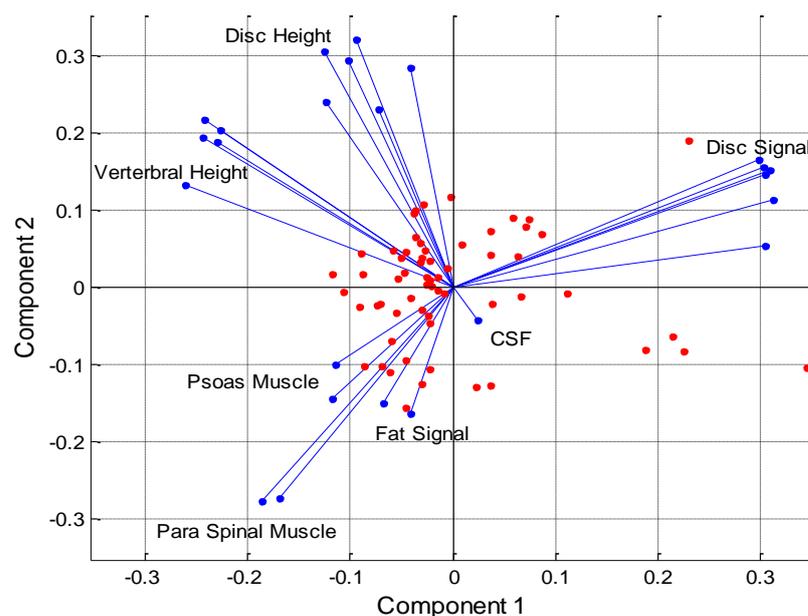


Figure 4.8: PCA plot of samples with variables

4.3 FACTOR ANALYSIS

Factor analysis is a statistical method used to describe variability among observed, correlated variables in terms of a potentially lower number of unobserved variables called factors [11]. Factor analysis is in some way related to principal component analysis but the two are not identical [12]. Principal component analysis is used to find optimal ways of combining variables into a small number of subsets, while factor analysis may be used to identify the structure underlying such variables and to estimate scores to measure latent factors themselves [13]. A generic relationship among the variables is presented in the previous section. In factor analysis it is assumed that the covariance matrix has the following form:

$$\Sigma_x = \Lambda \Lambda^T + \psi \quad (4.3)$$

where Λ is the loading matrix, and the elements of the diagonal matrix ψ are the specific variances. If x is a vector of observed variables and μ is a constant vector of means, then the factor analysis model can also be represented as:

$$x = \mu + \Lambda f + e \quad (4.4)$$

where Λ is a constant d -by- m matrix of factor loadings, f is a vector of independent standardized common factors, and e is a vector of independent specific factors. In this section, factor analysis is used to obtain the numerical values of the significance and correlation seen among the features. Table 4.4 below gives the factor loadings by considering only the first three factors.

Table 4.4: Loadings of first three factors

Factor loadings	Variables	Factor 1	Factor 2	Factor 3
Vertebral height (Mid)	L1	0.0019	0.8857	0.0308
	L2	-0.0035	0.9415	0.0476
	L3	0.0426	0.9946	-0.0730
	L4	0.0029	0.9306	0.0264
	L5	-0.1210	0.8539	-0.0290
Disc height (Mid)	T12 L1	-0.1110	0.0622	0.6307
	L1 L2	0.0660	0.1004	0.8193
	L2 L3	-0.0307	0.0446	0.8996
	L3 L4	-0.0687	-0.0708	0.8840
	L4 L5	0.0170	-0.0965	0.7540
	L5 S1	-0.1250	-0.1200	0.6653
Disc signal	T12 L1	0.9271	-0.0178	0.0043
	L1 L2	0.9711	0.0430	-0.0492
	L2 L3	0.9916	0.0443	0.0095
	L3 L4	0.9824	0.0130	-0.0292
	L4 L5	0.8838	-0.0986	-0.0503
	L5 S1	0.7773	-0.1804	-0.0877
Para spinal muscle	SI Right	-0.5270	-0.0863	-0.1911
	SI Left	-0.5833	-0.0928	-0.1751
Psoas muscle	Right	-0.2450	-0.0052	-0.0171
	Left	-0.2237	-0.0339	0.0892
Subcutaneous fat	Right	-0.1887	-0.0742	-0.1325
	Left	-0.1879	-0.0005	-0.1198
CSF	at L3	0.1536	0.0310	-0.0545

It can be seen from table 4.4 that the first factor has very high loadings of disc signal intensities. Looking at the loadings of factor one; it was found that the disc signal intensities were the most significant feature that varies with the age. A general rule of thumb is that any variable having a loading value greater than or equal to 0.7 is significant. However, this level is very high and most researchers use 0.4 as an appropriate level for real-life data analysis [14]. The loadings for disc signal intensities in the order of their significance are given below with L2-L3 being the most significant:

Disc Signal L2 L3 → 0.99155

Disc Signal L3 L4 → 0.98236

Disc Signal L1 L2 → 0.97114

Disc Signal T12 L1 → 0.92709

Disc Signal L4 L5 → 0.88377

Disc Signal L5 S1 → 0.77728

Other notable variables for factor-one are:

Paraspinal Muscle Left → - 0.58326

Paraspinal Muscle Right → - 0.52695

Psoas Muscle Right → - 0.24495

Psoas Muscle Left → - 0.22365

Subcutaneous Fat Right → - 0.18871

Subcutaneous Fat Left → - 0.18790

Cerebrospinal fluid at L3 → 0.15359

This shows that PSM SI left-right, Psoas left-right, and subcutaneous fat left-right, are negatively correlated with the disc signal intensities. For a unit increase in the

disc signal intensity, there is half a unit decrease in the para spinal muscle signal intensity and a quarter unit decrease in the psoas signal. Similarly, looking at the loading of factor 2, the vertebral height is the only significant variable. All other variables have very small loading values, which can be neglected. The loadings for factor 2 in the order of their significance are:

Vertebral Height L3 → 0.99457
Vertebral Height L2 → 0.94146
Vertebral Height L4 → 0.93064
Vertebral Height L1 → 0.88570
Vertebral Height L5 → 0.85393

By inspecting the loadings of factor 3, disc heights is the most significant variable.

The disc heights are listed below on the basis of their significance.

Disc height L2 L3 → 0.89956
Disc height L3 L4 → 0.88404
Disc height L1 L2 → 0.81926
Disc height L4 L5 → 0.75404
Disc height L5 S1 → 0.66527
Disc height T12 L1 → 0.63065

4.4 HIERARCHICAL CLUSTERING

Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar to each other than to those in other groups [15]. In data mining, hierarchical clustering is a method of cluster analysis that seeks to build a hierarchy of clusters [16]. Hierarchical clustering

groups data over a variety of scales by creating a cluster tree or dendrogram. The tree is not a single set of clusters, but rather a multilevel hierarchy, where clusters at one level are joined as clusters at the next level [17]. This allows us to decide the level or scale of clustering that is most appropriate for the application. The clustering was performed in Matlab.

4.4.1 *AGGLOMERATIVE CLUSTERING*

In this section, an agglomerative clustering technique is used. It is a "bottom up" approach in which each observation starts in its own cluster, and pairs of clusters are merged as one move up the hierarchy. It can produce an ordering of the objects, which may be informative for data display.

Agglomerative Clustering Algorithm:

The following procedure [18] is used to perform agglomerative hierarchical cluster analysis on the lumbar spine data set:

- 1). First the similarity/dissimilarity between every pair of objects in the data set is found. In this step, the Euclidean formula is used to compute the distance between objects.
- 2). Then the objects are grouped into a binary, hierarchical cluster tree. In this step, pairs of objects that are in close proximity are linked together using single linkage function which is also called as nearest neighbor linkage. The distance $D(A, B)$ between clusters A and B is given by $D(A, B) = \min_{a \in A, b \in B} d(a, b)$. This distance information is used to determine the proximity of objects to each other. As objects

are paired into binary clusters, the newly formed clusters are grouped into larger clusters until a hierarchical tree is formed.

3). Finally, the hierarchical tree is cut into the desired clusters, pruning branches off the bottom of the hierarchical tree, and assigning all the objects below each cut to a single cluster. This creates a partition of the data. These clusters can be created by detecting natural groupings in the hierarchical tree or by cutting off the hierarchical tree at any desired point.

The scatter plot of principal component1 (PC1) vs. principal component2 (PC2) is given in figure 4.9 below.

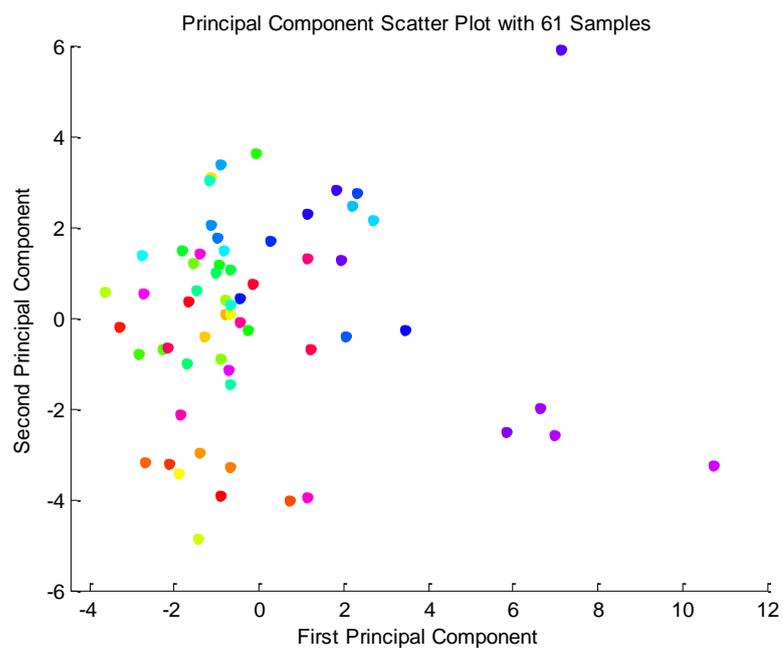


Figure 4.9: Principal component1 vs.principal component2 scatter plot

4.4.2 CLUSTER ANALYSIS

PCA transformed the input features into principal components. Using these principal components, the agglomerative clustering was performed. The Euclidean

distance between the samples was calculated and the samples that are similar or close to one another were linked or merged together to form a cluster. In this way, clusters close to one another were merged at each level. Figure 4.10 below shows a 2D scatter plot of the first two principal components (which account for almost 2/3 of the variation in the data set) with ten clusters. Samples of same age decade tend to lie close to one another with few exceptions. From figure 4.10 below, it can be seen that samples of age 2 (purple cluster) and age 8 (blue cluster) exhibit a different behaviour from the rest, and hence are put in a different cluster. Similarly, a green cluster can also be seen having samples ranging from 80 to 91 years of age scoring low on both principal components.

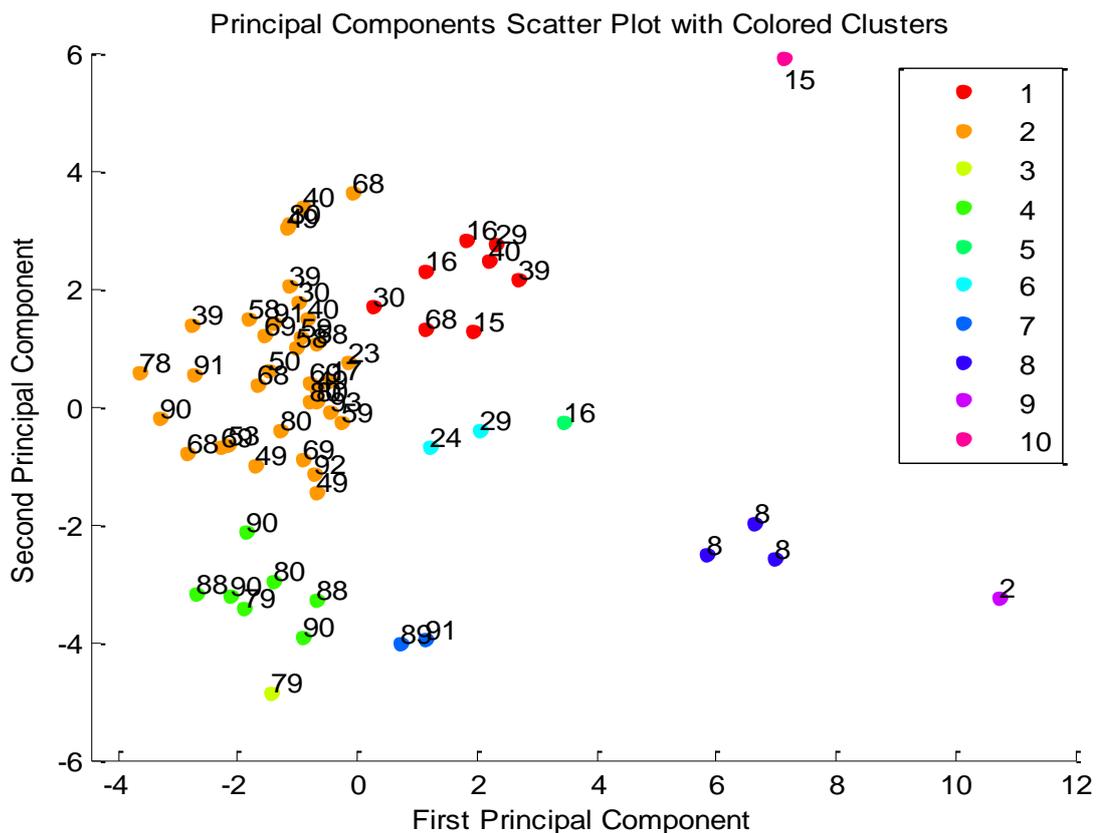


Figure 4.10: PC1 vs. PC 2 with 10 clusters

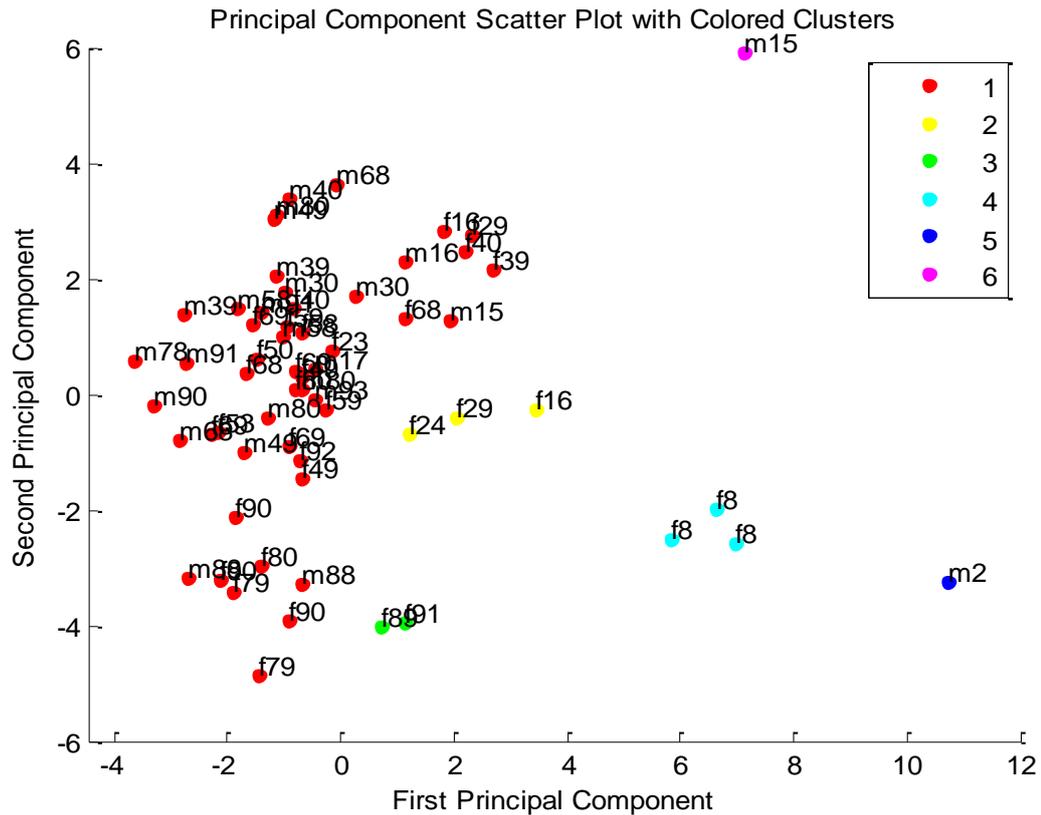


Figure 4.11 (b): 6 clusters with age and gender labelled

Looking at figure 4.11 (a), cluster one (red) consists of 2 female samples aged 89 and 91. They score slightly higher on PC1 as compared to the other samples of the same age decade. Similarly, cluster 3 of female samples aged 8, and cluster 4 of males aged 2 are isolated from the other clusters. This shows that samples tend to score low on PC1 with increasing age. From the factor analysis, it was observed that factor 1 can represent the disc signal intensity, so it can be concluded that disc signal intensities tends to decrease with increasing age.

The tree was further trimmed and 6 clusters were formed as shown in figure 4.11 (b). Here it can be seen that all four clusters remains exactly the same except the largest cluster. The largest cluster in figure 4.11 (a) (green) is broken into two clusters. These two new clusters are represented by red and yellow in figure 4.11

(b). The yellow samples corresponding to females aged 16, 24, and 29 were isolated from the rest of their age group. These three samples score lower on PC2 as compared to the samples from the same age decade. Factor 2 reflects vertebral heights, so it can be inferred that these three samples have lower vertebral heights as compared to the other samples from the same age decade. Figure 4.12 below shows the agglomerative clustering with 7 and 8 clusters where big groups are further divided. Forming more number of clusters gives the composition of larger groups. Several trials were conducted by changing the number of clusters so that a logically appropriate number of groups can be formed which can describe spinal changes in different phases of natural ageing.

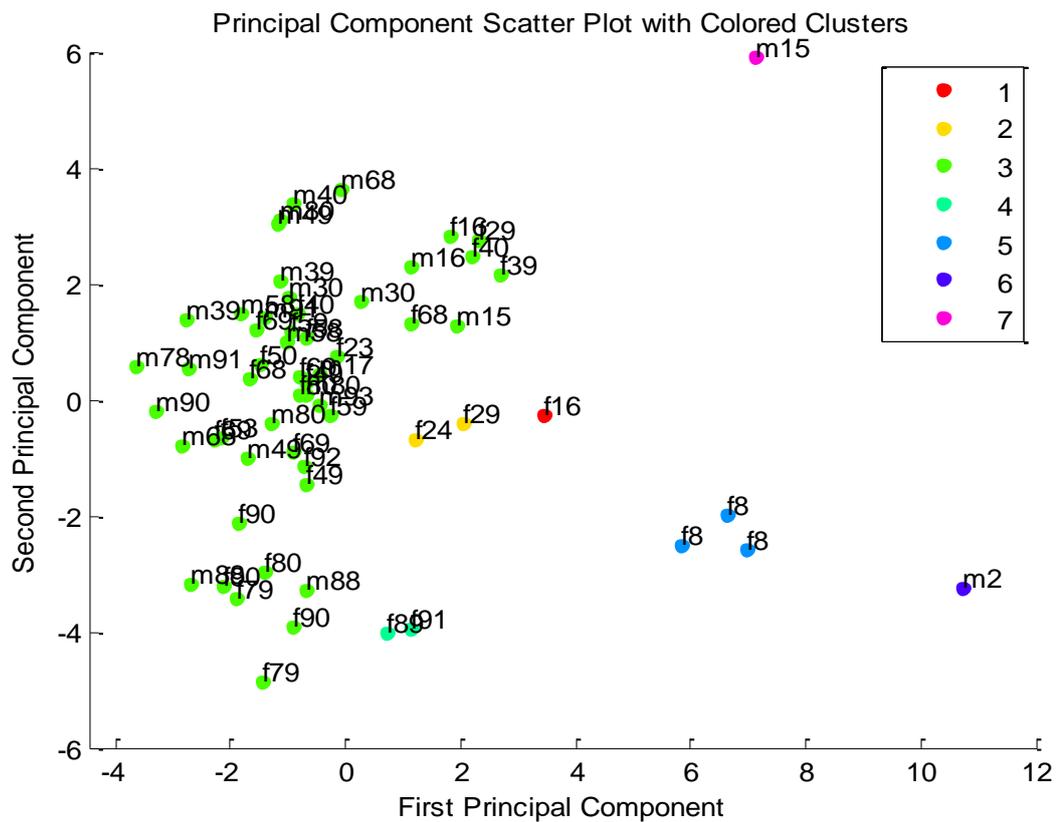


Figure 4.12 (a): Cluster analysis with 7 clusters

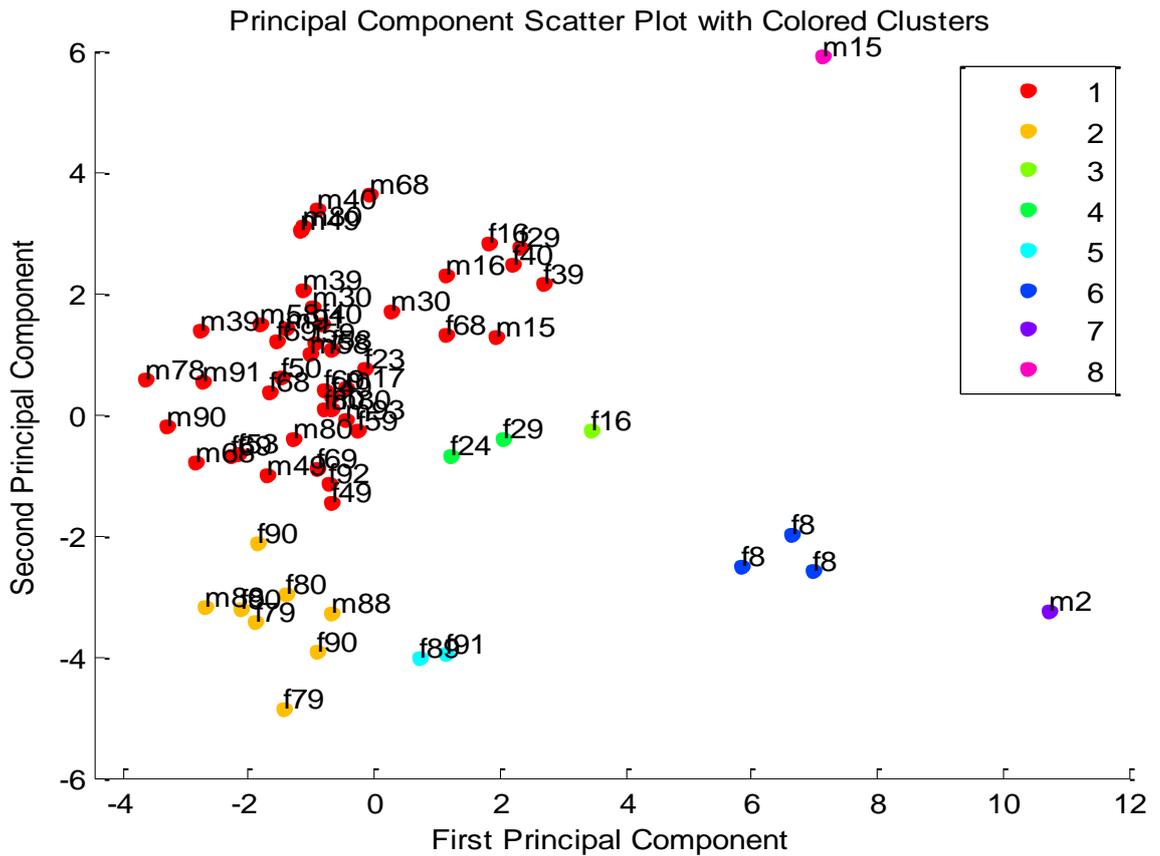


Figure 4.12 (b): Clustering with 8 clusters

Finally, the number of cluster was set to ten. As before, we divided data set into ten age decades to investigate the difference in patterns among these age decades. It can be seen from figure 4.13 (b) that most of the samples in the 40's, 50's, and 60's are located close to each other. This shows that there is very little change in the lumbar spine features during ages 40-60. However, the samples with age in the 10's, 20's, 80's, and 90's show some distinct changes in spinal features. This suggests that in these age decades, spinal features shows considerable growth and degeneration, respectively. The dendrogram is shown in figure 4.13 (a). It shows the samples that are merged at each stage. The dissimilarity measure increases as we go up the hierarchy.

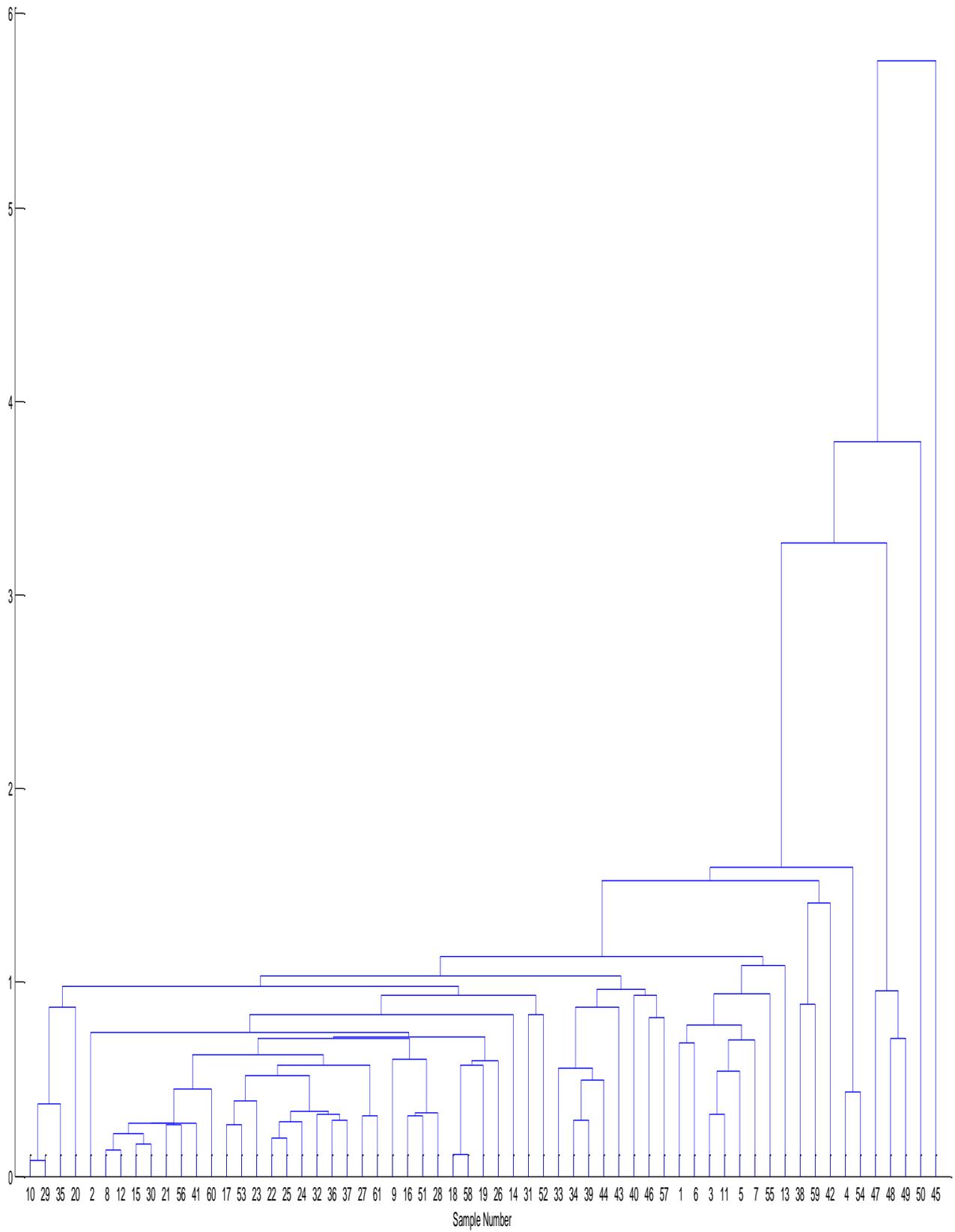


Figure 4.13 (a): Dendrogram with sample number on x-axis and dissimilarity

measure on y-axis

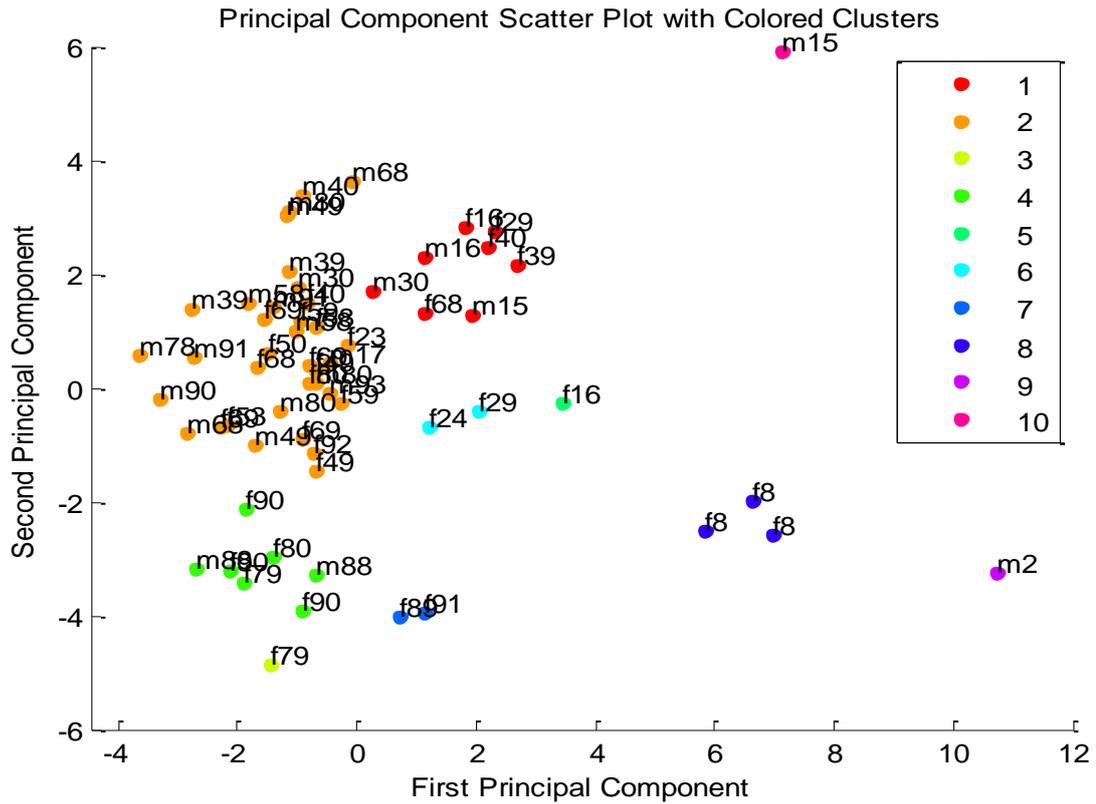


Figure 4.13 (b): 10 Clusters with age and gender labelled

4.5 PCA BASED ANOMALY DETECTION

Anomaly detection, also known as outlier detection, is the search for items or events that do not conform to an expected pattern [20]. Anomaly detection methods have been designed and well described for various professional domains. They are becoming very popular in the medical field [21], [22]. There are three broad categories of anomaly detection: supervised, unsupervised and semi supervised. In this chapter, a cluster distance based unsupervised technique is used for anomaly detection. PCA has provided a 2D representation of the multivariate data and clustering has grouped similar samples together. Anomalies can be detected by the visual inspection of clusters. Clusters that exhibit a desired pattern might contain noise or contain misfit samples, which can eventually lead to

inaccurate results. Considering the two clusters formed in figure 4.14 below, it can be seen that the red cluster has only one sample, which is far from all other samples. It scores very highly on the 1st and 2nd principal components, so this sample is an outlier adding ambiguity and noise to the data.

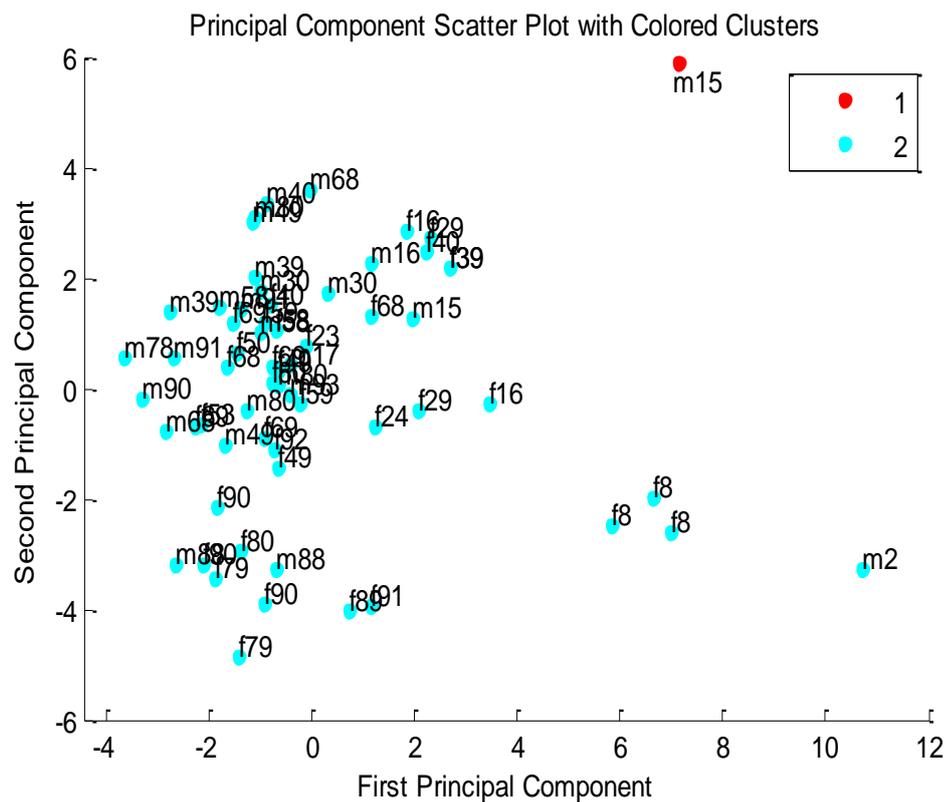


Figure 4.14: PC1 vs. PC2 with 2 clusters

Closely studying the features of this specific sample shown in red, it was noticed that the disc signal intensity of this sample is almost twice as large as those of the other samples in the same age decade. A comparison of the disc signal intensity of the detected sample (male age 15) and the average values between ages 10-20 years is shown in table 4 below. This unusually high reading of disc signal intensity could be due the onset of some spinal disease.

Table 4.5: Comparison of anomaly and average values of its respective age decade

Variables	Average Value	Anomaly Value
Disc Signal T12-L1	182.92	374.5
Disc Signal L1-L2	209.44	373.4
Disc Signal L2-L3	190.86	436.8
Disc Signal L3-L4	188.8	460
Disc Signal L4-L5	165.84	473
Disc Signal L5-S1	168.9	374.2

Similarly, 9 clusters are shown in figure 4.15 below. Other than the previously discussed cluster (male aged 15), another distinct cluster shown in red was formed, consisting of one sample of a female age 79. This sample scores very low on both principal components. Upon inspecting the spinal scores of this sample, it was found that it has an unusually low reading on the disc signal intensity, the disc height and the vertebral height. Again, these unusually low values could be due to some spinal disease. By using the suggested clustering technique, such unusual samples can be easily identified and eliminated. In this section, the anomalies from only first two principal components were reported as these two components account for 68% of the variance in the dataset. Anomalies from remaining less significant components can be identified in the same way but due to the small sample size, the anomalies from other less significant components were not taken into account in this thesis. This technique can be used to classify the problematic spines (when a visual inspection of lumbar spine MRI can easily overlook such findings).

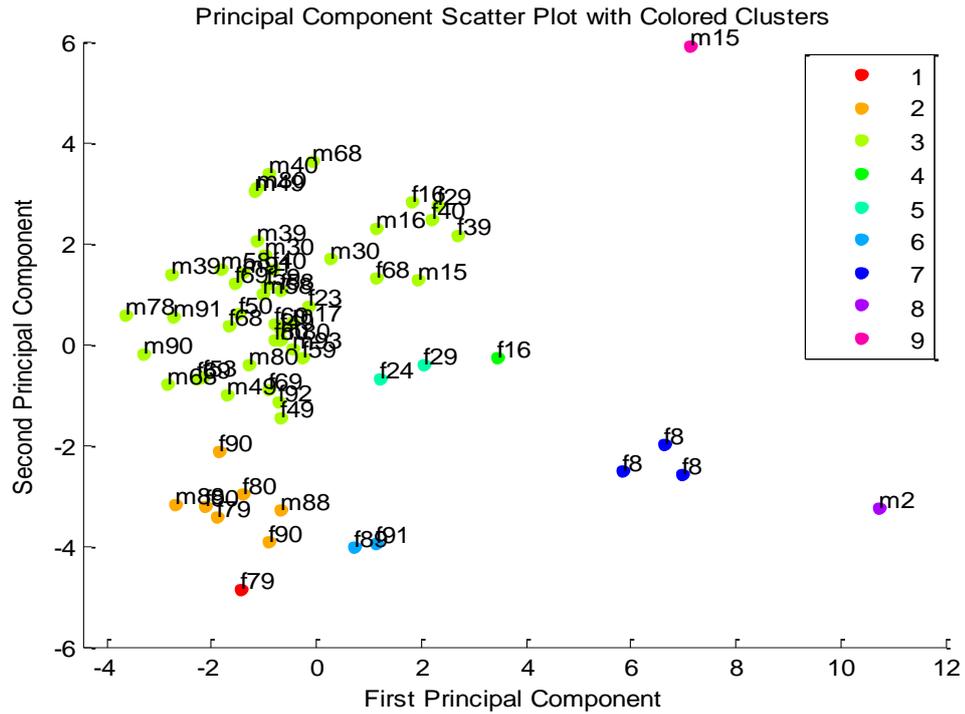


Figure 4.15: 9 Clusters with age and gender labelled

4.6 SUMMARY

Principal component analysis (PCA) on extracted MRI data was described in this chapter. PCA reduced the dimension of the inputs by a factor of 8 thus enabling us to visualize multivariate data. Further exploratory data analysis was conducted by analysing PCA plots and component loadings. Agglomerative clustering was used for anomaly detection. This chapter also presented the results of a factor analysis that helped in better understanding the relationships among the spinal features.

REFERENCES

- [1] Jackson, J. Edward. A user's guide to principal components. Vol. 587. John Wiley & Sons, (2005).
- [2] Krzanowski, W. J. "Principles of Multivariate Analysis—A User's Perspective (Revised ed.) Oxford University Press." (2000).
- [3] Online course material, STAT-505, "Applied Multivariate Statistical Analysis, Lesson 7: Principal Components Analysis (PCA)", The Pennsylvania State University. USA.
- [4] Abdi, Hervé, and Lynne J. Williams. "Principal component analysis." Wiley Interdisciplinary Reviews: Computational Statistics 2, no. 4 (2010): 433-459.
- [5] Jolliffe I.T., "Principal Component Analysis", Springer Series in Statistics, 2nd ed., Springer, NY, (2002).
- [6] Jolliffe, Ian. Principal component analysis. John Wiley & Sons, Ltd, 2005.
- [7] Jolliffe I.T., "Principal Component Analysis", Springer Series in Statistics, 2nd ed., Springer, NY, (2002), XXIX, ISBN 978-0-387-95442-4.
- [8] Jones, Bradley. MATLAB: Statistics Toolbox; User's Guide. MathWorks, (1997).
- [9] Suhr, Diana D. "Principal component analysis vs. exploratory factor analysis." SUGI 30 Proceedings (2005): 203-230.
- [10] Shlens, Jonathon. "A tutorial on principal component analysis." Systems Neurobiology Laboratory, University of California at San Diego 82 (2005).

- [11] Van Belle, Gerald, Lloyd D. Fisher, Patrick J. Heagerty, and Thomas Lumley. *Biostatistics: a methodology for the health sciences*. Vol. 519. John Wiley & Sons, (2004): 584-639.
- [12] Martinez, Wendy L., Angel Martinez, and Jeffrey Solka. *Exploratory data analysis with MATLAB*. CRC Press, (2004).
- [13] Kim, Jae-On, and Charles W. Mueller, eds. *Introduction to factor analysis: What it is and how to do it*. No. 13. Sage, (1978).
- [14] Costello, A. B., and J. W. Osborne. "Best practices in exploratory factor analysis: four recommendations for getting the most from your analysis." *Pract Assess Res Eval*, no. 10 (2005): 1-9.
- [15] Kaufman, Leonard, and Peter J. Rousseeuw. *Finding groups in data: an introduction to cluster analysis*. Vol. 344. John Wiley & Sons, (2009).
- [16] Johnson, Stephen C. "Hierarchical clustering schemes." *Psychometrika* 32, no. 3 (1967): 241-254.
- [17] Jain, Anil K., M. Narasimha Murty, and Patrick J. Flynn. "Data clustering: a review." *ACM computing surveys (CSUR)* 31, no. 3 (1999): 264-323.
- [18] Jones, Bradley., *MATLAB: Statistics Toolbox User's Guide R2012b*, Revised version 8.3, MathWorks, (1993).
- [19] Fraley, Chris, and Adrian E. Raftery. "How many clusters? Which clustering method? Answers via model-based cluster analysis." *The computer journal* 41, no. 8 (1998): 578-588.
- [20] Chandola, Varun, Arindam Banerjee, and Vipin Kumar. "Anomaly detection: A survey." *ACM Computing Surveys (CSUR)* 41, no. 3 (2009): 15.

- [21] Beutel, Jacob, and Richard G. Stafford. "Application of neural networks as an aid in medical diagnosis and general anomaly detection." U.S. Patent 5,331,550, issued July 19, (1994).
- [22] Wong, Weng-Keen, Andrew Moore, Gregory Cooper, and Michael Wagner. "Bayesian network anomaly pattern detection for disease outbreaks." In ICML, (2003): 808-815.

5

ESTIMATION OF AGE AND GENDER FROM LUMBAR SPINE FEATURES

5.1 MACHINE LEARNING IN MEDICINE

5.2 ARTIFICIAL NEURAL NETWORK (ANN)

5.3 RESULTS FROM NEURAL NETWORK MODELLING

5.3.1 SPINAL AGE ESTIMATION

5.3.2 GENDER ESTIMATION FROM SPINAL FEATURES

5.3.3 BOTH AGE AND GENDER ESTIMATION

5.4 CROSS VALIDATION

5.4.1 CROSS VALIDATION TECHNIQUES

5.4.2 REPEATED K-FOLD CROSS VALIDATION

5.5 PRINCIPAL COMPONENT NEURAL NETWORK MODEL

5.6 FUZZY INFERENCE SYSTEM (FIS)

5.6.1 MAMDANI FUZZY INFERENCE SYSTEM

5.6.2 SUGENO FUZZY INFERENCE SYSTEM

5.7 HYBRID MODEL

5.7.1 ADAPTIVE NEURO FUZZY INFERENCE SYSTEM (ANFIS)

5.7.2 ANFIS WITH SUBTRACTIVE CLUSTERING

5.8 SUMMARY

REFERENCES

5.1 MACHINE LEARNING IN MEDICINE

Machine learning is widely used in medical informatics, diagnosis, pattern recognition and clinical decision making [1]. It finds its application in almost every area of medical science ranging from genomics, study of human brain dynamics, analysis of biological relationships, image informatics, infectious disease study, to disease diagnosis [2]. The area of machine learning provides methods, techniques, and applications that are capable of solving diagnostic and prognostic problems in various medical domains. These methods/techniques are increasingly used for the analysis of important clinical parameters, predictions of disease progression, extraction of medical knowledge, and for overall patient management [3]-[5]. They are also used for data analysis, anomaly detection, interpretation of continuous data, and for intelligent alarming resulting in effective and efficient monitoring. Systems based on machine learning facilitate healthcare professionals, ultimately improving the efficiency and quality of medical care. Neural networks are known for their strength when it comes to prediction and estimation. Artificial neural networks (ANNs) represent a powerful tool to facilitate physicians perform diagnosis and other implementations. ANNs have several advantages including their ability to process large volumes of data, a reduced likelihood of overlooking relevant information and a reduction in diagnosis time [6].

In this chapter, an artificial neural network is used for “spinal age” estimation. In the previous chapter, PCA was used to study the correlations among the components of multivariate data and to reduce redundancy by projecting the data over a proper basis. In this chapter, two neural network models are presented, one

with a given data set (full features) and the other with reduced dimensions (principal components). This new principal component neural network (PCNN) architecture [7] reduces the complexity of the model by minimizing the redundancy, hence increasing the overall efficiency in terms of computation.

5.2 ARTIFICIAL NEURAL NETWORKS (ANNs)

(Artificial) neural networks (ANNs), or simply neural networks (NNs), represent a modelling approach inspired by the action of neurons in the human brain. In the way that the human mind can solve complex tasks and learn through a complex network of interacting neurons, an ANN can be used to model complicated functions. More specifically, an ANN is a combination of artificial neurons that are interconnected via weights. The neurons are arranged in layers, starting with an input layer (the physical inputs) and ending with an output layer (the physical outputs). The ANNs can progressively learn from examples (data) to provide a mapping between inputs and outputs of an experiment (or simulation), and then produce reliable predictions for new inputs. In most cases (as is assumed here), information is propagated in one direction only, that is, from the input layer to the output layer, and neurons in the same layer are not connected by weights. This means that neurons in layer i provide 'inputs' to neurons in layer $i+1$, and so on. An activation function is associated with each neuron [8], [9]. This function acts on a weighted combination of the inputs from each neuron to which the neuron in question is connected. The output becomes the input to the neurons in subsequent layers.

An ANN can be categorized from the pattern of the interconnected neurons, the training method used (the way the weights are determined) and its activation functions. There are different approaches to constructing an ANN. A simple ANN can be seen in Figure 5.1, where I denotes the inputs of the network, w are the weights, a is the activation function and O is the output.

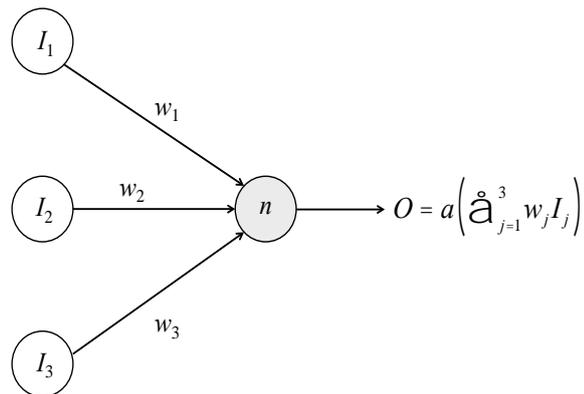


Figure 5.1: A simple neural network

Neuron n receives three separate inputs with different weights (levels of importance in terms of their effect on the output) attached to each. The neuron is associated with an activation function a , which acts on the input to the neuron, $w_1 I_1 + w_2 I_2 + w_3 I_3$. The output to this simple neural network is therefore

$$O = a\left(\hat{a}_{j=1}^3 w_j I_j\right).$$

In many cases the activation function is a non-decreasing function of the input to the neuron. Although there are other activation functions, the most often used are the sigmoid, Gaussian and sine function. In back-propagation neural networks, the log-sigmoid transfer function is convenient because it is differentiable. Using this transfer function, the input may have any finite value, while the range of the output

is (0, 1). The most frequently used functions in neural networks can be seen in Figure 5.2 below.

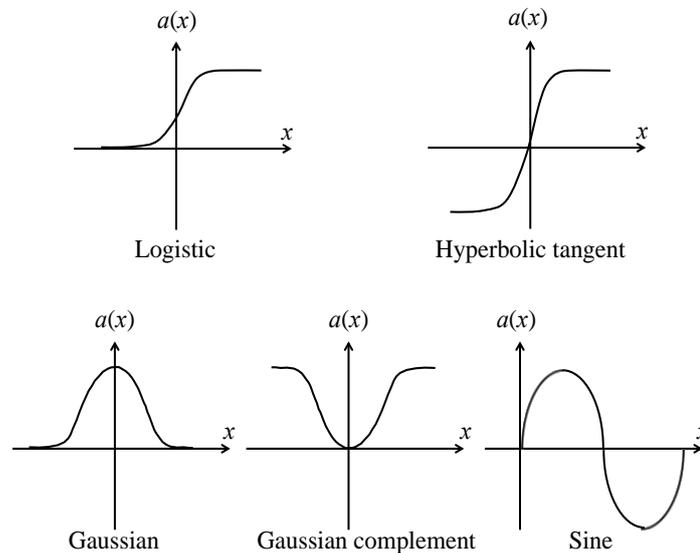


Figure 5.2: Some nonlinear neuron activation functions

These activation functions are nonlinear, which allows the neural network to learn nonlinear mappings between the inputs and outputs. Logistic functions are often used in neural networks to introduce nonlinearity in the model and/or to clamp signals within a specified range. The hyperbolic tangent function produces positive numbers between -1 and 1. The hyperbolic tangent activation function is most useful for training data that is between 0 and 1. Because the hyperbolic tangent activation function has a continuous derivative, it can be used with gradient descent based training methods. A Gaussian activation function can be used when finer control is needed over the activation range. It is the classic bell shaped curve, which maps high values into low values, and maps mid-range values onto high values. The Gaussian complement function tends to bring out meaningful characteristics in the extremes of the data. The sine activation function is rarely

used but can possibly be helpful in recognizing radially-distributed data. An ANN is usually used to solve complex tasks and to achieve this aim it can consist of several layers. Such networks are called *multi-layer neural networks*.

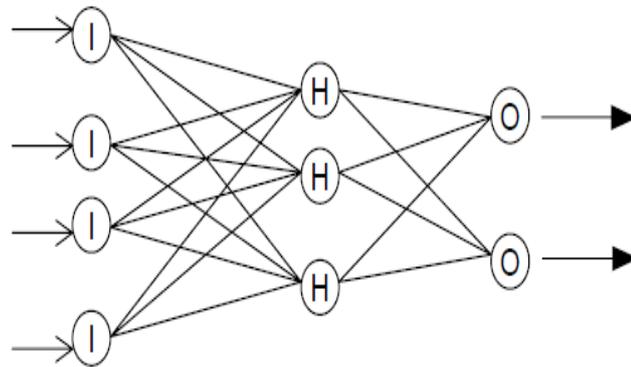


Figure 5.3: Multilayer perceptron model

The multilayer perceptron (MLP) model illustrated in Figure 5.3 is used to solve nonlinear classification tasks. It can be seen that there are three different types of neurons; input (I), hidden (H) and the output (O) neurons. Input neurons receive data from the environment. Output neurons provide the numerical output of the network and hidden layers process the inputs by multiplying each by a corresponding weight and summing these products (as before). In other words, the neurons organized into layers where the output of the first layer is the input of the second hidden layer. For instance, if the input is x_k , $k = 1, \dots, K$, its input is multiplied by a weight w_{ki} and then they summed, often with a bias θ_i yielding an output n_i . This output is used as an input for the activation function g . For this example it can be assumed the activation function is a hyperbolic tangent. Hence the output of node i is $y_i = g_i = g(\sum_{j=1}^K w_{ji}x_j + \theta_i)$. Connecting more nodes leads to a multi-layer perceptron network.

5.2.1 NEURAL NETWORK TOPOLOGIES

As mentioned above, neural networks can be categorized based on the pattern of the interconnected neurons. This leads to two different topologies. First, the feed-forward network, where the data processing goes from the input to the output without the existence of a feedback. Classical examples of feed-forward networks are the perceptron and the adaline networks. The second category belong the recurrent networks, which contain a feedback. Examples of recurrent ANNs are the Kohonen and Hopfield networks. Hence, a neuron model and the architecture of a neural network can illustrate how the network converts a number of inputs into an output.

5.2.2 TRAINING METHODS FOR A NEURAL NETWORK

An ANN has to be designed in a way that the application of a given data set at the input leads to the desired output. One method to achieve this is to set the weights of the network a priori while the other method is to train the network by introducing training parameters to adjust its weights according to a learning rule. The two main learning categories are the supervised learning method and the unsupervised methods.

In supervised learning methods, the neural network is fed with the inputs and the matched output where the matching can be achieved by an external supervisor. The training includes initially randomizing the weights and comparing each output with a given target (known value) for different inputs. Due to the random initial assignment of the weights the error initially large. Using the mean square error

(MSE) indicator $MSE = \frac{1}{2N} \sum_i^N (t_i - o_i)^2$ we can minimize the error to an acceptable predefined value by adjusting the weights (see later). In the mean square error indicator, N is the total number of training patterns, t_i denotes the target value and o_i represents the predicted output.

In unsupervised learning method the output reacts to a specific input pattern. This in turn means that there is no need for training data to clarify in which group each output belongs. The net modifies the weights to adjust the input vector with the best-fitted output.

5.2.3 *BACK-PROPAGATION*

Back-propagation learning, which is used to minimize the total squared error of the output, is the most common supervised learning rule in multilayer feed-forward neural networks. The training set is a dataset of desired outputs for different inputs. One of the example cases is fed to the network and produces an output, which in turn is compared with the desired output data and the mean square error is calculated. The calculated error is used to adjust the weights in each layer. The aforementioned process is repeated until the error becomes lower than a predefined tolerance.

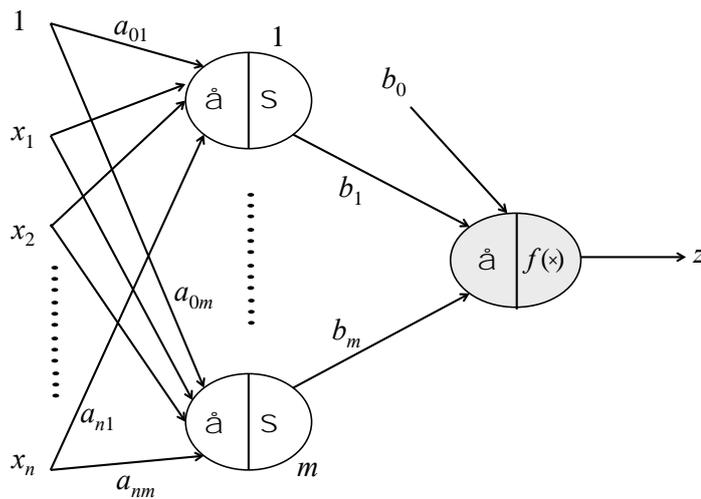


Figure 5.4: Multilayer perceptron and weights free parameters to be adapted

Considering the neural network in Figure 5.4, in back-propagation the weights of the hidden neurons and the output have to be adjusted simultaneously. To achieve this, it is necessary to calculate the error derivative of the weights. If the derivative w.r.t. an output weight is $\frac{\partial E}{\partial b}$ and the derivative w.r.t. the hidden neurons weights is $\frac{\partial E}{\partial a}$, the mean square error (E) is not directly linked to a and b , but it is linked to z (the network output), while z is related to the weighted sum of hidden neurons. Hence, the chain rule can be used to calculate the $\frac{\partial E}{\partial b}$. This process is called back-propagation.

Back-propagation can be used in batch training (where the weights are updated after processing all of the dataset) and in incremental training (where the weights are updated after each point in the dataset is presented to the network). In batch training, the gradients of all data points are stored and then the average or the resultant gradient is calculated. Using this gradient we can minimize the error in the direction of steepest decrease or some other gradient based method. The total

gradient can be expressed mathematically as $d_m = \hat{a}_{n=1}^N [\nabla E / \nabla w_m]_n$, where d_m denotes the total gradient, m is the epoch (one pass of the whole training set) and $\nabla E / \nabla w_m$ represents the gradient for each epoch.

Momentum is another method used for stability when weights reach their optimal value. In this method, the average of the past weights changes with every weight augmentation normalizing the net weight change. The basic principle of the momentum method is the usage of the exponential average of the previously calculated weights to route the present weight change. This process can be described mathematically as $\Delta w_m = \mu \Delta w_{m-1} - (1 - \mu) \varepsilon_m^w$, where μ denotes momentum, which varies between 0 and 1, Δw_{m-1} is the previous weight change. On the right part of the equation the second part represents the usual amount of weight change for epoch m considering the current total derivative. The term $(1 - \mu)$ indicates the link between the current weight and the past change. To conclude, momentum stabilizes the learning process in a way that if the previous change was in the same direction as the current direction, the weight change is accelerated. On the other hand, if the previous change was in the opposite direction, the current change is restrained.

Typically in feed-forward neural networks the batch training is preferred since there is no need of a constant learning rate. A low learning rate can affect the performance of the network because of the slow learning process. On the other hand, a high learning rate may results to no training at all.

5.2.4 *APPLICATIONS OF NEURAL NETWORKS*

The concept of ANNs is around fifty years old but only gained widespread adoption in the last twenty years, coinciding with developments in computer hardware during the same period. The number of applications continues to rise. Due to their ability to model complex and highly nonlinear systems, ANNs have found use in many areas. For example, neural networks can be used in signal processing to remove the echo from an output signal, the noise in a telephone line and at the same time they are used as echo cancellers in satellite links [15].

Many of the ANN applications are related to pattern recognition, such as voice, face and hand writing recognition. They can also be applied to digital signal processing for analysing, compression and video encoding. In the same field they can be used to track the orbit of a subject or even event recognition in a video. ANNs can be used in finance and operational research to model and predict times series of shareholder and financial indicators. They have been used extensively in robotics and space technology. Robotics systems that search for independent planets are based on ANNs; perform a variety of different processes. In industry they can be used to help in the detection and rejection of faulty parts in a supply chain [16]. This brief list mentions only few of the fields in which neural networks have been applied (others include meteorology, bioengineering and civil engineering). A more complete list is beyond the scope of this thesis.

5.2.5 BACK-PROPAGATION WITH LEVENBERG-MARQUARDT ALGORITHM

The back-propagation training method was used in this work, along with the Levenberg-Marquardt optimization algorithm [17]. The Levenberg-Marquardt algorithm can achieve a second order speed to train the network without the need of calculating the Hessian matrix $H = J^T J$, where J is the Jacobian matrix of the first order sum square errors. The derivation of the Levenberg-Marquardt optimization algorithm is divided into four parts: the steepest descent algorithm, Newton's Method, Gauss-Newton's method and finally Levenberg-Marquardt. To begin with, the steepest descent algorithm is a first order algorithm meaning it calculates first order derivatives of the error. Let the gradient be $g = \partial E(x, w) / \partial w = [\partial E / \partial w_1, \partial E / \partial w_2, \dots, \partial E / \partial w_N]^T$, where E is the error, w denotes the weight and x is the input vector. Taking into account the definition of the gradient, the update rule for the algorithm is $w_{k+1} = w_k - s g_k$, where s is the step size. In Newton's algorithm, the gradient for each neuron can be expressed as:

$$\begin{cases} g_1 = g_{1,0} + \frac{\partial g_1}{\partial w_1} \Delta w_1 + \frac{\partial g_1}{\partial w_2} \Delta w_2 + \dots + \frac{\partial g_1}{\partial w_N} \Delta w_N \\ \vdots \\ g_N = g_{N,0} + \frac{\partial g_N}{\partial w_1} \Delta w_1 + \frac{\partial g_N}{\partial w_2} \Delta w_2 + \dots + \frac{\partial g_N}{\partial w_N} \Delta w_N \end{cases} \quad (5.1)$$

The gradient definition and the above vector can be combined to provide the second order derivative of the total square error

$$\nabla g_i / \nabla w_j = \nabla \left(\nabla E / \nabla w_j \right) / \nabla w_j = \nabla^2 E / \left(\nabla w_i \nabla w_j \right). \text{ Hence:}$$

$$\begin{cases} g_1 = g_{1,0} + \frac{\partial^2 E}{\partial w_1^2} \Delta w_1 + \frac{\partial^2 E}{\partial w_1 \partial w_2} \Delta w_2 + \dots + \frac{\partial^2 E}{\partial w_1 \partial w_N} \Delta w_N \\ \vdots \\ g_N = g_{N,0} + \frac{\partial^2 E}{\partial w_N \partial w_1} \Delta w_1 + \frac{\partial^2 E}{\partial w_N \partial w_2} \Delta w_2 + \dots + \frac{\partial^2 E}{\partial w_N^2} \Delta w_N \end{cases} \quad (5.2)$$

If it is assumed that the gradient is zero (to obtain the minimum of the sum error function E) the above equation can be written as:

$$\begin{cases} 0 = g_{1,0} + \frac{\partial^2 E}{\partial w_1^2} \Delta w_1 + \frac{\partial^2 E}{\partial w_1 \partial w_2} \Delta w_2 + \dots + \frac{\partial^2 E}{\partial w_1 \partial w_N} \Delta w_N \\ \vdots \\ 0 = g_{N,0} + \frac{\partial^2 E}{\partial w_N \partial w_1} \Delta w_1 + \frac{\partial^2 E}{\partial w_N \partial w_2} \Delta w_2 + \dots + \frac{\partial^2 E}{\partial w_N^2} \Delta w_N \end{cases} \quad (5.3)$$

Since there are N parameters, there are N equations, and, therefore, the value for each weight is calculated and expressed in a matrix form:

$$\begin{bmatrix} -g_1 \\ \vdots \\ -g_N \end{bmatrix} = \begin{bmatrix} -\frac{\partial E}{\partial w_1} \\ \vdots \\ -\frac{\partial E}{\partial w_N} \end{bmatrix} = \begin{bmatrix} \frac{\partial^2 E}{\partial w_1^2} & \dots & \frac{\partial^2 E}{\partial w_1 \partial w_N} \\ \vdots & \dots & \vdots \\ \frac{\partial^2 E}{\partial w_N \partial w_1} & \dots & \frac{\partial^2 E}{\partial w_N^2} \end{bmatrix} X \begin{bmatrix} \Delta w_1 \\ \vdots \\ \Delta w_N \end{bmatrix} \quad (5.4)$$

From this point the Hessian matrix can be calculated:

$$H = \begin{bmatrix} \frac{\partial^2 E}{\partial w_1^2} & \dots & \frac{\partial^2 E}{\partial w_1 \partial w_N} \\ \vdots & \dots & \vdots \\ \frac{\partial^2 E}{\partial w_N \partial w_1} & \dots & \frac{\partial^2 E}{\partial w_N^2} \end{bmatrix} \quad (5.5)$$

So the update rule for this method is $w_{k+1} = w_k - H_k^{-1} g_k$. Comparing the update rule of Newton's and the steepest descent algorithm it can be noted that the better approximation of Newton's method is due to the use of the inverse Hessian matrix in the calculation of the weights.

Although Newton's method gives better results, the calculation of the Hessian matrix is complicated due to the use of second order derivatives of the error function. The Gauss Newton method introduces the Jacobian matrix to make the calculation process simpler. The Jacobian matrix has the following form:

$$J = \begin{bmatrix} \frac{\partial e_{1,1}}{\partial w_1} & \dots & \frac{\partial e_{1,1}}{\partial w_N} \\ \vdots & \dots & \vdots \\ \frac{\partial e_{P,M}}{\partial w_1} & \dots & \frac{\partial e_{P,M}}{\partial w_N} \end{bmatrix} \quad (5.6)$$

Where $e_{P,M}$ denotes the training error of the M outputs for the applied pattern P ; in other words $e_{P,M} = d_{P,M} - o_{p,m}$, where d is the desired output and o is the original output.

In the Gauss-Newton method, the gradient is given by the following formula:

$$g_i = \frac{\partial E}{\partial w_i} = \frac{\partial \left(\frac{1}{2} \sum_{p=1}^P \sum_{m=1}^M e_{p,m}^2 \right)}{\partial w_i} = \sum_{p=1}^P \sum_{m=1}^M \left(\frac{\partial e_{p,m}}{\partial w_i} e_{p,m} \right) \quad (5.7)$$

The relationship between the Jacobian matrix and the gradient is $g = Je$, while the update rule for Gauss-Newton method is $w_{k+1} = w_k - (J_k^T J_k)^{-1} J_k e_k$. It is obvious that this method does not require the second derivative of the total square error due to the use of the Jacobian matrix. On the other hand, the Gauss-Newton method suffers from the same convergence problem as that of the Newton algorithm.

The Levenberg-Marquardt algorithm [17], [18] is based on the Newton algorithm and it is the most effective optimization technique for fast training of a neural network. This technique can calculate the Jacobian matrix without calculating the

complex Hessian matrix [19]. In other words, the Levenberg-Marquardt algorithm uses the above approximation of the Hessian matrix and it can be expressed as $\chi_{k+1} = \chi_k - [J^T J + \mu I]^{-1} J^T e$, where μ denotes a zero scalar and e is the error matrix of the network.

More specifically, for the calculation of the Jacobian matrix, the following steps are involved. The Jacobian matrix is in the following form: $(i, j) = \partial e_i / \partial w_j$, where e denotes the error and w is the weight. The error can be defined as the difference between the desired output and the calculated output given by the network:

$$e_k = o_d - o_k \quad (5.8)$$

where e_k represents the error, o_d is the desired output and o_k denotes the actual output. For every iteration from k to $k - 1$, there is a change in the outputs (due to the adjustment of the weights) equal to:

$$\partial o = o_d - e_k - o_{k-1} \quad (5.9)$$

Combining the last two equations, the relationship between ∂o and ∂e is:

$$\partial o = o_d - o_{k-1} - e_k = e_{k-1} - e_k \Rightarrow \partial o = -\partial e \quad (5.10)$$

From this point, the Jacobian matrix can be written using the output change instead of the total squared error in the form $J(i, j) = -\partial o_i / \partial w_j$ or $o = f(net)$, where

$$net = w_1 x_1 + w_2 x_2 + \dots + w_n x_n \quad (5.11)$$

Taking the last equation into account, the Jacobian matrix can be written as:

$$-\frac{\partial o_i}{\partial w_j} = -\frac{\partial o_j}{\partial net} \frac{\partial net}{\partial w_j} \Leftrightarrow \frac{\partial net}{\partial w_j} = x_j \quad (5.12)$$

Hence, the final form of the Jacobian matrix used in the Levenberg-Marquardt algorithm is:

$$J(i, j) = -\frac{\partial o_i}{\partial net} x_j \quad (5.13)$$

This algorithm is one of the fastest training methods for feed-forward neural networks with a number of weights up to several hundreds. The limitation of this algorithm is that due to the use of Jacobian matrix it can be used only to calculate the mean of the squared error.

5.3 RESULTS OF NEURAL NETWORK

MODELLING

The proposed model is based on a two-layer, feedforward ANN, trained using backpropagation (Levenberg-Marquardt algorithm for the rules update). The vector-valued inputs in \mathbb{R}^{24} were the measured spine features and the (scalar) outputs were the spinal ages. The number of neurons in the hidden layer was set to 10 as discussed below. The workflow for modelling the system was divided into seven main steps as: collecting the data, creating the network, configuring the network, initializing the weights and biases, training the network, validating the network and hence using the network.

61 samples were used to train, validate and test the neural network model. Input vectors and corresponding targets were randomly divided into three sets. As the sample size is very small, so the major portion of the samples were used for training the network, so that it can learn better. An equal number of samples were used for validation and testing. The division of data is as following:

- 70% of the data (43 samples) were used for training. These were presented to the network during training, and the network was adjusted according to Levenberg-Marquardt
- 15% of the data (9 samples) was used for validation. This data was used to measure network generalization error, and to halt training when generalization error fell below a set tolerance
- The last 15% (9 samples) were used as a completely independent test of network generalization. This data had no effect on training and so provided an independent measure of network performance after training

For the initial trial, the number of hidden neurons was set to 10. The training was automatically terminated when the change in the generalization error fell below a set tolerance [10], which is indicated by an increase in the mean square error of the validation samples. There are some empirically-derived rules-of-thumb; the most used is the optimal size of the hidden layer is usually between the size of the input and size of the output layers [11]. For most problems, good performance is usually achieved by setting the hidden layer configuration using the following rules:

(i) The number of hidden layers equals one; and

(ii) The number of neurons in that layer is the mean of the neurons in the input and output layers.

There are many other methods for determining the number of neurons to use in the hidden layers, such as: the number of hidden neurons should be $2/3$ the size of the input layer, plus the size of the output layer [12]. In most of the situations, there is no theoretical way to determine the best number of hidden units without training several networks and estimating the generalization error for each one. If there are too few hidden units, then there will be a high training error and high generalization error due to under fitting and high statistical bias. If there are too many hidden units, then there will be low training error but high generalization error due to over fitting and high variance.

Some studies have attempted to explain the effect of the number of hidden units on the bias/variance trade-off [13]. The selection of hidden units mainly depends on: the numbers of input and output units, the number of training cases, the amount of noise in the targets, the complexity of the function or classification to be learned, the architecture, the type of hidden unit activation function, and the training algorithm [14]. In this thesis, two neural network models were presented with 10 and 12 hidden neurons and the performance was evaluated for both the models. These models were further cross validated by using repeated k-fold cross validation. All models were implemented in the Matlab Neural Network Toolbox.

5.3.1 SPINAL AGE ESTIMATION

The model presented here does not incorporate PCA or dimensionality reduction. Figure 5.5 below shows the random distribution of samples for neural network modeling labeled with their corresponding age. Roughly, there are 5-6 samples of each age cluster. The youngest person with lumbar MRI scans is 2 years old and the oldest is 93 years old.

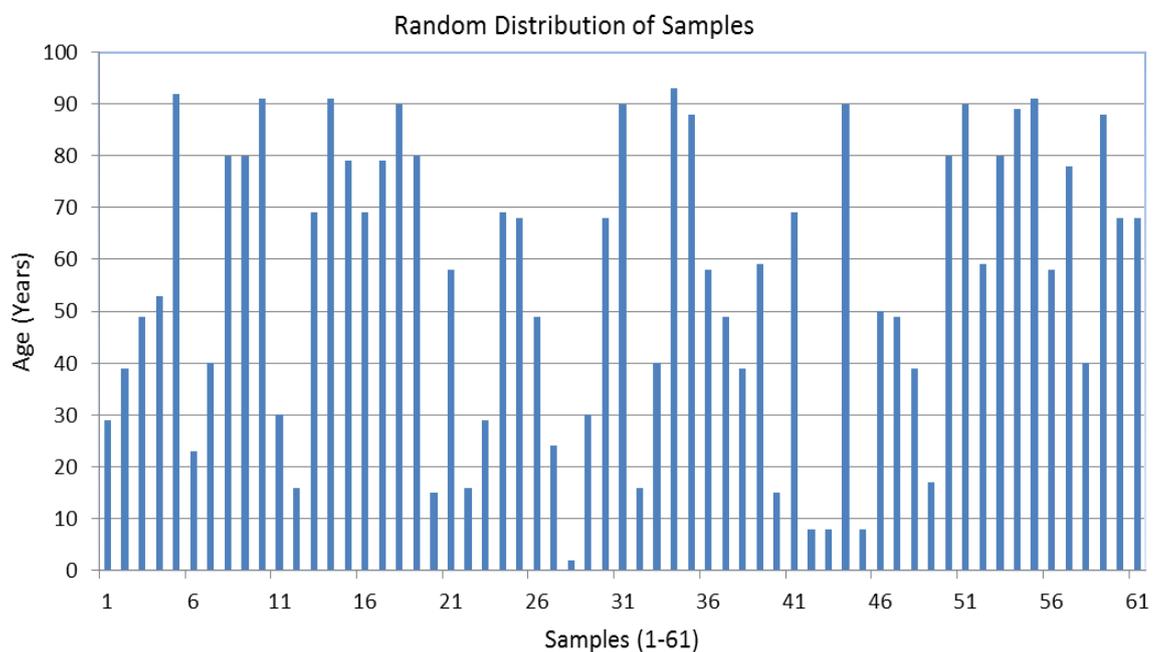


Figure 5.5: Random distribution of samples for neural network modeling

The performance of the ANNs changed with the number of hidden neuron. After several trials by changing the number of hidden neurons, a good approximation was obtained with 10 neurons. The network was trained using Levenberg-Marquardt [17]. This method is scale invariant, so no standardization of the inputs and data was required [18]. The performance of the network was measured in terms of mean squared Error (MSE).

In first model, “spinal age” is predicted from 24 spinal features by using one hidden layer and 10 hidden neurons. The training and performance states are shown below in figure 5.6, in which the blue line shows the training mse, green line shows the validation mse and a red line shows testing mse.

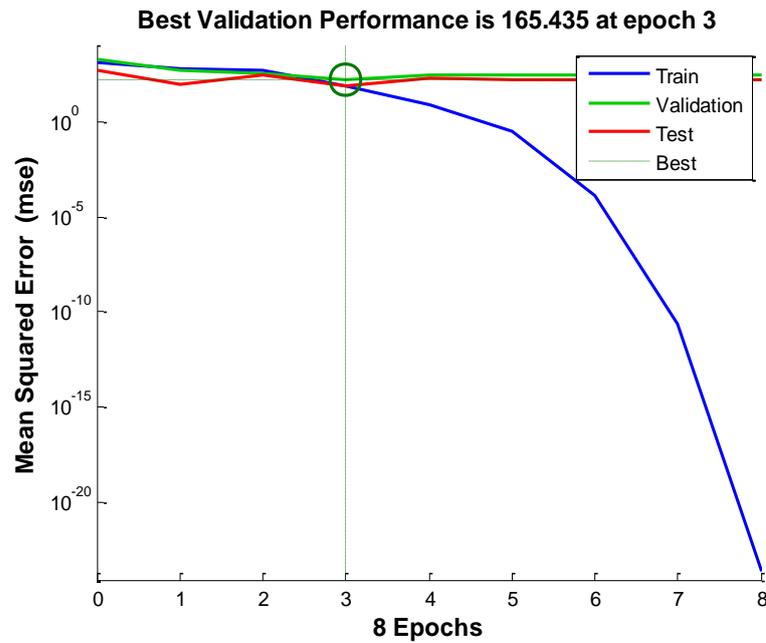


Figure 5.6: (a) Mean square error of neural network model

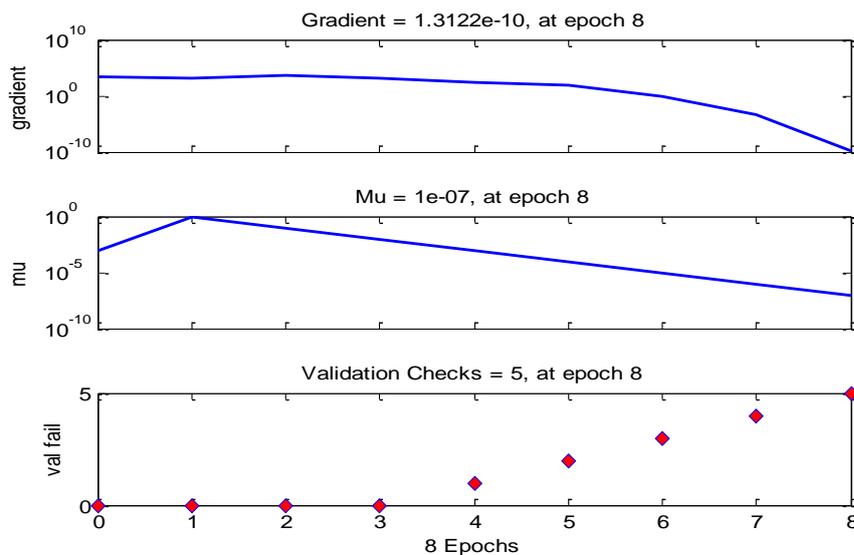


Figure 5.6: (b) Neural network training state

Figure 5.7, shows the regression of the neural network model. The blue line shows the regression of training set which is $R=0.97124$ and a green line shows the regression of validation set which is $R=0.93328$ and the red line shows the regression of testing set which is $R=0.9474$. The overall regression was $R=0.94471$.

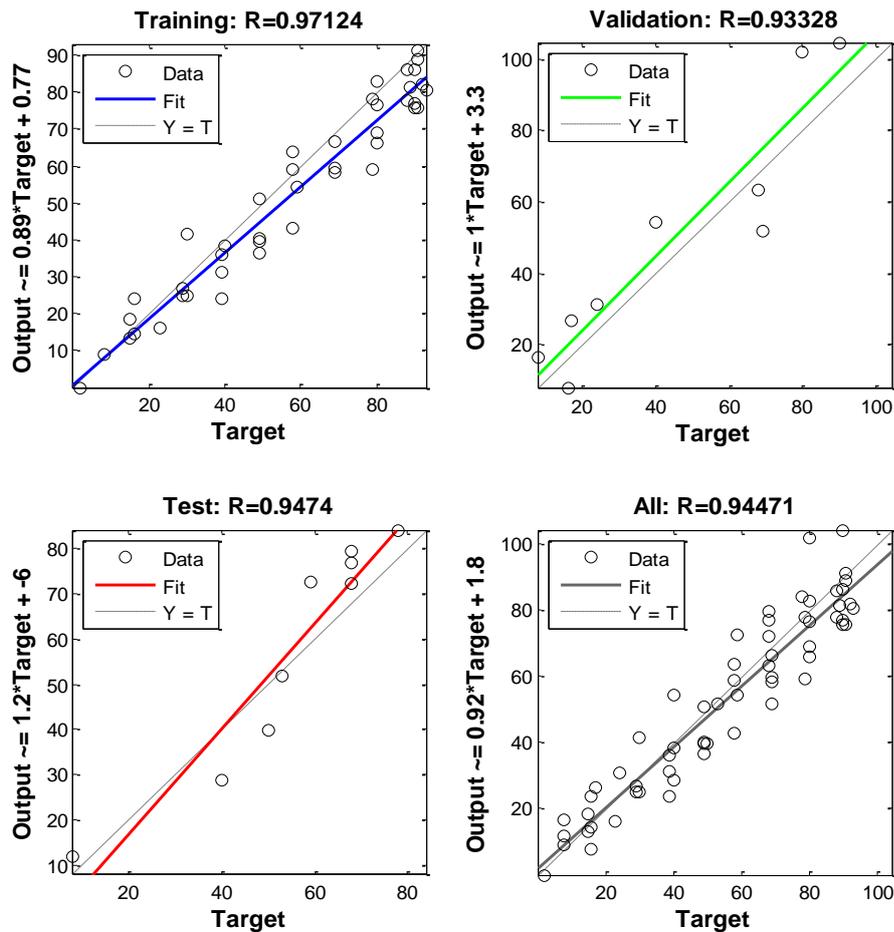


Figure 5.7: Regression of first neural network model

Figure 5.8 below, shows the error histogram. It is the difference between actual and the neural network estimated values. It can be seen clearly that most (roughly 90%) of the samples (out of 61) are estimated within a ± 15 year margin. This error histogram is helpful in the outlier detection by filtering out the samples that are different from rest of the data.

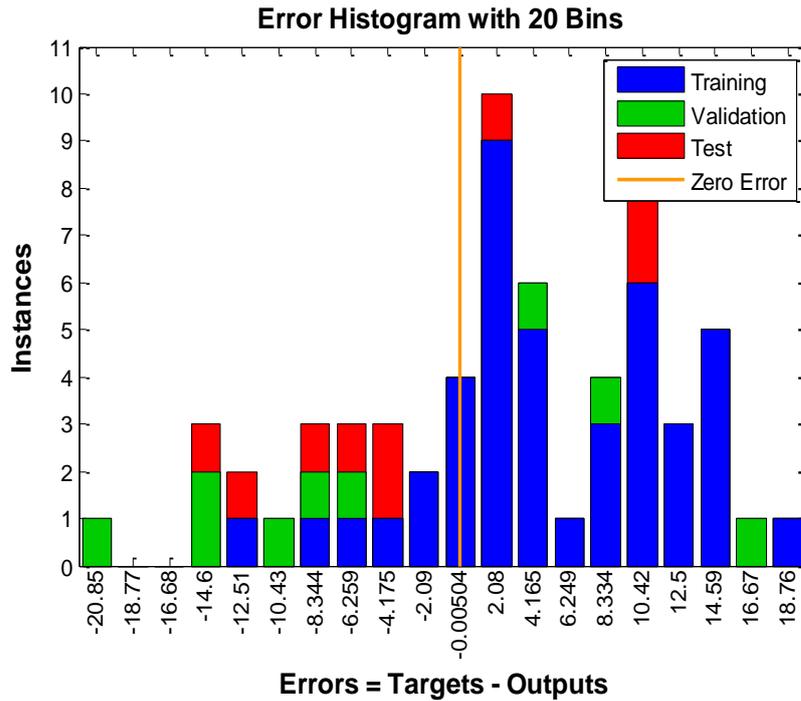


Figure 5.8: Error histogram of the neural network estimation

In this second model, the number of hidden neurons was changed to 12. The training state and the performance of this model is illustrated in figure 5.9 below.

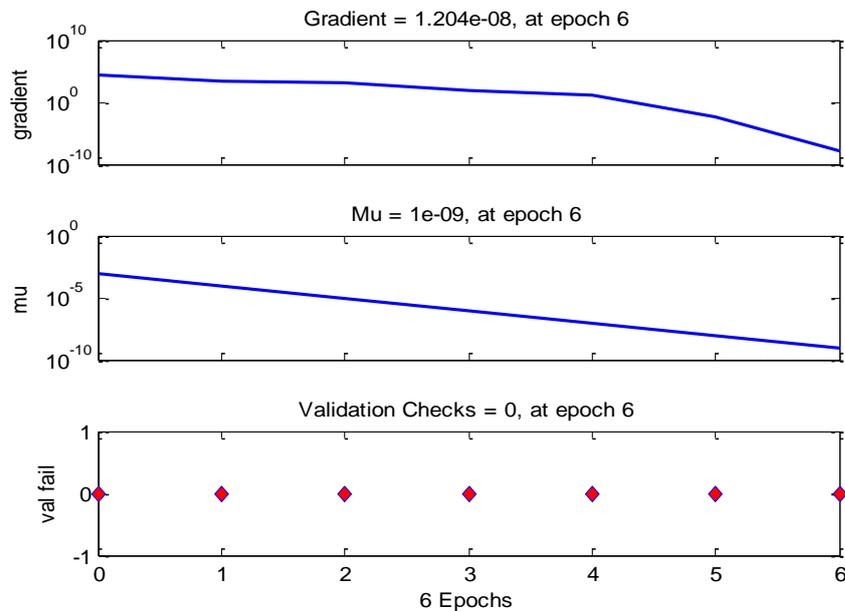


Figure 5.9: (a) Training state of neural network

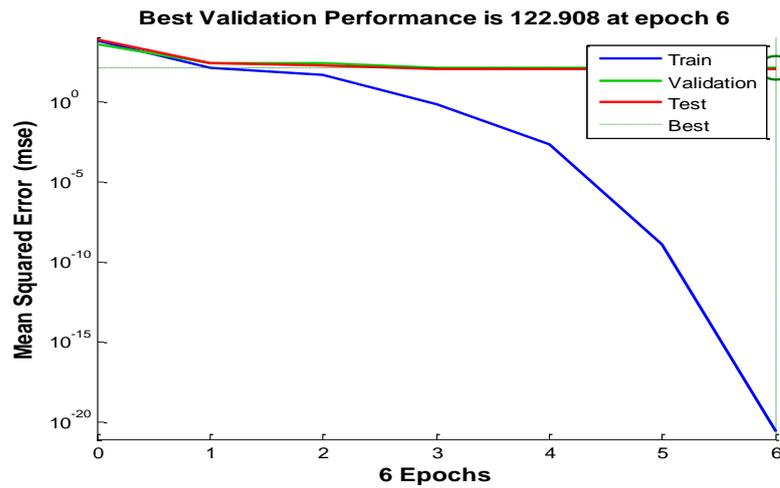


Figure 5.9: (b) Performance

In this model, regression of training set is ($R=1$), regression of validation set is ($R=0.91313$) and regression of testing set is ($R=0.89249$). The overall regression (considering all 61 samples) was $R=0.9786$. These regressions are shown in figure 5.10 below. This model shows a better overall regression than the previous one.

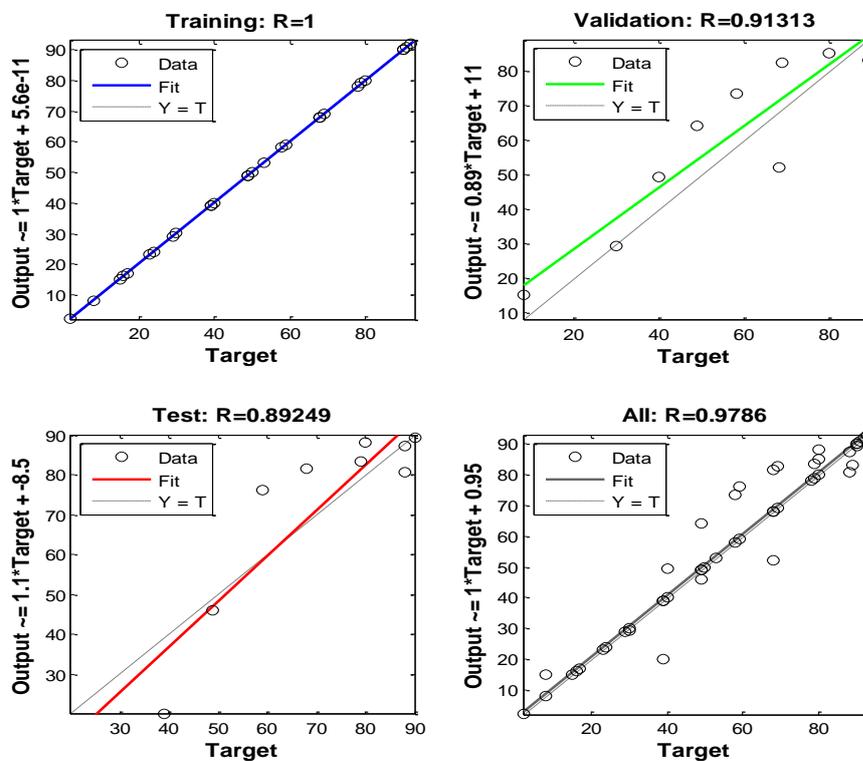


Figure 5.10: Regression of the second neural network model

The error histogram of this model is shown in figure 5.11 below.

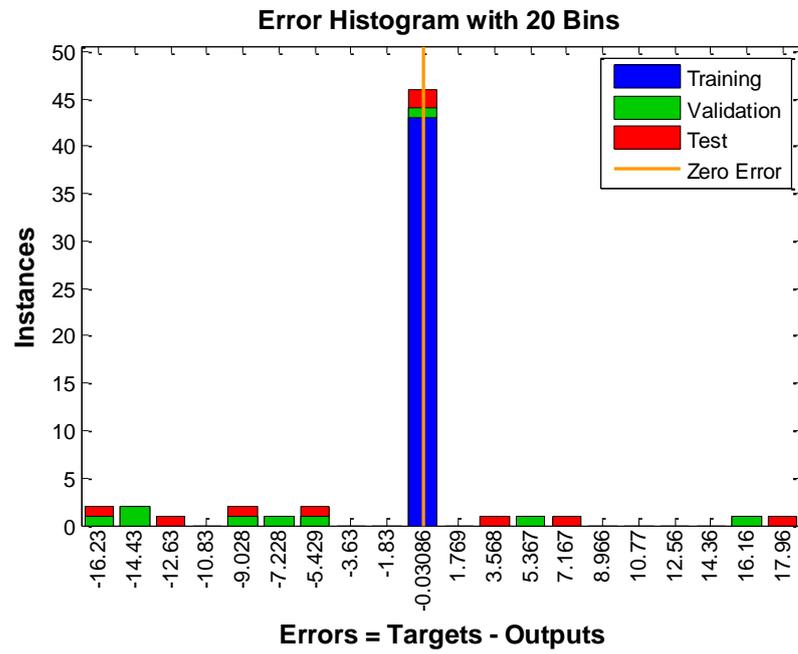


Figure 5.11: (a) Error histogram of model with 12 neurons

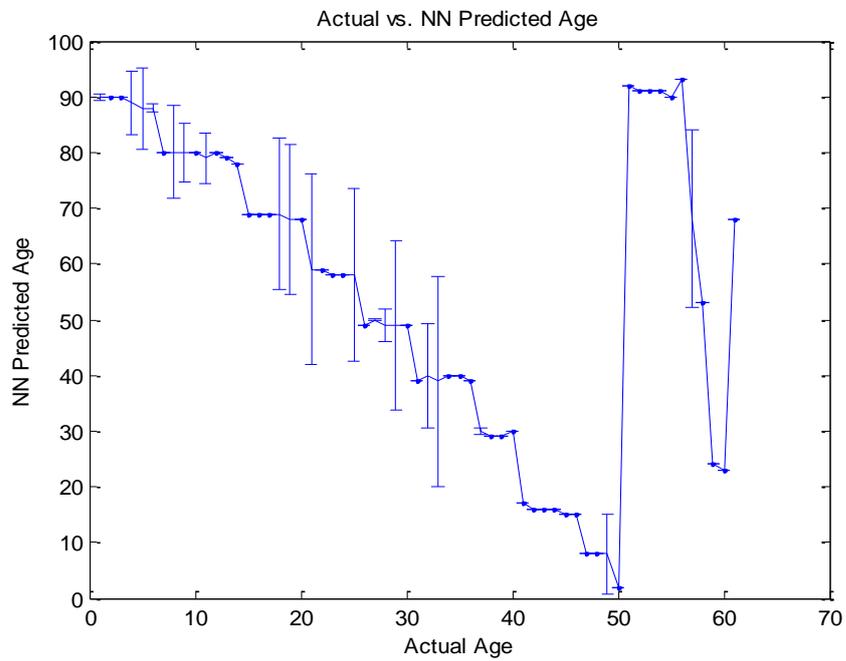


Figure 5.11: (b) Actual vs. ANN estimated spinal age

5.3.2 GENDER ESTIMATION FROM SPINAL FEATURES

It was found in PCA analysis that male and female spine differs in their characteristics. Notably, male samples were found having slightly higher vertebral heights as compared to the female samples of the same age. In this section a feedforward neural network was constructed with 24 input features and one output to predict gender on the basis of given spinal features. This gender prediction model has the same learning method and performance measures as the previously discussed model. One hidden layer and 12 hidden neurons were used. In the model, the output gender is assigned numerical values, with 1 representing female and 2 representing male. The performance and training states are shown below in figure 5.12.

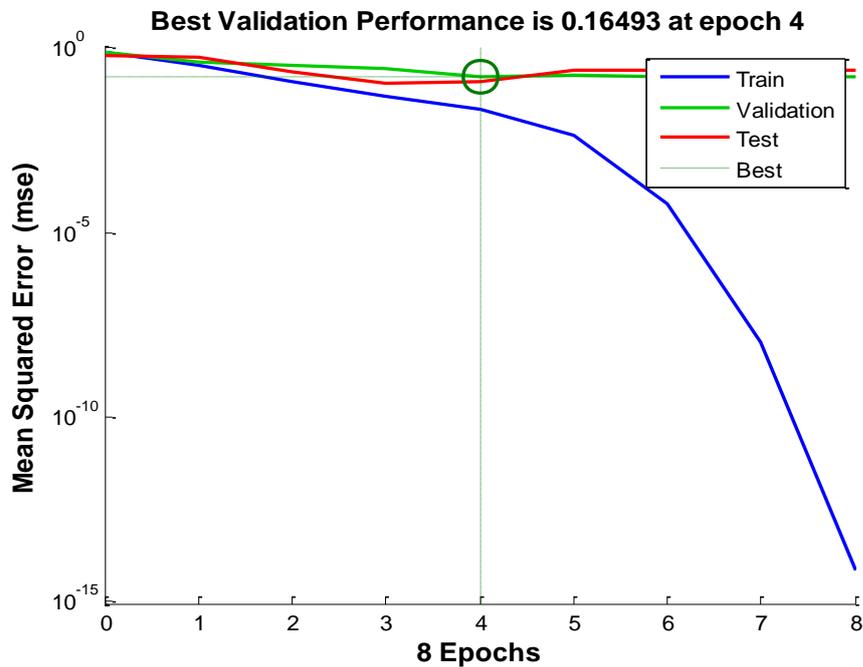


Figure 5.12: (a) Performance of gender estimation model

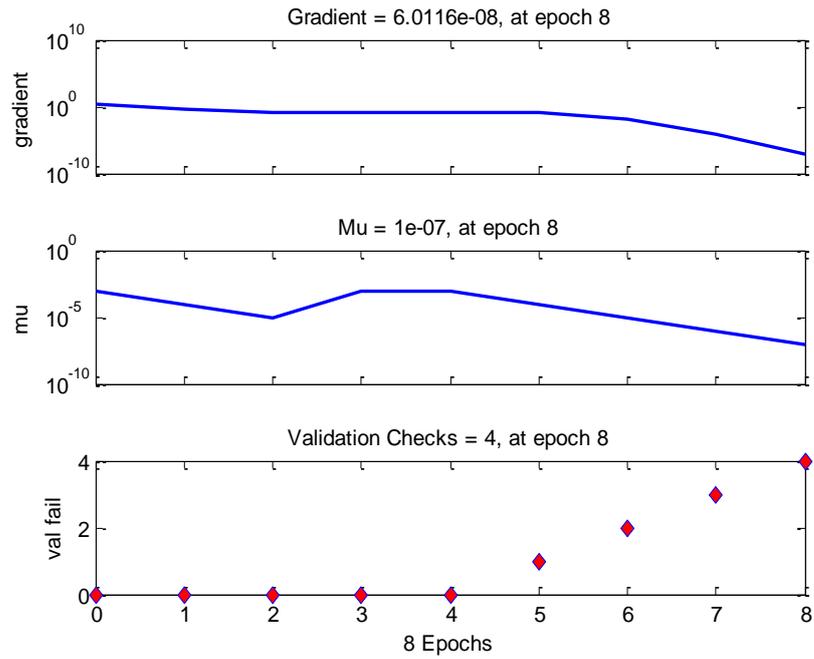


Figure 5.12: (b) Training state of gender estimation model

In this model, regression of the training set is ($R=0.96389$), the regression of the validation set is ($R=0.61622$) and the regression of the testing set is ($R=0.70727$). The overall regression was $R=0.8793$. These regressions are shown in figure 5.13.

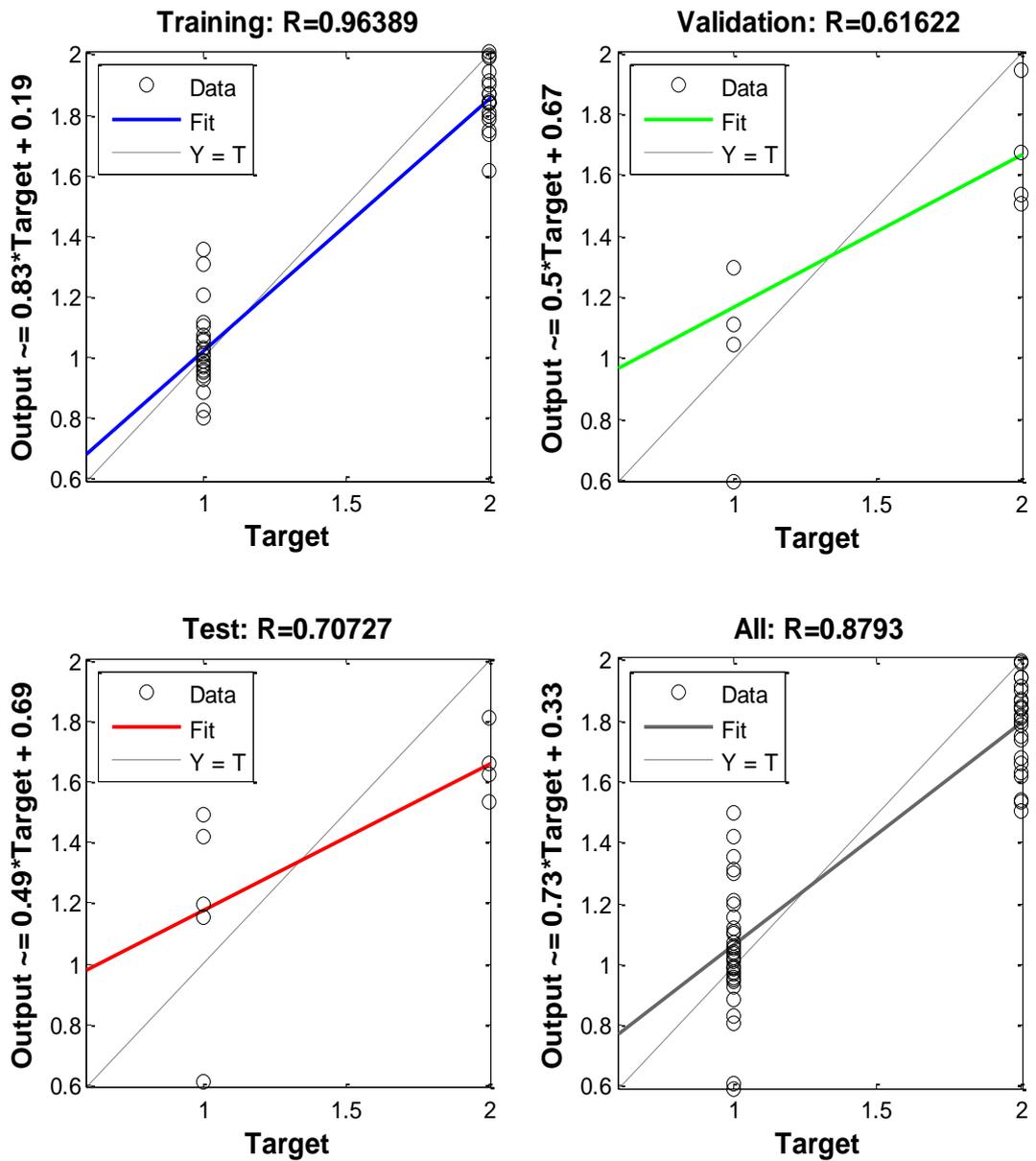


Figure 5.13: Regression of gender predicting neural network

The error histogram is shown below in figure 5.14. Here, any sample having value less than or equal to 1.5 is classified as female whereas samples having values greater than 1.5 are classified as male. The results show that 58 out of 61 samples are classified correctly, yielding a model accuracy of 95%.

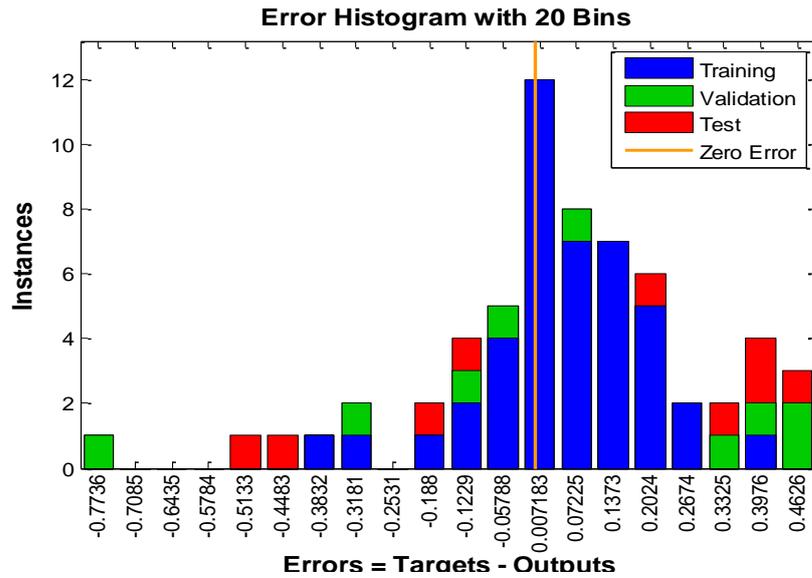


Figure 5.14: Error histogram for gender prediction neural network

5.3.3 BOTH AGE AND GENDER ESTIMATION

In the final model, both age and gender are predicted simultaneously from the spinal features. The ANN structure is the same as before (24 inputs, one hidden layer, 12 hidden neurons) except the number of outputs is now 2. This model is built with 24 inputs, one hidden layer, 12 hidden neurons, and 2 outputs. The performance and training stated are shown in figure 5.15.

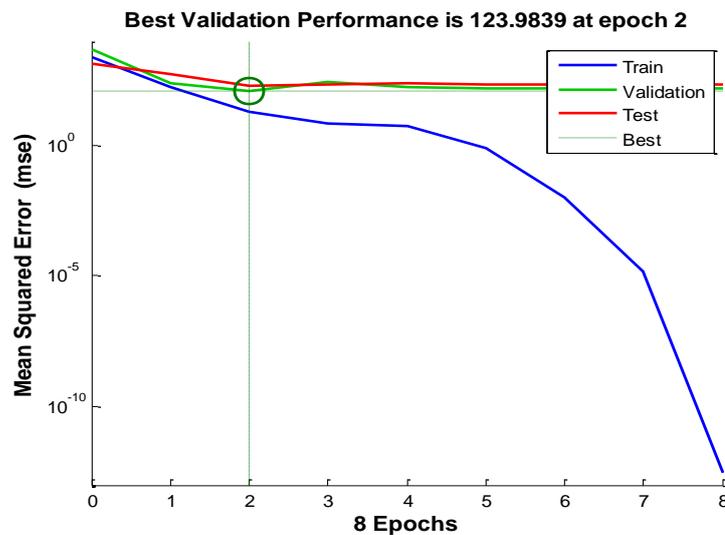


Figure 5.15: (a) Performance of age-gender model

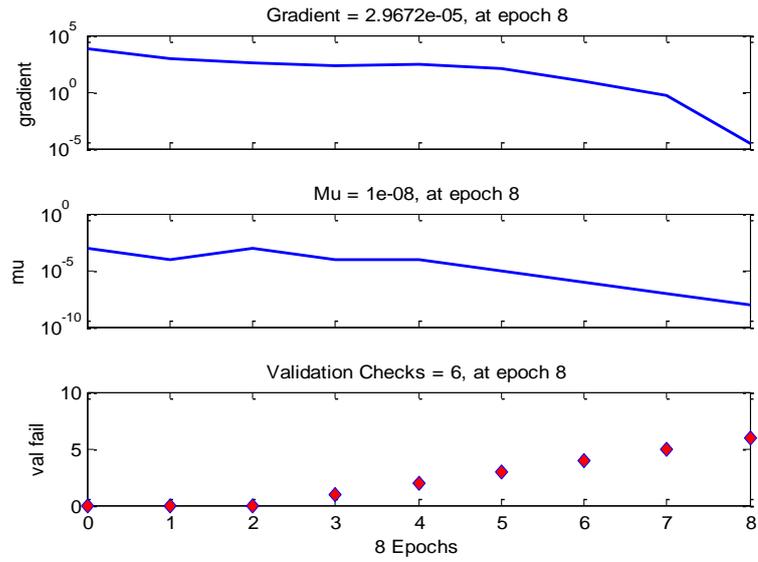


Figure 5.15: (b) Training state of age-gender model

The regression of network is shown in figure 5.16.

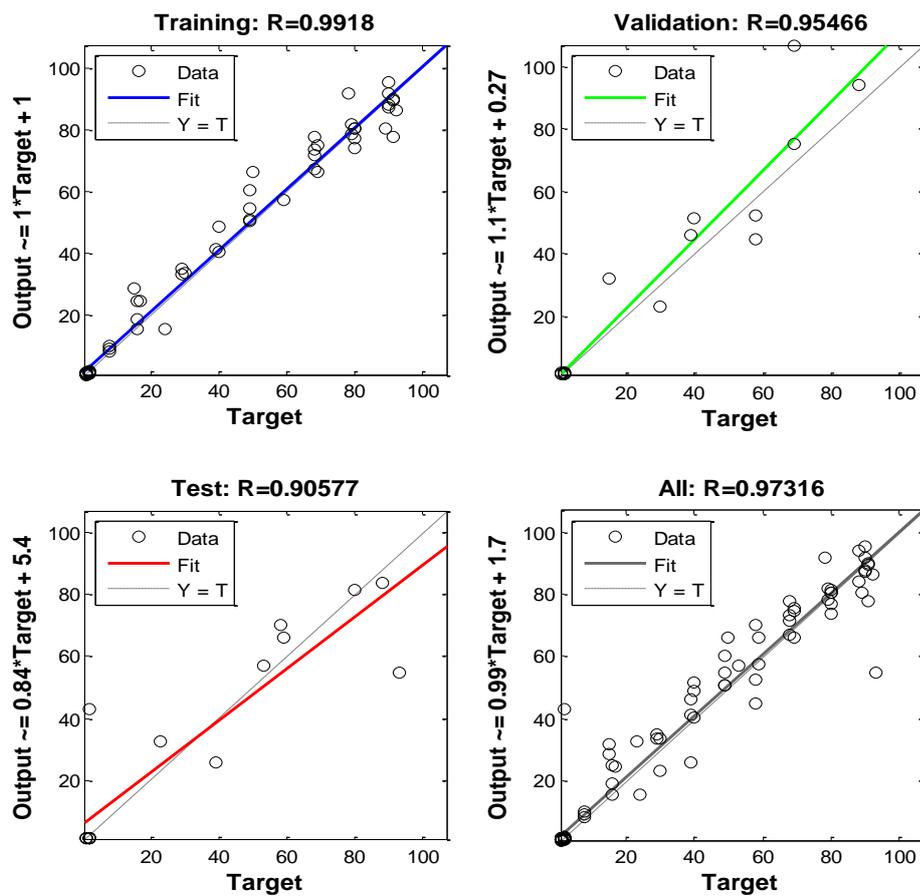


Figure 5.16: Regression of age-gender predicting neural network

In this model, the regression of the training set was ($R=0.9918$), the regression of the validation set was ($R=0.95466$) and the regression of the testing set was ($R=0.90577$) as shown in figure 5.16. The overall regression (considering all 61 samples) was $R=0.97317$. The error histogram of this network output is shown in figure 5.17 below. It can be seen that only three samples are not predicted satisfactorily (having difference of -38.9 and 36.23) while all other samples are predicted with considerable accuracy of (± 15) years.

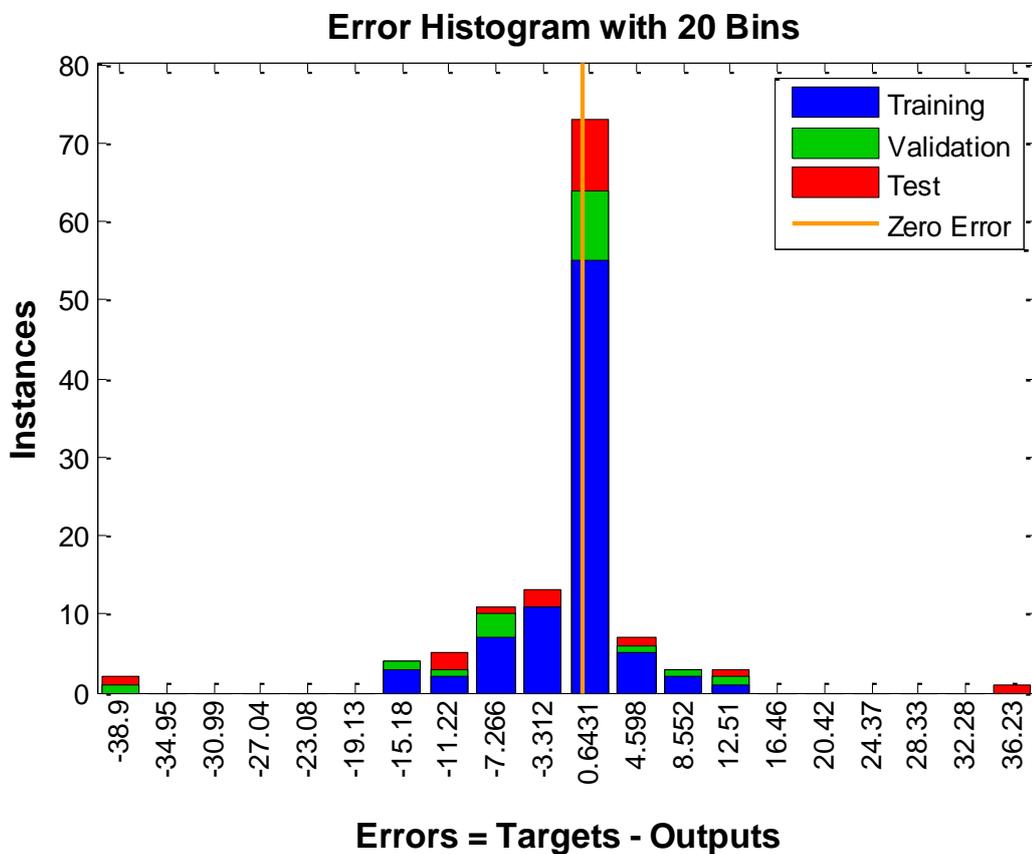


Figure 5.17: Error histogram for an age-gender predicting neural network

5.4 CROSS-VALIDATION

Cross validation is one of many statistical approaches to evaluating and comparing learning algorithms by dividing a large data set into segments. To train and validate

the model, several cross-overs need to be performed, so each set of data will be validated against the test set.

5.4.1 CROSS-VALIDATION TECHNIQUES

There are different forms of cross-validation such as k-fold, leave-one-out and repeated random sub-sampling cross-validation, each one with advantages and disadvantages.

K-fold cross-validation

In this type of cross-validation the original data set is divided into K random equal subsets. From these sub-sets one is selected to be used as the test dataset while the rest will be used to train the model. The process is repeated for the number of folds (K) as shown in figure 5.18. The main advantage of K-fold of cross-validation is that all of the samples are used to train and validate the model.

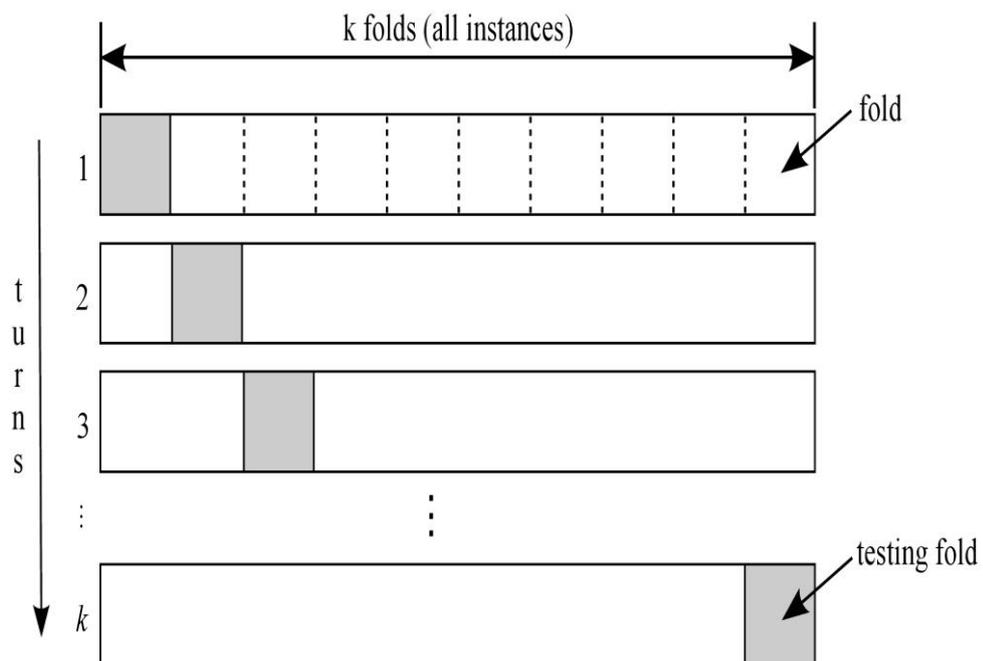


Figure 5.18: K-fold cross-validation

Leave-one-out cross-validation

Another case of K-fold cross-validation is the Leave-One-Out method where all the data points (samples) are used for training except one that is used for testing the model. This method, due to the high variance of the estimation, is not reliable.

Repeated random sub-sampling validation

In this method, in each round of cross-validation the data is reshuffled then the whole process is repeated. The data is randomly divided into test and training samples and for each division the model is fitted to the training data. The main advantage of this method is that its accuracy does not depend on the number of folds.

Resubstitution validation

In this method all the data are used for both the training and testing process. The validation uses all the available data but obviously may suffer from over fitting.

5.4.2 REPEATED K-FOLD CROSS-VALIDATION

To train and validate the neural network model, a repeated 10-fold cross validation was employed in this thesis. The anomaly identified in the previous chapter was removed and the remaining 60 samples were randomly divided into 10 equal subgroups each having 6 samples. From these 10 subgroups one was selected as the test dataset, while the rest were used to train the model. The process is repeated for all of the 10 folds. The difference between the actual age and neural network predicted age was recorded in each fold. The results from the first iteration (repetition) of the repeated 10 fold cross validation are presented in table 5.1.

Table 5.1: Results from first iterations of 10-fold cross validation

Folds	Age	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6	MSE	RMSE
Fold 1	Actual Age	39.00	80.00	69.00	58.00	16.00	68.00	90.59	9.52
	NN Predicted Age	34.00	61.83	73.57	50.52	26.55	68.59		
	Error	5.00	18.17	-4.57	7.48	-10.55	-0.59		
Fold 2	Actual Age	2.00	24.00	17.00	91.00	8.00	23.00	107.22	10.35
	NN Predicted Age	-14.15	23.06	24.84	76.94	0.06	30.70		
	Error	16.15	0.94	-7.84	14.06	7.94	-7.70		
Fold 3	Actual Age	68.00	16.00	91.00	16.00	92.00	39.00	98.47	9.92
	NN Predicted Age	77.92	7.51	83.02	17.63	94.40	57.66		
	Error	-9.92	8.49	7.98	-1.63	-2.40	-18.66		
Fold 4	Actual Age	49.00	29.00	68.00	80.00	89.00	29.00	95.60	9.78
	NN Predicted Age	37.04	43.84	62.80	86.84	93.34	39.86		
	Error	11.96	-14.84	5.20	-6.84	-4.34	-10.86		
Fold 5	Actual Age	90.00	50.00	40.00	8.00	39.00	79.00	110.89	10.53
	NN Predicted Age	78.06	52.56	61.93	13.80	39.38	77.81		
	Error	11.94	-2.56	-21.93	-5.80	-0.38	1.19		
Fold 6	Actual Age	69.00	80.00	15.00	49.00	30.00	78.00	89.94	9.48
	NN Predicted Age	74.84	63.79	17.05	59.94	32.97	67.52		
	Error	-5.84	16.21	-2.05	-10.94	-2.97	10.48		
Fold 7	Actual Age	49.00	88.00	69.00	90.00	88.00	30.00	196.99	14.04
	NN Predicted Age	58.55	70.32	80.02	75.61	71.37	43.16		
	Error	-9.55	17.68	-11.02	14.39	16.63	-13.16		
Fold 8	Actual Age	58.00	40.00	79.00	80.00	69.00	90.00	108.77	10.43
	NN Predicted Age	62.08	52.80	97.91	74.78	66.31	98.95		
	Error	-4.08	-12.80	-18.91	5.22	2.69	-8.95		
Fold 9	Actual Age	59.00	68.00	58.00	80.00	59.00	49.00	134.94	11.62
	NN Predicted Age	68.55	45.47	59.96	85.22	57.47	35.68		
	Error	-9.55	22.53	-1.96	-5.22	1.53	13.32		
Fold 10	Actual Age	93.00	8.00	53.00	91.00	40.00	90.00	139.30	11.80
	NN Predicted Age	69.28	12.22	61.34	96.63	28.72	95.17		
	Error	23.72	-4.22	-8.34	-5.63	11.28	-5.17		

Error is measured in terms of the root mean square error (RMSE) on the test dataset. The RMSE is calculated in each fold and final RMSE of the single iteration was obtained by averaging the RMSE of all 10 folds. The averaged RMSE obtained in first iteration was 11.802 as shown in figure 5.19.

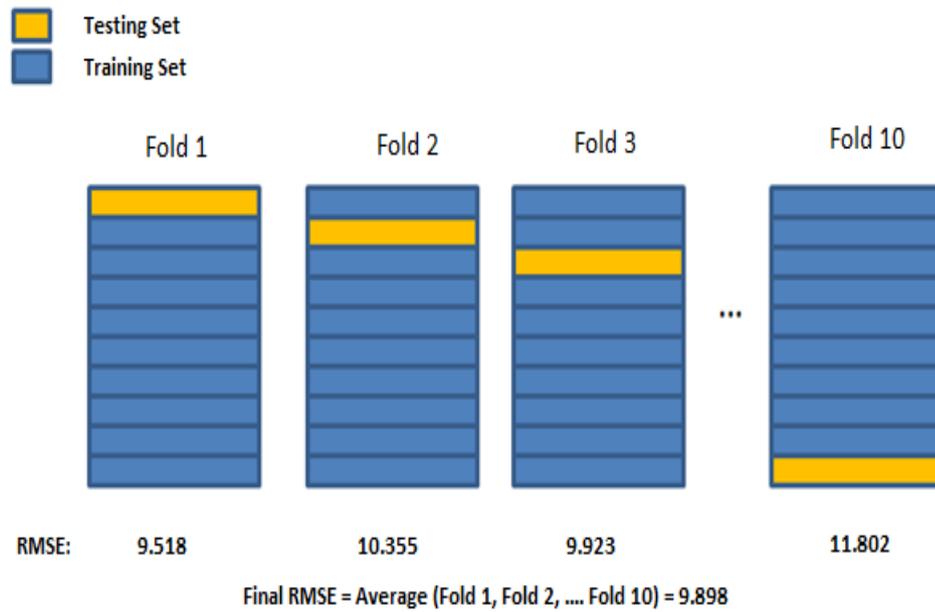


Figure 5.19: 1st iteration of 10-fold cross validation

Again, the data was randomly divided into 10 subgroups and 10-fold cross validation was performed. The 10-fold cross validation was repeated 5 times. The results from second iteration of 10 fold cross validation are shown in table 5.2. The average RMSE on the test dataset obtained in the second iteration was 11.50.

Table 5.2: Results from second iterations of 10-fold cross validation

Folds	Age	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6	MSE	RMSE
Fold 1	Actual Age	58.00	17.00	80.00	8.00	15.00	49.00	99.95	10.00
	NN Predicted Age	40.61	25.40	77.87	6.28	13.83	63.77		
	Error	17.39	-8.40	2.13	1.72	1.17	-14.77		
Fold 2	Actual Age	39.00	90.00	30.00	80.00	58.00	8.00	144.53	12.02
	NN Predicted Age	35.30	67.88	33.42	81.80	73.11	-2.99		
	Error	3.70	22.12	-3.42	-1.80	-15.11	10.99		
Fold 3	Actual Age	29.00	30.00	88.00	40.00	69.00	53.00	122.77	11.08
	NN Predicted Age	44.92	36.17	84.08	25.29	60.45	41.15		
	Error	-15.92	-6.17	3.92	14.71	8.55	11.85		
Fold 4	Actual Age	2.00	24.00	16.00	92.00	50.00	69.00	192.12	13.86
	NN Predicted Age	16.99	44.94	19.84	71.03	55.59	67.05		
	Error	-14.99	-20.94	-3.84	20.97	-5.59	1.95		
Fold 5	Actual Age	69.00	16.00	23.00	68.00	39.00	90.00	173.37	13.17
	NN Predicted Age	72.46	25.70	39.47	65.55	38.63	64.37		
	Error	-3.46	-9.70	-16.47	2.45	0.37	25.63		

Fold 6	Actual Age	39.00	91.00	68.00	79.00	59.00	93.00	217.14	14.74
	NN Predicted Age	33.01	74.99	86.33	66.14	74.60	76.70		
	Error	5.99	16.01	-18.33	12.86	-15.60	16.30		
Fold 7	Actual Age	80.00	16.00	89.00	49.00	58.00	68.00	88.43	9.40
	NN Predicted Age	66.26	17.74	89.81	66.66	60.22	63.40		
	Error	13.74	-1.74	-0.81	-17.66	-2.22	4.60		
Fold 8	Actual Age	8.00	49.00	90.00	59.00	91.00	29.00	126.72	11.26
	NN Predicted Age	7.45	39.19	113.44	67.41	96.42	25.22		
	Error	0.55	9.81	-23.44	-8.41	-5.42	3.78		
Fold 9	Actual Age	40.00	80.00	88.00	69.00	79.00	68.00	88.96	9.43
	NN Predicted Age	51.54	79.54	72.65	58.35	71.83	68.01		
	Error	-11.54	0.46	15.35	10.65	7.17	-0.01		
Fold 10	Actual Age	80.00	49.00	78.00	40.00	91.00	90.00	100.80	10.04
	NN Predicted Age	72.01	52.72	88.51	43.45	93.02	110.02		
	Error	7.99	-3.72	-10.51	-3.45	-2.02	-20.02		

The results of the third iteration of 10 fold cross validation are presented in table

5.3. Here the average RMSE was 11.247.

Table 5.3: Results from third iterations of 10-fold cross validation

Folds	Age	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6	MSE	RMSE
Fold 1	Actual Age	23.00	39.00	79.00	58.00	91.00	16.00	108.06	10.40
	NN Predicted Age	34.69	26.68	65.89	55.66	77.49	16.32		
	Error	-11.69	12.32	13.11	2.34	13.51	-0.32		
Fold 2	Actual Age	49.00	40.00	69.00	78.00	91.00	80.00	118.78	10.90
	NN Predicted Age	54.91	27.07	58.83	97.71	91.89	75.74		
	Error	-5.91	12.93	10.17	-19.71	-0.89	4.26		
Fold 3	Actual Age	39.00	80.00	88.00	53.00	16.00	69.00	99.29	9.96
	NN Predicted Age	33.00	67.18	81.02	58.23	6.70	53.74		
	Error	6.00	12.82	6.98	-5.23	9.30	15.26		
Fold 4	Actual Age	16.00	8.00	49.00	30.00	59.00	93.00	171.06	13.08
	NN Predicted Age	21.16	0.61	61.95	38.49	63.35	66.80		
	Error	-5.16	7.39	-12.95	-8.49	-4.35	26.20		
Fold 5	Actual Age	8.00	79.00	80.00	17.00	90.00	24.00	248.08	15.75
	NN Predicted Age	-15.18	70.47	66.49	26.41	72.42	41.27		
	Error	23.18	8.53	13.51	-9.41	17.58	-17.27		
Fold 6	Actual Age	69.00	91.00	29.00	80.00	58.00	15.00	117.79	10.85
	NN Predicted Age	84.47	103.64	14.71	75.44	50.92	9.31		
	Error	-15.47	-12.64	14.29	4.56	7.08	5.69		
Fold 7	Actual Age	40.00	2.00	50.00	90.00	30.00	8.00	102.49	10.12
	NN Predicted Age	57.77	-0.61	55.50	100.75	18.96	12.96		

	Error	-17.77	2.61	-5.50	-10.75	11.04	-4.96		
Fold 8	Actual Age	29.00	92.00	68.00	90.00	80.00	89.00	167.10	12.93
	NN Predicted Age	30.78	74.48	75.49	70.24	66.00	96.08		
	Error	-1.78	17.52	-7.49	19.76	14.00	-7.08		
Fold 9	Actual Age	68.00	58.00	39.00	49.00	88.00	68.00	80.46	8.97
	NN Predicted Age	59.27	57.39	30.79	51.03	74.33	55.84		
	Error	8.73	0.61	8.21	-2.03	13.67	12.16		
Fold 10	Actual Age	90.00	69.00	68.00	59.00	49.00	40.00	90.34	9.50
	NN Predicted Age	100.94	74.87	86.63	54.95	53.94	40.27		
	Error	-10.94	-5.87	-18.63	4.05	-4.94	-0.27		

Table 5.4 shows the results of the fourth iteration of the 10 fold cross validation.

Here the average RMSE was 10.907. It was observed that samples with very low age (less than 10) or very high age (greater than 80) account for most of the error in the neural network prediction.

Table 5.4: Results from third iterations of 10-fold cross validation

Folds	Age	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6	MSE	RMSE
Fold 1	Actual Age	16.00	80.00	40.00	69.00	90.00	8.00	84.20	9.18
	NN Predicted Age	8.93	72.18	55.12	77.77	81.49	12.02		
	Error	7.07	7.82	-15.12	-8.77	8.51	-4.02		
Fold 2	Actual Age	29.00	17.00	58.00	16.00	39.00	80.00	150.36	12.26
	NN Predicted Age	36.86	33.76	45.77	30.29	52.07	74.10		
	Error	-7.86	-16.76	12.23	-14.29	-13.07	5.90		
Fold 3	Actual Age	89.00	49.00	90.00	15.00	92.00	8.00	167.46	12.94
	NN Predicted Age	105.34	65.44	78.19	29.72	81.78	5.37		
	Error	-16.34	-16.44	11.81	-14.72	10.22	2.63		
Fold 4	Actual Age	49.00	40.00	80.00	91.00	58.00	90.00	94.65	9.73
	NN Predicted Age	43.78	35.40	84.00	70.22	51.07	85.11		
	Error	5.22	4.60	-4.00	20.78	6.93	4.89		
Fold 5	Actual Age	53.00	39.00	88.00	49.00	68.00	8.00	128.47	11.33
	NN Predicted Age	46.28	27.40	75.04	59.83	74.19	-8.35		
	Error	6.72	11.60	12.96	-10.83	-6.19	16.35		
Fold 6	Actual Age	24.00	40.00	30.00	59.00	68.00	29.00	126.66	11.25
	NN Predicted Age	11.51	31.72	26.98	73.42	82.81	38.95		
	Error	12.49	8.28	3.02	-14.42	-14.81	-9.95		
Fold 7	Actual Age	88.00	30.00	80.00	79.00	79.00	91.00	96.60	9.83
	NN Predicted Age	91.98	38.60	72.36	79.48	65.86	74.92		
	Error	-3.98	-8.60	7.64	-0.48	13.14	16.08		

Fold 8	Actual Age	2.00	91.00	59.00	23.00	58.00	93.00	111.29	10.55
	NN Predicted Age	13.70	102.12	67.09	28.12	72.11	82.21		
	Error	-11.70	-11.12	-8.09	-5.12	-14.11	10.79		
Fold 9	Actual Age	16.00	69.00	80.00	78.00	69.00	68.00	146.07	12.09
	NN Predicted Age	9.21	55.14	59.52	86.03	81.40	67.09		
	Error	6.79	13.86	20.48	-8.03	-12.40	0.91		
Fold 10	Actual Age	50.00	49.00	68.00	39.00	90.00	69.00	98.12	9.91
	NN Predicted Age	45.14	44.38	82.72	24.73	98.02	76.69		
	Error	4.86	4.62	-14.72	14.27	-8.02	-7.69		

The results of the final iteration of 10 fold cross validation are presented in table

5.5. In this iteration, the average RMSE was 11.219.

Table 5.5: Results from fifth iterations of 10-fold cross validation

Folds	Age	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6	MSE	RMSE
Fold 1	Actual Age	88.00	89.00	8.00	90.00	69.00	30.00	170.97	13.08
	NN Predicted Age	73.45	89.93	16.67	65.25	80.20	29.94		
	Error	14.55	-0.93	-8.67	24.75	-11.20	0.06		
Fold 2	Actual Age	39.00	79.00	68.00	16.00	69.00	68.00	85.08	9.22
	NN Predicted Age	28.59	95.72	77.66	21.24	68.79	66.71		
	Error	10.41	-16.72	-9.66	-5.24	0.21	1.29		
Fold 3	Actual Age	16.00	68.00	90.00	49.00	58.00	88.00	132.98	11.53
	NN Predicted Age	21.72	50.67	69.75	51.68	58.10	94.89		
	Error	-5.72	17.33	20.25	-2.68	-0.10	-6.89		
Fold 4	Actual Age	2.00	80.00	93.00	92.00	49.00	68.00	180.10	13.42
	NN Predicted Age	17.69	82.10	68.37	77.98	44.67	65.14		
	Error	-15.69	-2.10	24.63	14.02	4.33	2.86		
Fold 5	Actual Age	40.00	91.00	90.00	59.00	59.00	53.00	80.52	8.97
	NN Predicted Age	48.47	97.73	88.88	72.77	69.43	61.15		
	Error	-8.47	-6.73	1.12	-13.77	-10.43	-8.15		
Fold 6	Actual Age	80.00	39.00	8.00	29.00	24.00	69.00	100.59	10.03
	NN Predicted Age	81.43	37.19	-4.47	47.70	32.93	72.65		
	Error	-1.43	1.81	12.47	-18.70	-8.93	-3.65		
Fold 7	Actual Age	69.00	39.00	78.00	40.00	49.00	91.00	174.18	13.20
	NN Predicted Age	61.28	43.74	75.87	45.84	40.93	61.69		
	Error	7.72	-4.74	2.13	-5.84	8.07	29.31		
Fold 8	Actual Age	58.00	90.00	80.00	16.00	30.00	17.00	111.65	10.57
	NN Predicted Age	50.64	86.95	56.87	17.28	34.54	24.00		
	Error	7.36	3.05	23.13	-1.28	-4.54	-7.00		
Fold 9	Actual Age	15.00	91.00	80.00	8.00	50.00	79.00	125.60	11.21
	NN Predicted Age	19.63	68.22	93.63	3.20	50.49	81.02		

	Error	-4.63	22.78	-13.63	4.80	-0.49	-2.02		
Fold 10	Actual Age	80.00	23.00	49.00	58.00	29.00	40.00	120.14	10.96
	NN Predicted Age	59.32	30.81	53.45	56.43	17.85	49.24		
	Error	20.68	-7.81	-4.45	1.57	11.15	-9.24		

Table 5.6 shows the MSE and RMSE of fold 1-10 for all 5 iterations. The average MSE of 5 iterations of 10 fold cross validation was 126.33 with a standard deviation of 7.43. The average RMSE of 5 iterations of 10 fold cross validation was 11.12 with a standard deviation of 1.63. By comparing the root mean square errors produced in each fold, it can be seen that the neural network model presented in this thesis is consistent in terms of performance.

Table 5.6: Results from all 5 iterations of 10-fold cross validation

	Iteration 1		Iteration 2		Iteration 3		Iteration 4		Iteration 5	
	MSE	RMSE								
Fold 1	90.59	9.52	99.95	10.00	108.06	10.40	84.20	9.18	170.97	13.08
Fold 2	107.22	10.36	144.53	12.02	118.79	10.90	150.36	12.26	85.08	9.22
Fold 3	98.47	9.92	122.77	11.08	99.29	9.96	167.46	12.94	132.99	11.53
Fold 4	95.60	9.78	192.12	13.86	171.06	13.08	94.65	9.73	180.10	13.42
Fold 5	110.89	10.53	173.37	13.17	248.08	15.75	128.47	11.34	80.52	8.97
Fold 6	89.94	9.48	217.14	14.74	117.79	10.85	126.66	11.25	100.59	10.03
Fold 7	196.99	14.04	88.43	9.40	102.49	10.12	96.60	9.83	174.18	13.20
Fold 8	108.77	10.43	126.72	11.26	167.10	12.93	111.29	10.55	111.65	10.57
Fold 9	134.94	11.62	88.96	9.43	80.46	8.97	146.07	12.09	125.60	11.21
Fold 10	139.30	11.80	100.80	10.04	90.34	9.51	98.12	9.91	120.14	10.96
Avg.	117.27	10.75	135.48	11.50	130.35	11.25	120.39	10.91	128.18	11.22
Std.	32.74	1.40	45.31	1.90	51.09	2.07	27.88	1.26	36.36	1.61

5.5 PRINCIPAL COMPONENT NEURAL NETWORK MODEL

Here, the data is first standardized using the standard deviation method. Then the principal components were computed [19]. From the previous PCA analysis, it can be seen that much of the variance (about 87%) is explained by first three components. The first component accounts for 36% of the variance, the second component accounts for 30% and the third component accounts for about 21%. The 24 input features are now replaced by 3 principal components.

By using the principal components, a new neural network model is built. Unlike the previous neural network model, this new principal component neural network uses only three input features (the first three principal components). There are three inputs, one output layer and one hidden layer. Using the conventional rule discussed in previous section for selecting the number of hidden neurons (hidden neurons= mean of input and output), the number of hidden neurons was set to 2. However, the network failed to provide acceptable performance. Gradually the number of neurons was increased and it was found that good performance was obtained by setting the number of neurons to 10.

61 samples were fed to the neural network, out of which 75% of were used for building (training) the network, 15% for validating the network and the remaining 15% for testing the network. Figure 5.20 shows the training state and mean square error for the network.

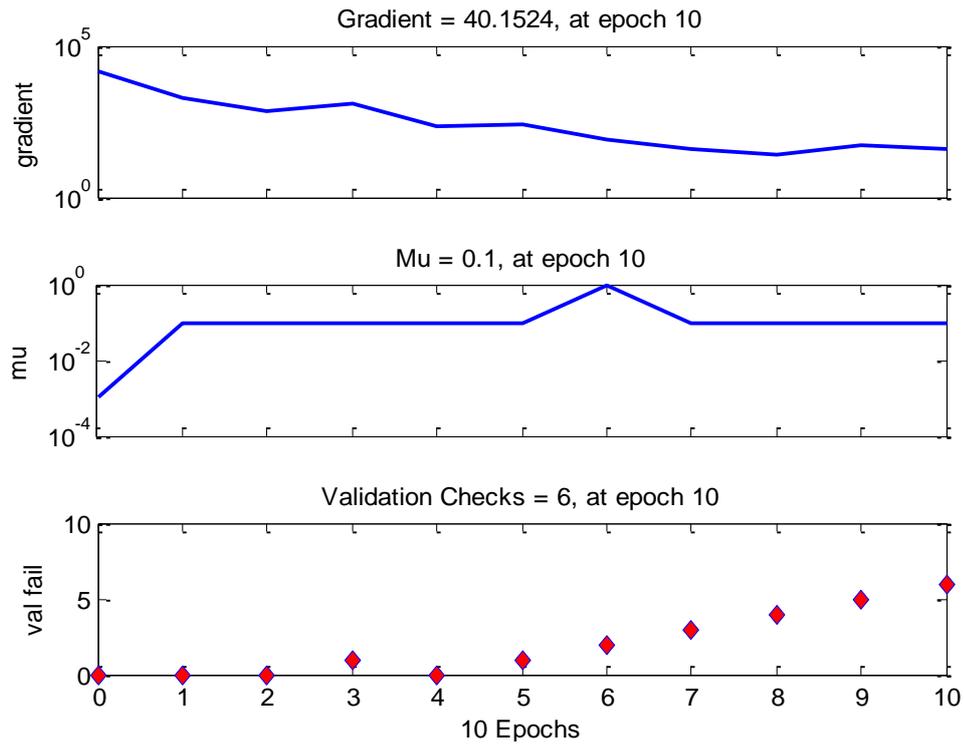


Figure 5.20: (a) Training state of PCNN

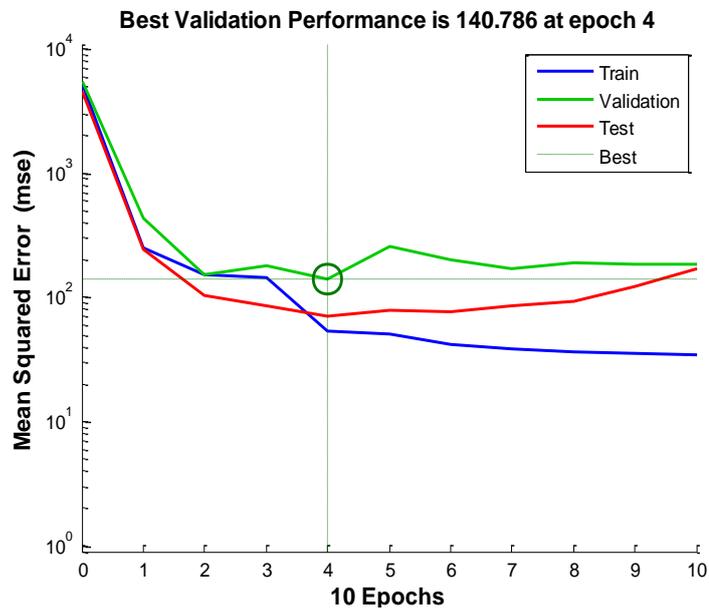


Figure 5.20: (b) Performance of PCNN

Figure 5.21, shows the regression of the principal component neural network. The overall regression of the principal component neural network is $R=0.95356$. The regression for the training, validation and testing is $R=0.96586$, $R=0.93115$ and

R=0.92412, respectively. It can be seen that both the model provides almost identical results. PCNN (with overall regression=0.95356) performs as well as the NN model (regression=0.97316). Although PCNN has a slightly lower accuracy (almost 2%) it is worth noting that the PCNN has 8 times fewer inputs. Although this is a pilot study with a very small sample size, for a sample of 10,000 MRI scans this reduction in dimensionality would be highly significant. In this case, the sample size for ordinary neural network will be $(24 \times 10,000 = 240,000)$, while for PCNN this size will be $(3 \times 10,000 = 30,000)$. This PCNN is computationally efficient and will be very helpful when using cascaded neural networks. The error histogram of this model is shown in figure 5.22 below which give the difference between actual and neural network estimated values.

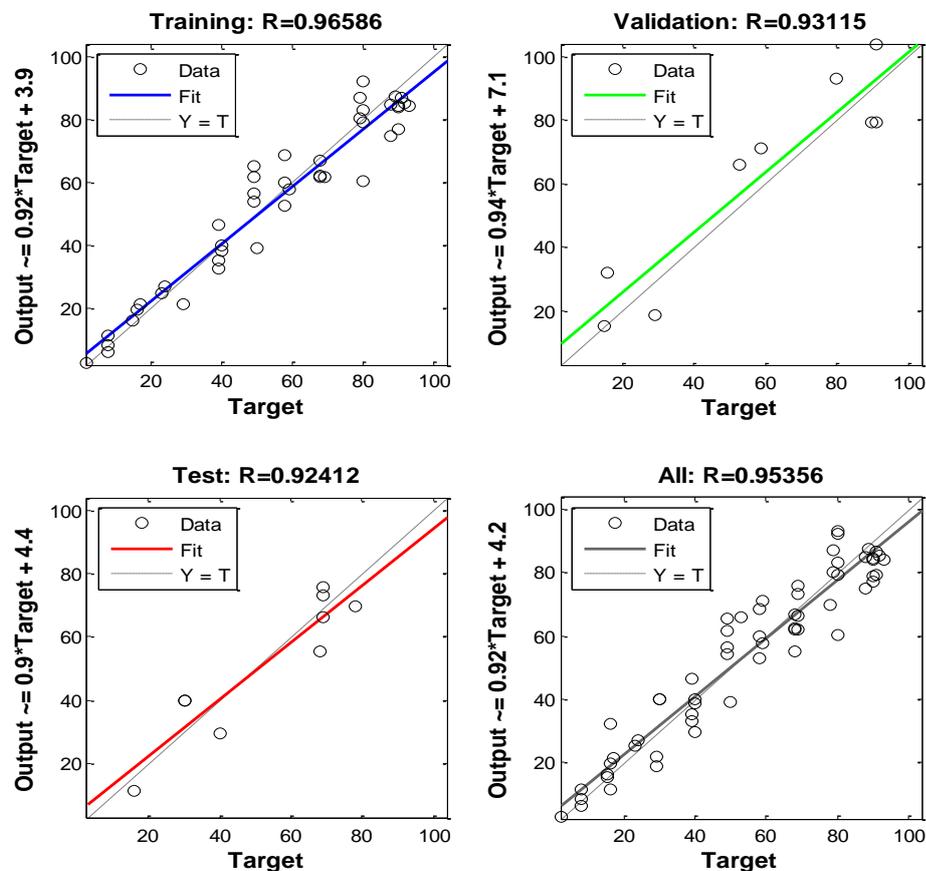


Figure 5.21: Regression of principal component neural network

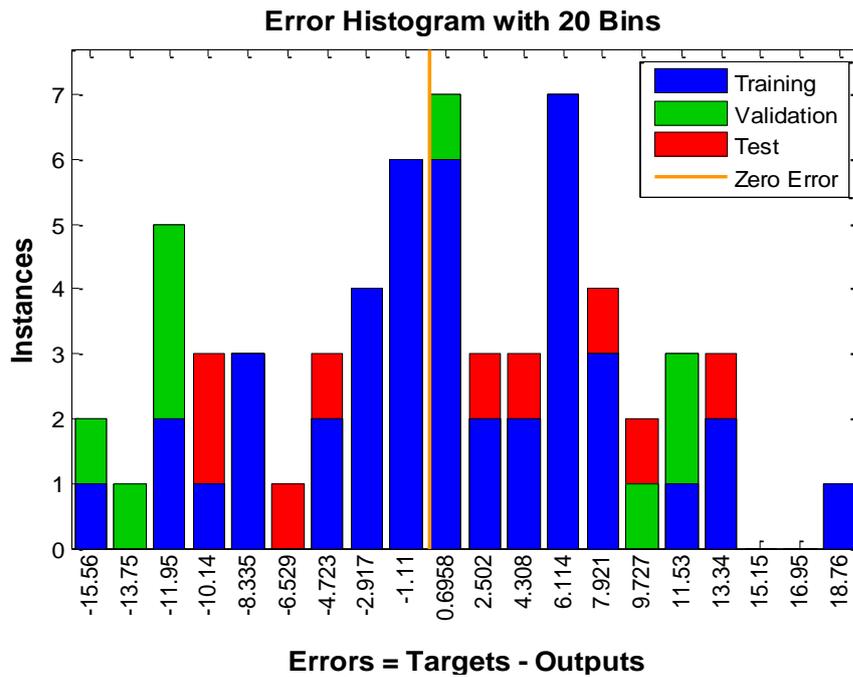


Figure 5.22: Error histogram for principal component neural network

5.6 FUZZY INFERENCE SYSTEM

A process for formulating the mapping between an input space and an output using fuzzy logic is known as a fuzzy inference system. The mapping provides a basis for decision making and pattern recognition. Fuzzy inference processes involve a collection of membership functions, logical operations, and linguistic (if-then) rules. A fuzzy inference system comprises four parts: fuzzification, implication, aggregation and defuzzification. Details of these procedures are explained in chapter 2. Neural network model described in previous section worked very well but due to its black box nature, it failed to provide any useful information about the correlation between inputs and output. In this section fuzzy inference system is used to understand the relationship between inputs and output by extracting linguistic rules. There are two main types of fuzzy inference systems: a Mamdani type fuzzy inference and Sugeno type fuzzy inference system.

5.6.1 MAMDANI FUZZY INFERENCE SYSTEM

Mamdani type fuzzy inference is the most commonly used fuzzy method. This method was proposed by Ebrahim Mamdani [20] in 1975, in order to control a steam engine and boiler combination. He achieved this by creating a set of linguistic control rules obtained from experienced human operators. Fuzzy logic was introduced by Lotfi Zadeh in 1973 through a paper on fuzzy algorithms for complex systems and decision processes [21]. Mamdani's fuzzy inference was built on the fuzzy logic presented by Lotfi Zadeh. The difference between Mamdani and Sugeno type fuzzy inference is the shape of the output. In Mamdani-type inference, the output membership functions are fuzzy sets that need defuzzification to produce a final output whereas in Sugeno type fuzzy inference, output membership functions are either linear or constant. A Mamdani type fuzzy inference system was built in Matlab with 24 inputs (spinal features) and 1 output (spinal age). Each input has three membership functions defined as low, medium and high. Triangular membership functions are used here. The range of each input variable is determined from data set and defined in the Mamdani FIS. The range of inputs 1-5 (vertebral heights) is set 10-30, input 6-11 (disc heights) as 1-17, inputs 12-17 (disc signals) as 20-500, inputs 18-19 (PSM) as 10-425, inputs 20-21 (Psoas) as 20-180, inputs 22-23 (Subcutaneous Fat Signal) as 150-1000, and input 24 (CSF) have range 150-1500. 21 linguistic if-then rules are used in this FIS. These generic rules are designed on the basis of information obtained from PCA, factor analysis, and SOM clustering used in the next chapter. Figure 5.23 shows a general representation of the Mamdani type fuzzy inference system.

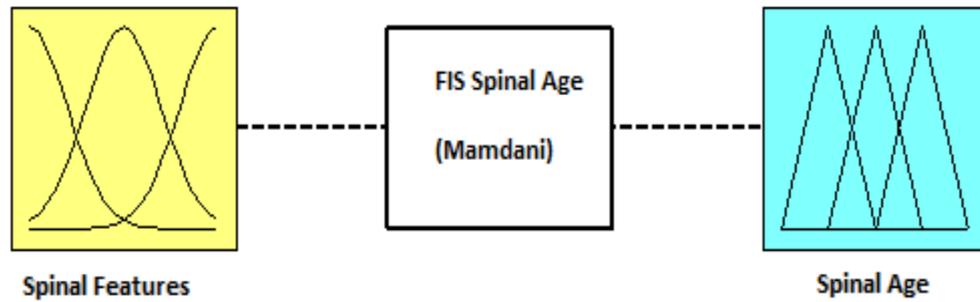


Figure 5.23: Mamdani type fuzzy inference system for spinal age estimation

Since FIS models lack the ability to learn from data, a neural network is used to make it adaptive by tuning the membership function to map the input to outputs correctly. Since Sugeno type fuzzy inference systems have linear or constant output membership functions, they can be used easily with neural networks for adaptation. So in this chapter, a Sugeno type fuzzy inference is preferred. However, Mamdani-type FIS can be easily transformed to Sugeno FIS by determining the centroid of the output membership functions of Mamdani FIS [22].

5.6.2 *SUGENO FUZZY INFERENCE SYSTEM*

Sugeno fuzzy inference systems were introduced in 1985 [23]. The only difference between Mamdani and Sugeno methods are in output membership functions. In Sugeno FIS, outputs are either linear or constant. Sugeno fuzzy inference systems are computationally efficient compared to Mamdani fuzzy inference systems. The same 24 inputs and single output are used as in the Mamdani FIS. The Sugeno type fuzzy inference system is illustrated in figure 5.24.

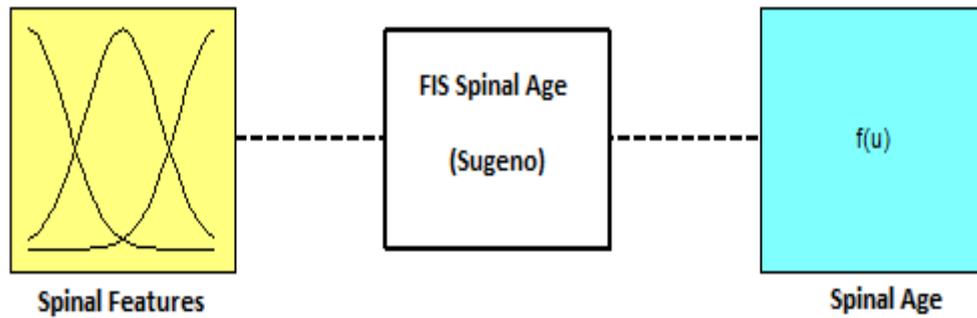


Figure 5.24: Sugeno type fuzzy inference system

For all inputs, three membership functions were used: low, medium, and high. The shapes of the input membership functions were triangular. Each input has its specific range as defined in case of Mamdani FIS. For example, input membership functions for vertebral height L1 are shown in figure 5.25 below.

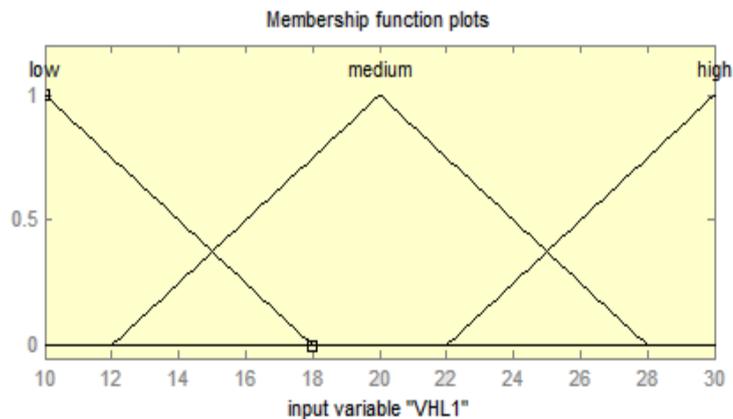


Figure 5.25: Input membership functions for Sugeno type FIS

21 rules were used to build the Sugeno FIS. These rules were made on the basis of knowledge extracted from PCA, FA and SOM analysis. Each rule produces specific results, so there were 21 output membership functions. These output membership functions were constant values in term of age.

On the basis of the fuzzy rules, the relationship of a single input can be visualized against any other input. For example, a surface view of vertebral height L1 vs. L2 is shown in figure 5.26 (a), and surface view of vertebral height L1 vs. CSF is shown in figure 5.26 (b) below. It can be seen from figure 5.26 (a) that both vertebral heights L1 and L2 increase and decrease together. Similarly, the relationships between all other inputs can be visualized. If required, rules can be amended to represent a meaningful relationship between input features.

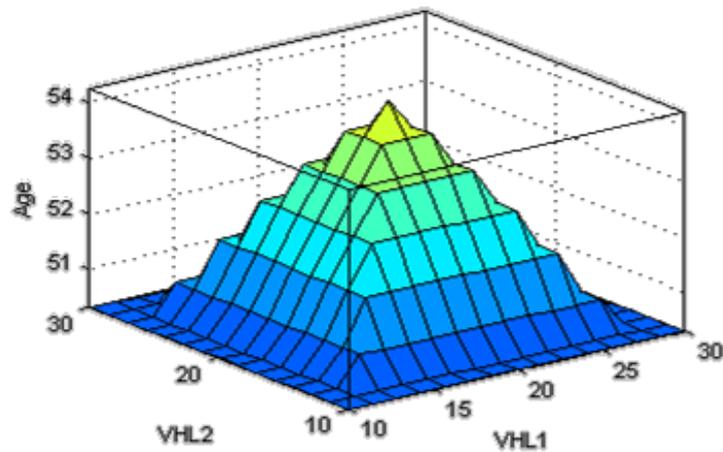


Figure 5.26: (a) Surface view of Sugeno type FIS inputs (VHL1 and VHL2, vs. age)

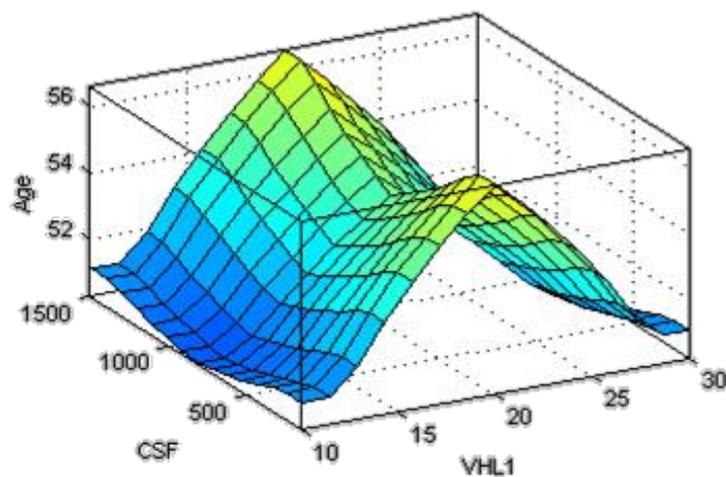


Figure 5.26: (b) Surface view of Sugeno type FIS inputs (VHL1 and CSF, vs. age)

5.7 HYBRID MODEL

Fuzzy inference systems provide a linguistic and easy to understand model but they lack learning ability. Also, the tuning of fuzzy inference system is mainly based on a trial and error method, which is time consuming. Other intelligent system techniques can be used to overcome the tuning problem of fuzzy inference systems and their lack of ability to learn. Fusion of two or more intelligent techniques produces a hybrid intelligent model. Hybrid intelligent systems are often more efficient and accurate compared to the stand-alone systems. Fuzzy inference systems have been used with neural networks, genetic algorithms, and many other techniques. The fusion of a neural network with fuzzy inference makes system adaptive [24]. Such systems are called adaptive neuro-fuzzy inference systems (ANFIS).

5.7.1 ADAPTIVE NEURO-FUZZY INFERENCE SYSTEM (ANFIS)

A fuzzy inference system (FIS) can be constructed and tuned by using a given input-output data set. In such case membership function parameters are tuned by using either a backpropagation algorithm alone or in combination with a least squares type of method. This procedure allows fuzzy systems to learn from the data.

In this chapter the Sugeno type fuzzy inference system is trained with a neural network. Inputs are fed to the Sugeno FIS and the outputs are compared to the targets. The error (difference between target and output values) is calculated and FIS parameters are readjusted in an attempt to minimize the error (usually the square error). The model structure for FIS has 24 inputs, 72 input membership

functions, 21 rules, and 21 output membership functions as shown in figure 5.27 below. In the ANFIS model, 41 samples were used for training, 10 for testing and 10 for validation. These samples were trained by using a hybrid optimization method which is a combination of least-squares and backpropagation gradient descent methods. Here, backpropagation is used for the parameters linked to the input membership functions, and the least squares estimation is used for the parameters linked to the output membership functions.

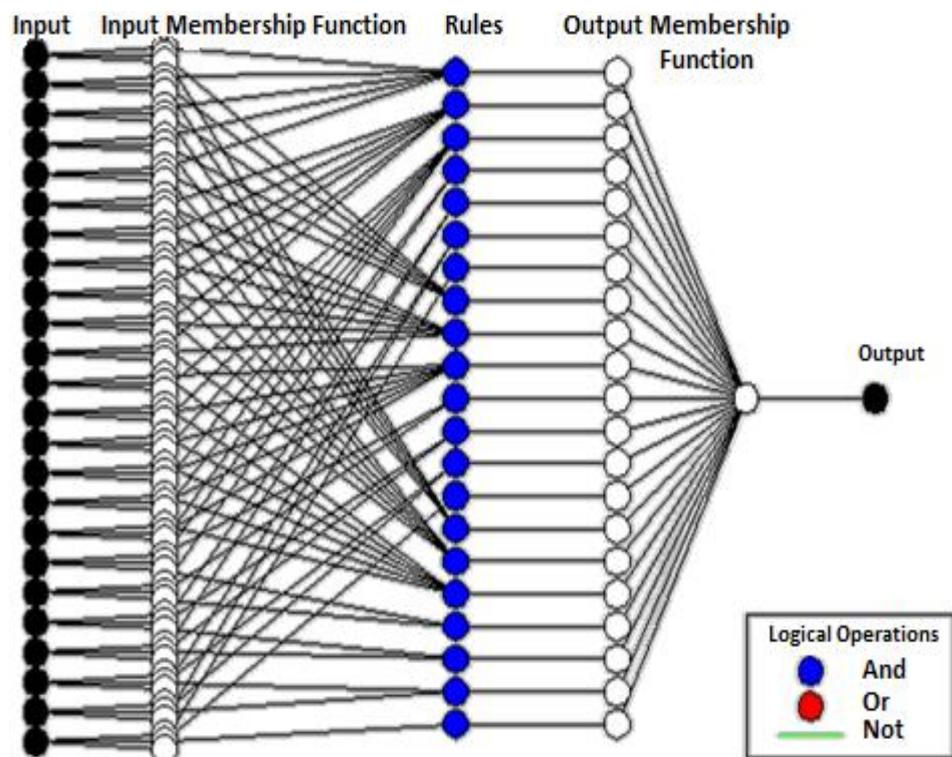


Figure 5.27: ANFIS model structure

Figure 5.28 illustrates the training of the Sugeno FIS with 41 samples. In this figure, blue circles shows actual outputs and the red "+" signs show the Sugeno FIS outputs. The horizontal axis shows the number of samples and on vertical axis is the age in years.

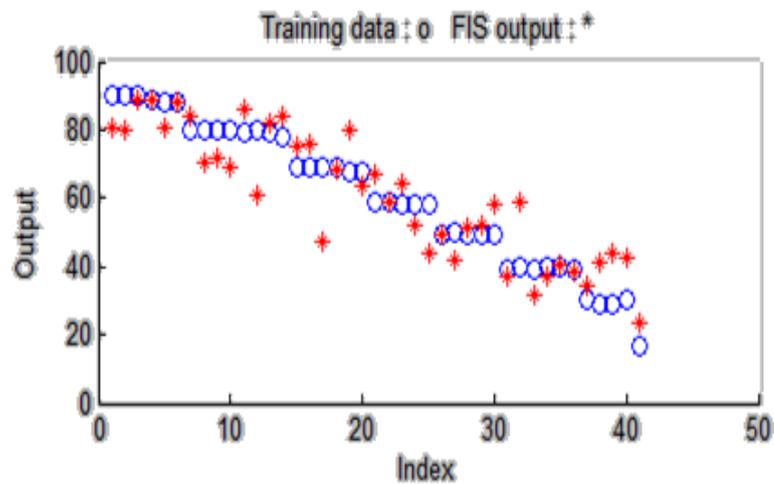


Figure 5.28: ANFIS training with 41 samples

The ANFIS model was validated with 10 samples and tested with another 10 samples. The validation of the FIS is illustrated in figure 5.29 (a) where blue dots show the original outputs and red “*” shows output of the Sugeno FIS. The average validation error is 27.2066, which is high. It can be seen from the figures that almost all of the samples are classified with a higher than actual age. It is also worth noting that 9/10 of the validation samples have age less than 20, i.e., almost all the samples in the validation set belong to the same age group. It is very difficult to estimate the age from spinal features, especially in the age group of less than 10 years. Because spinal features develop rapidly in this age group, assessments are not necessarily reliable. Moreover, this research is mainly focused on back pain and spinal diseases that usually affect middle or late age patients, so the results and spinal variations in mature and elderly samples are specially taken into account.

The performance of the Sugeno FIS on the test set is shown in figure 5.29 (b), in which blue “+” indicate the test samples and the red “*” indicate the FIS output.

The average test error is 19.2319. It can be seen that 7 out of 10 samples are estimated with close to the correct age.

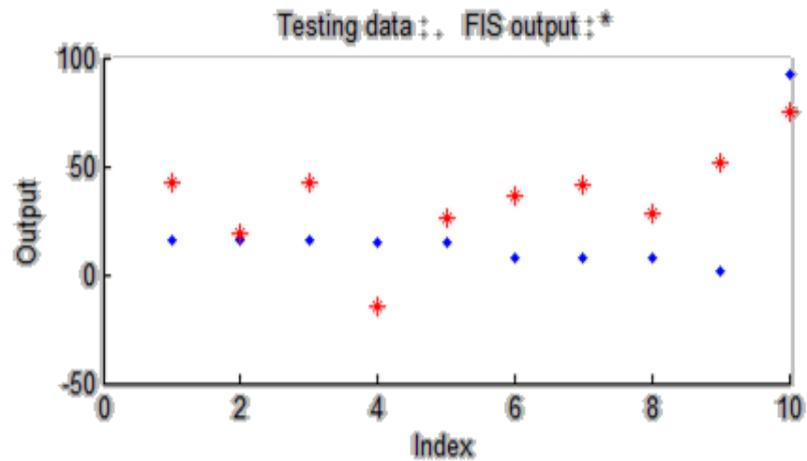


Figure 5.29: (a) Validation of ANFIS

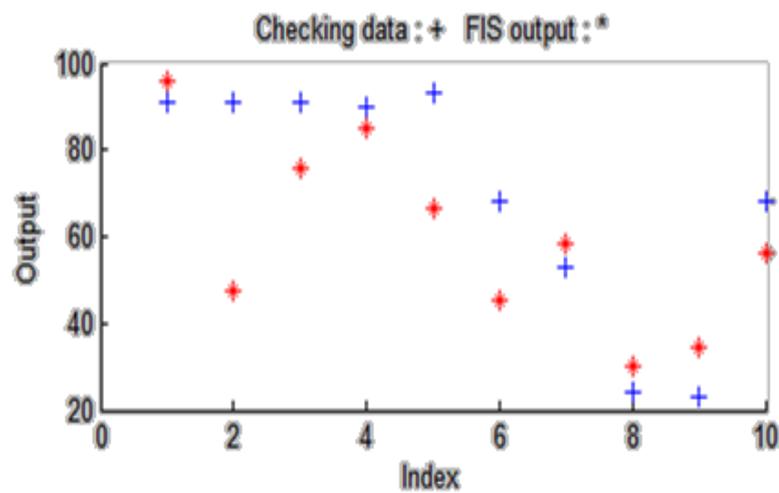


Figure 5.29: (b) Testing of ANFIS

Another model was considered by training the Sugeno FIS with back propagation method for all parameters (a steepest descent method). This method produced very similar results to the previously used hybrid method consisting of backpropagation for the parameters associated with the input membership functions, and a least squares estimation for the parameters associated with the output membership functions.

5.7.2 ANFIS WITH SUBTRACTIVE CLUSTERING

It is difficult to choose the shape and parameters of membership functions by simply inspecting the dataset. Furthermore, it is difficult to produce rules that interpret all the characteristics of the variables in the model. The ANFIS model discussed in the previous section is based on a Sugeno FIS, modelled by human interpretation and knowledge extracted from the results of the previous data analysis. However, the resulting membership functions, parameter ranges, and fuzzy rules might not cover all of the characteristics of the variables in the data.

In this section, another model is presented in which subtractive clustering [25] is applied on the dataset to generate an initial FIS. The purpose of clustering is to find natural groups in the dataset and understand their behaviour. The subtractive clustering method estimates the number of clusters and defines cluster centres in the dataset. Using the cluster information from subtractive clustering, a Sugeno FIS was generated having a minimum number of rules to present the fuzzy qualities associated with each of the clusters. The FIS generated by subtractive clustering has 24 inputs, 41 rules, and one output.

The new Sugeno FIS was trained using a hybrid method (combination of least-squares and backpropagation gradient descent method). 41 samples were used for training, 10 for testing, and 10 for checking the new FIS. Training of FIS is illustrated in figure 5.30 (a) below, with an average error of 1.2135×10^{-6} . Testing of this new FIS yielded an average error of 0.4593 as shown in figure 5.30 (b). Checking of new FIS produced an average error of 0.8267 as shown in figure 5.30 (c).

Training of this new FIS alters the shape of the membership functions and their parameters. After training, the updated relationship between the inputs and inputs vs. output are understood with the help of a surface view. Figure 5.31 (a) below gives surface view of first two inputs vs. output.

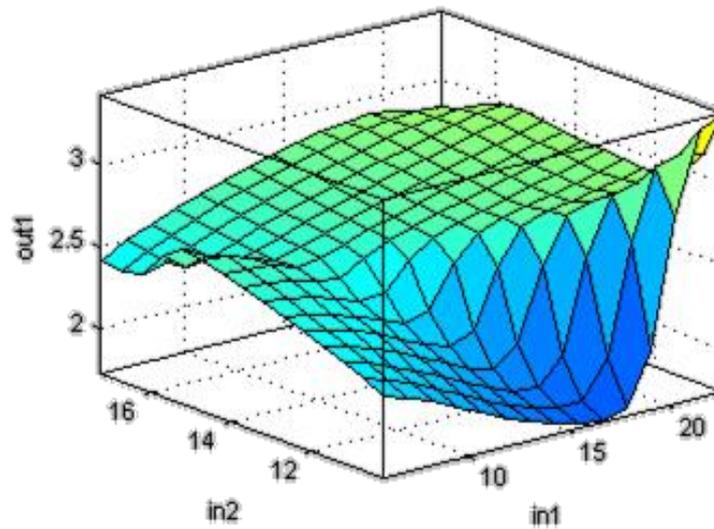


Figure 5.31: (a) Surface view of new FIS (in1, in2 vs. out1)

The surface view of input 1 (vertebral height L) and input 3 (vertebral height L3), vs. output is shown in figure 5.31 (b). Similarly, the surface view of input 18 (PSM-Left) and input 24 (CSF), vs. output is shown in figure 5.31 (c) below.

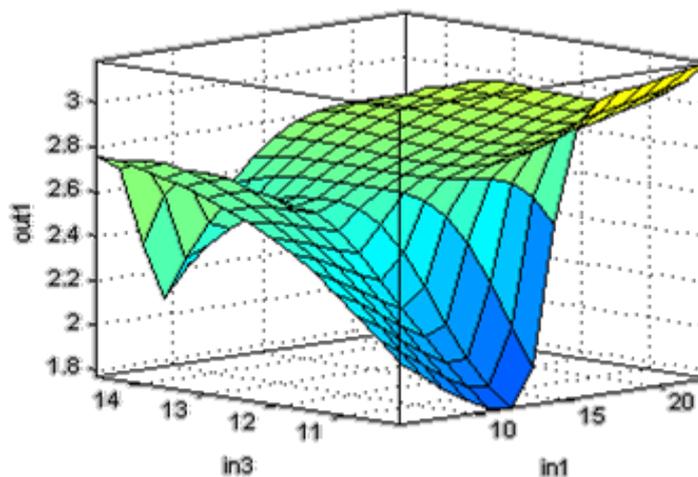


Figure 5.31: (b) Surface view of new FIS (in1, in3 vs. out1)

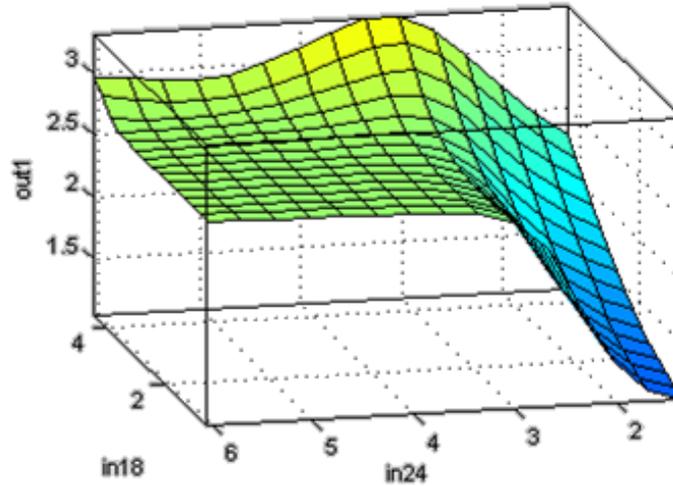


Figure 5.31 (c) Surface view of new FIS (in18 and in24, vs. out1)

The Sugeno fuzzy inference system generated through subtractive clustering showed better performance than the Sugeno fuzzy inference system designed previously. The relationships among the data were observed by exploring fuzzy if-then rules. These linguistic rules describe most of the patterns seen in the data.

5.8 SUMMARY

In this chapter, a feed forward neural network model was presented that was capable of predicting the spinal age from the extracted features of the lumbar spine. This model was validated through repeated 10-fold cross validation. A new term “Spinal Age” was introduced as an identifier to explain the overall behaviour of the spine. In this chapter, a principal component neural network model was also presented along with the comparison of both neural network models. Furthermore, a fuzzy and a hybrid intelligent model (adaptive neuro-fuzzy inference system) were presented in this chapter to extract simple linguistic rules from the data. Finally an expert system was developed that was capable of predicting the spinal age from given spinal features.

REFERENCES

- [1] Cleophas, Ton J., and Aeilko H. Zwinderman. Machine Learning in Medicine. Springer, (2013).
- [2] Magoulas, George D., and Andriana Prentza. "Machine learning in medical applications." In Machine Learning and its applications, pp. 300-307. Springer Berlin Heidelberg, (2001).
- [3] Baxt, William G. "Application of artificial neural networks to clinical medicine." The lancet 346, no. 8983 (1995): 1135-1138.
- [4] Lavrač, Nada. Machine learning for data mining in medicine. Springer Berlin Heidelberg, (1999): 47-62,
- [5] Magoulas, George D., and Andriana Prentza. "Machine learning in medical applications." In Machine Learning and its applications, Springer Berlin Heidelberg, (2001): 300-307.
- [6] Amato, Filippo, Alberto López, Eladia María Peña-Méndez, Petr Vaňhara, Aleš Hampl, and Josef Havel. "Artificial neural networks in medical diagnosis." Journal of Applied Biomedicine 11, no. 2 (2013): 47-58.
- [7] Diamantaras, Konstantinos I., and Sun Y. Kung. Principal component neural networks. New York: Wiley, (1996).
- [8] Popa, C., and Cosmin Cernăzanu-Glăvan. "Pattern neural networks: A case study." In Proc. 2nd Workshop on Software Services: Cloud Computing and Applications based on Software Services, Timisoara. (2011).

- [9] Math Works, Matlab Neural Network Toolbox, tutorial 2011a.
- [10] Al-Shayea, Qeethara Kadhim. "Artificial Neural Networks in Medical Diagnosis." *International Journal of Computer Science Issues (IJCSI)* 8, no. 2 (2011): 150-154.
- [11] Lawrence, Steve, C. Lee Giles, and Ah Chung Tsoi. "What size neural network gives optimal generalization? Convergence properties of backpropagation." Technical Report UMIACS-TR-96-22 and CS-TR-3617, Institute for Advanced Computer Studies, University of Maryland, College Park, (1998).
- [12] Panchal, Gaurang, Amit Ganatra, Y. P. Kosta, and Devyani Panchal. "Behaviour analysis of multilayer perceptrons with multiple hidden neurons and hidden layers." *International Journal of Computer Theory and Engineering* 3, no. 2 (2011): 332-337.
- [13] Geman, Stuart, Elie Bienenstock, and René Doursat. "Neural networks and the bias/variance dilemma." *Neural computation* 4, no. 1 (1992): 1-58.
- [14] Hagan, M.T., H.B. Demuth, and M.H. Beale, "Neural Network Design," Boston, MA: PWS Publishing, (1996).
- [15] Murray, Alan F., ed. *Applications of neural networks*. Boston: Kluwer Academic Publishers, (1995).
- [16] Widrow, Bernard, David E. Rumelhart, and Michael A. Lehr. "Neural networks: Applications in industry, business and science." *Communications of the ACM* 37, no. 3 (1994): 93-105.

- [17] Marquardt, Donald W. "An algorithm for least-squares estimation of nonlinear parameters." *Journal of the Society for Industrial & Applied Mathematics* 11, no. 2 (1963): 431-441.
- [18] Moré, Jorge J. "The Levenberg-Marquardt algorithm: implementation and theory." In *Numerical analysis*, pp. 105-116. Springer Berlin Heidelberg, (1978).
- [19] Shlens, Jonathon. "A tutorial on principal component analysis." *Systems Neurobiology Laboratory, University of California at San Diego* 82 (2005).
- [20] Mamdani, Ebrahim H., and Sedrak Assilian. "An experiment in linguistic synthesis with a fuzzy logic controller." *International journal of man-machine studies* 7, no. 1 (1975): 1-13.
- [21] Zadeh, Lotfi A. "Outline of a new approach to the analysis of complex systems and decision processes." *Systems, Man and Cybernetics, IEEE Transactions on* 1 (1973): 28-44.
- [22] Jassbi, Javad, S. H. Alavi, Paulo JA Serra, and Rita Almeida Ribeiro. "Transformation of a Mamdani FIS to first order Sugeno FIS." In *Fuzzy Systems Conference, 2007. FUZZ-IEEE 2007. IEEE International, IEEE*, (2007): 1-6.
- [23] Sugeno, Michio. *Industrial applications of fuzzy control*. Elsevier Science Inc., (1985).

- [24] Jang, J-SR. "ANFIS: adaptive-network-based fuzzy inference system." Systems, Man and Cybernetics, IEEE Transactions on 23, no. 3 (1993): 665-685.
- [25] Chiu, Stephen L. "Fuzzy model identification based on cluster estimation." Journal of intelligent and Fuzzy systems 2, no. 3 (1994): 267-278.

6

UNSUPERVISED PATTERN RECOGNITION IN AGEING SPINE

6.1 UNSEPERVISED LEARNING

6.2 SELF ORGANIZING MAPS (SOM)

6.3 SOM MOELLING AND ANALYSIS

6.3.1 SOM MODELLING

6.3.2 VISUAL CLUSTER ANALYSIS

6.4 CLUSTERING ANALYSIS FOR AGEING PATTERN OF SPINE

6.4.1 WARD'S CLUSTERING

6.4.2 MODIFIED SOM-WARD CLUSTERING

6.4.3 CHARACTERISTICS OF THE CLUSTERS

6.5 SUMMARY

REFERENCES

6.1 UNSUPERVISED LEARNING

Unsupervised learning is a type of machine learning used to find hidden structures in data without labelled responses. Since the examples given to the learner are unlabelled, there is no error or reward signal to evaluate a potential solution [1]. One of the most common unsupervised learning methods is cluster analysis. In this chapter unsupervised learning is used for exploratory data analysis and to find hidden patterns and groups in the data. These groups (clusters) are modelled using a measure of similarity, which is defined using metrics (distance functions) such as the Euclidean distance. A comparison between the characteristics of these clusters was conducted to understand the spinal variations among different groups. This is helpful in setting standards for spinal features among different age groups. Some common clustering algorithms include:

- **Hierarchical clustering:** builds a multilevel hierarchy of clusters by creating a cluster tree.
- **K-Means clustering:** partitions data into k distinct clusters based on a distance to the centroid of a cluster.
- **Gaussian mixture models:** clusters as a mixture of multivariate normal density components.
- **Self-organizing maps:** uses neural networks that learn the topology and distribution of the data.
- **Adaptive resonance theory (ART):** allows the number of clusters to vary with problem size and lets the user control the degree of similarity between

members of the same clusters by means of a user-defined constant called the vigilance parameter.

Self-organizing map (SOM) and adaptive resonance theory (ART) are the two commonly used types of used unsupervised learning algorithms in artificial neural networks. The SOM is a topographic representation where nearby locations in the map represents inputs that have similar properties. The ART model is user dependent that allows the number of clusters to vary depending on the problem and lets the user control the degree of similarity between members of the same clusters through vigilance parameter [2]. Unsupervised learning methods are widely used in medical domains. They are used in bioinformatics (for sequence analysis), genetic clustering (in data mining for sequence and pattern), in medical imaging (for image segmentation), and in computer vision (for object recognition) [3], [4].

6.2 SELF-ORGANIZING MAPS (SOMs)

A self-organizing map (SOM) is a data visualization technique that reduces the dimensionality of data through the use of self-organizing neural networks. It transforms complex and nonlinear statistical relationships between high-dimensional data into simple geometric relationships on a low-dimensional (usually 2D) map. SOM compresses the information at the same time preserving the most important topological and metric relationships of the original data on the display [5]. SOM reduces the dimensionality by producing a map of usually two dimensions, which plot the similarities of the data by grouping similar data items together. SOM thus reduces the dimensionality and displays the similarities.

The proposed model in this chapter has a set of 24 inputs. SOM groups or ranks each sample (lumbar MRI) on the basis of similarities in their 24 features and assigns certain location to each sample in the map. Training and mapping are the two main processes in SOM. In the training, it constructs the map from input samples. After training, it automatically classifies a new input sample in the mapping process. The map consists of several neurons, which are associated with a weight vector that has the same number of coordinates as the input sample and a position in the map. The neurons are arranged originally in physical positions according to a topology function, such as a grid, hexagonal or other topology. The purpose of SOM is to detect regularities and correlations in the input and also to recognize groups of similar input vectors [6], [7]. SOMs have the ability to adapt to the future responses of the input, e.g., the neurons of competitive networks physically near each other respond to similar input vectors.

The self-organization process involves four major procedures:

Initialization: First of all the connection weights are initialized with random values.

Competition: For every input vector, the neurons compute their respective values of a discriminant function which gives the basis for competition. The value of the discriminant function decides the winning neuron.

Cooperation: The winning neuron determines the spatial location of a topological neighbourhood of excited neurons, thereby providing the basis for cooperation among neighbouring neurons.

Adaptation: The excited neurons adjust their individual values of the discriminant function in relation to the input vector. These adjustments are made through associated connection weights in such a way that the response of the winning neuron to the successive application of a similar input vector is enhanced.

One of the great strengths of SOMs is that they require no target vector in training. A SOM learns to classify the training data without any external supervision and is therefore called an unsupervised neural network. In SOMs, weights are initialized for each node. If the input space is D dimensional we can write the input patterns as:

$$X = (X_1, \dots, X_D) \quad (6.1)$$

and the connection weights between the input unit i and the neuron j in the computation layer can be written as:

$$W_j = \{W_{ji}; j = 1, \dots, N; i = 1, \dots, D\} \quad (6.2)$$

where, “ N ” is the total number of neurons. The best matching unit can be calculated by iterating through all the nodes and calculating the Euclidean distance between weight vector of each node and the current input vector. The node with a weight vector closest to the input vector will be the best matching unit (BMU). SOM map is created from a two-dimensional lattice of nodes. Each node is fully connected to the input layer and has a specific topological position in the map. Each node contains a vector of weights having the same dimension as the input vectors.

In SOM, several units compete for the current object thus forming clusters. The data of 61 samples was fed into the model for training.

6.3 SELF-ORGANIZING MAPS (SOM)

MODELLING AND ANALYSIS

Self-organizing maps are increasingly in demand and intensively used in different engineering and medical domains. There are number of platforms available to design, implement and test self-organizing maps. The neural network toolbox in Matlab provides a basic level of SOM modelling and implementation environment. However, a comprehensive implementation of the SOM algorithm was created by the developers at Laboratory of Information and Computer Science in the Helsinki University of Technology Finland, in the form of the “SOM Toolbox” [8], [9]. This toolbox is freely available and is compatible with Matlab. This toolbox covers the deficiencies in Matlab, while enables the user to take advantage of Matlab’s extensive general functions and commands, particularly for visualisation. Another good desktop application for SOM and cluster analysis is Viscovery SOMine developed by Viscovery Software GmbH, Vienna, Austria [10]. The main advantage of this software is that it provides a good visual cluster analysis. A licenced version of SOMine was obtained to perform the cluster analysis. In this chapter, SOM analysis is conducted using all the three platforms.

6.3.1 *SOM MODELLING*

In clustering problems, a neural network is typically used to group data on the basis of similarity. The Matlab Neural Network Clustering Tool is useful in selecting data, creating and training a network, and evaluating its performance. Here, a self-organizing map (SOM) consists of a competitive layer, which can classify a dataset

of vectors with any number of dimensions into as many classes as the layer has neurons. The neurons are presented in a 2D arrangement, which allows the layer to form a representation of the distribution and a two-dimensional approximation of the topology of the dataset. The network is trained with the SOM batch algorithm.

Experiments can be conducted with and without standardization of data. However, in SOM analysis it is usually recommended to standardize the data. Often the data is measured in different units and have different ranges. If data is not in standardized form then this may influence the clustering and the ultimate shape of the output map. The standardization of data is achieved by dividing the samples with their respective standard deviation as previously mentioned in chapter 4.

Network Architecture

There are 24 input variables measured for each of the 61 samples. The recommendations for the optimal map size of SOMs differ among studies. It depends on the purpose of the SOM and the goal of data exploration [11]. Although there is no theoretical principle to determine the optimum map size, quantitative indicators such as quantization error, topographic error and eigenvalues have proven to be useful tools to determine the optimal number of map units [12]. If the map size is too small it may not explain all the differences in the data and if the map size is too large, these differences become too small. With the given data, ten experiments were performed to find the optimal map size ranging between 5x5 till 15x15. A good approximate was obtained with a 10x10 map size. Therefore, the size of a two dimensional map was set to 100 (10x10), which gave enough space to present the topology of 61 samples.

Training

Training was performed using a batch unsupervised weight/bias method [13], i.e., weight and bias are updated at the end of an entire pass through the input data. After each epoch, weight and bias are updated according to their learning function.

Training stops when any of the following conditions is met:

- The maximum number of epochs (repetitions) is reached.
- Performance is minimized to the goal.
- The maximum amount of time is reached.
- Validation performance has increased more than maximum fail times since the last time it decreased (when using validation).

Training multiple times generates different results due to different (random) initial conditions and sampling. Figure 6.1 below shows the SOM neighbourhood weight distance through colour coding.

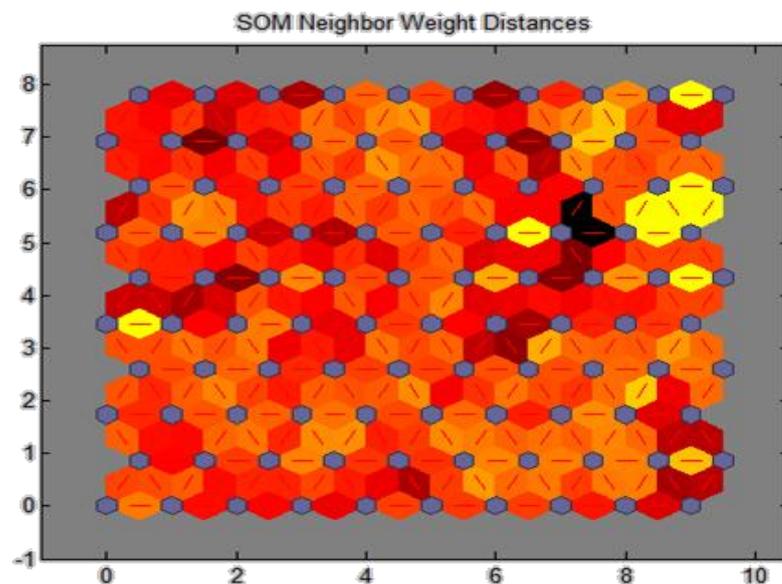


Figure 6.1: SOM neighbourhood weight distance

Here the blue hexagons represent the neurons and the red lines connect the neighbouring neurons. The colours of the cell containing the red lines indicate the distances between neurons where darker colours represent larger distances and the lighter colours represent smaller distances.

A group of light segments appear in the upper-right region, bounded by some darker segments. This grouping indicates that the network has clustered the data into groups. The light colour shows that the weights are closer together in this region. Similarly, darker segments show that the distances are larger. Figure 6.2 (a) below shows SOM neighbourhood connections (hexagon) and figure 6.2 (b) gives SOM layer, with each neuron showing the number of input vectors that it classifies.

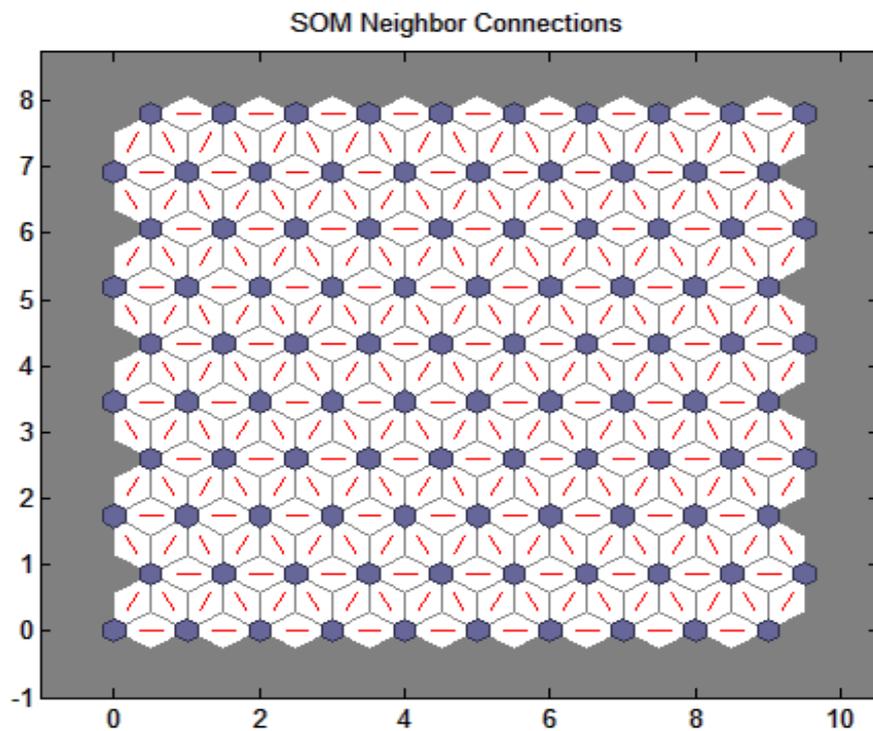


Figure 6.2: (a) SOM neighbour connections

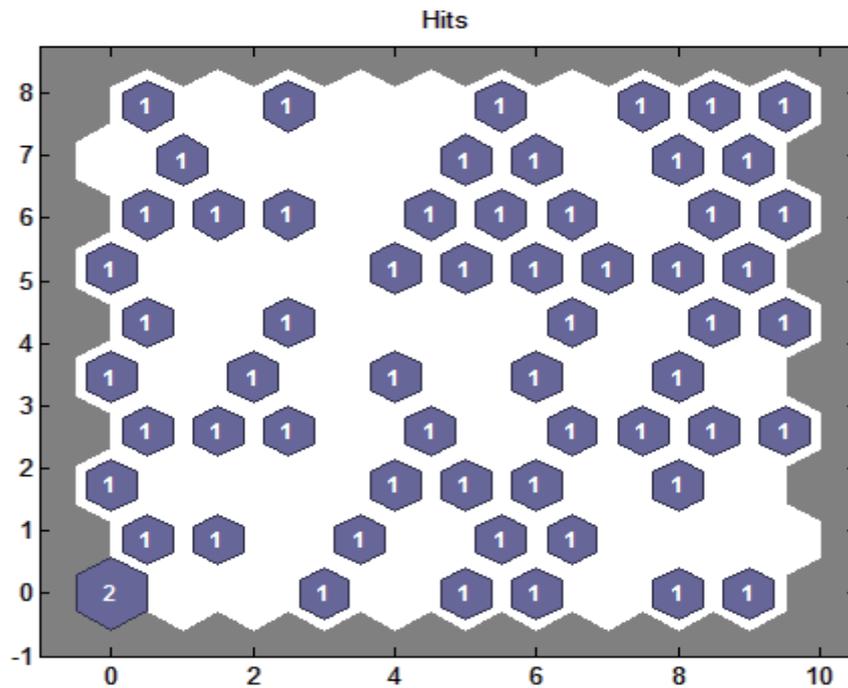


Figure 6.2: (b) SOM sample hits

There is a weight plane for each coordinate of the input vector (24 in this case). The weight plane for raw data (non-standardized) with map size=100 is shown in figure 6.3 below. These are visualizations of the weights that connect each input to each of the neurons. Darker colours represent larger weights. If the connection patterns of two or more inputs are very similar, it can be assumed that the inputs are highly correlated.

It can be seen from figure 6.3 below, that inputs 1, 2, 3, 4, and 5 have connections that are very similar. These are the vertebral heights L1, L2, L3, L4, and L5 respectively found highly correlated. Similarly inputs 6-11 corresponding to disc heights T12-L1, L1-L2, L2-L3, L4-L5, and L5-S1 are also highly correlated. Moreover, these two set of input variables: vertebral heights (input 1-5) and disc heights (input 6-11) are found somewhat correlated. Disc heights (input 6-11) have darker neurons in the map as compared to vertebral heights (input 1-5), which shows that

the neurons in disc heights are located apart from each other. Inputs 12-17 corresponding to disc signal intensities have a majority of dark colour segments showing that neurons are located far from each other. Similarly, inputs 18-21 corresponding to para spinal muscles and psoas muscle both have light and dark clusters with dark clusters in the majority. Inputs 22 and 23 are of subcutaneous fat signals, which have both light and dark clusters, with light clusters in majority. The last input represents weights for CSF having a majority of dark clusters.

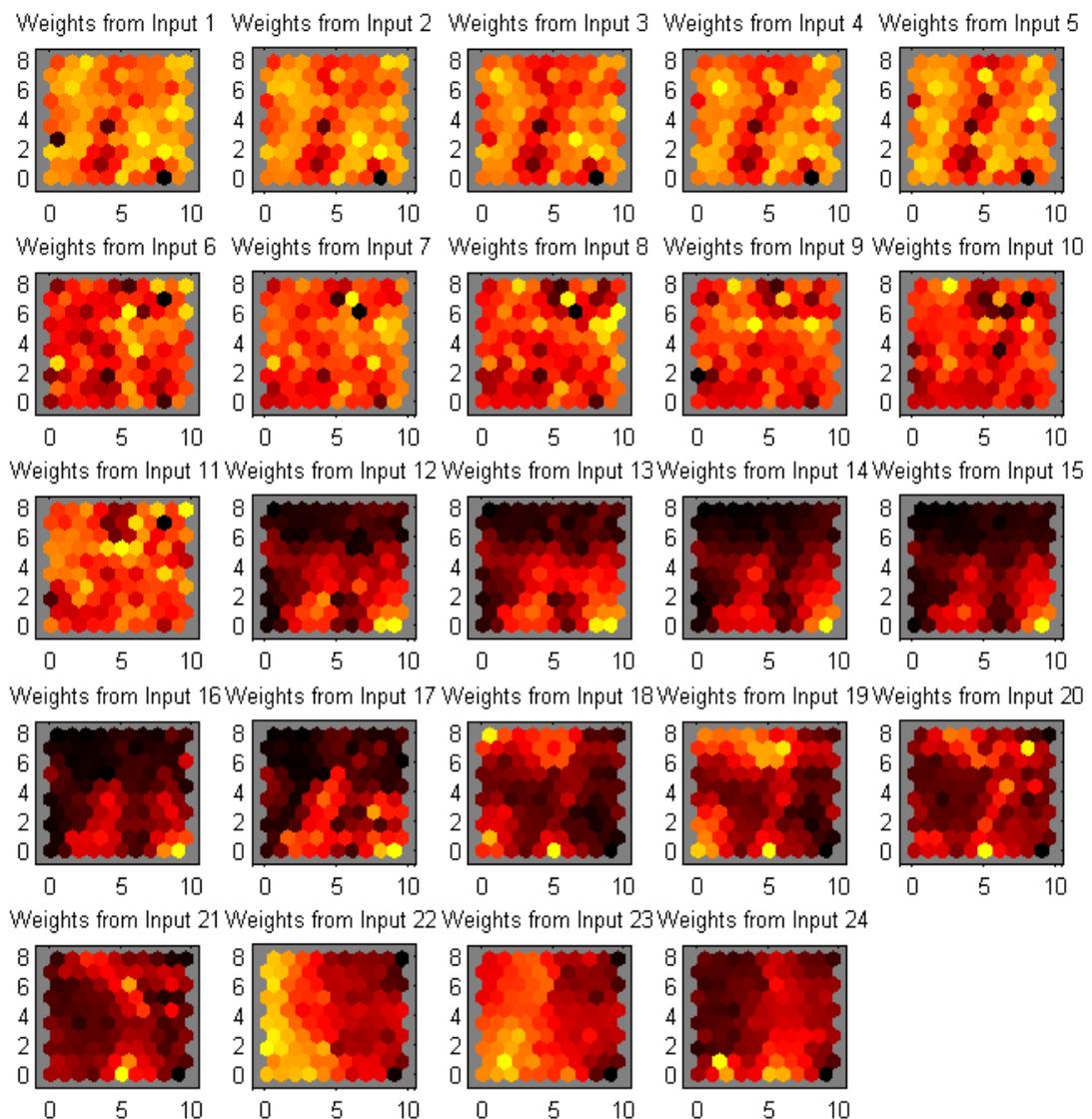


Figure 6.3: SOM input weight plane for non-standardized data with map size 100

Another SOM model is built with the standardized data. The results of the standardized SOM analysis are shown in figure 6.4. By comparing figure 6.3 and 6.4 it can be seen that there is little difference in the behaviour of each input. In figure 6.3 dark and light neurons are scattered throughout the map especially in input (1-11). However in figure 6.4, clusters become finer. This can be seen in figure 6.4 against inputs (1-11).

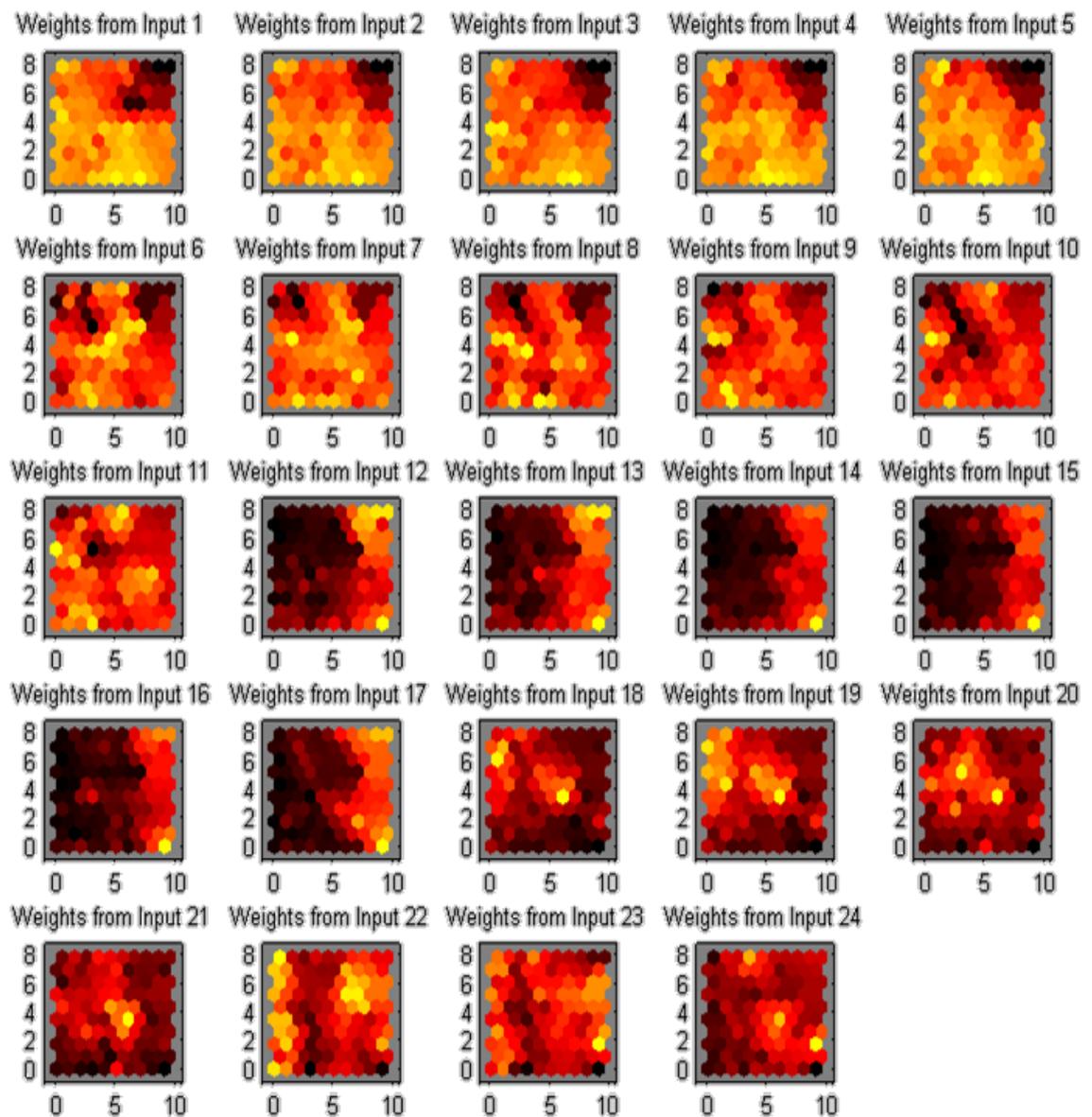


Figure 6.4: Weight planes for standardized data with map size 100

Here, two distinct clusters can be seen in inputs (1-5) corresponding to vertebral heights (L1-L5). The bigger cluster has almost 90% light (yellow) colour neurons whereas smaller cluster have very few dark colour neurons. The majority of light colour shows that the neurons are located very close to each other or in other words there is very little variation in vertebral heights. Hence, the majority of variation is seen in disc signal intensities (input 12-17), para spinal muscles (input 18 and 19), psoas signal (input 20 and 21) and CSF (input 24). A slight variation is seen in the subcutaneous fat signal (input 22 and 23) and disc heights (input 6-11) and least variation is seen in the vertebral heights (input 1-5).

Another common way of representing SOMs is through the unified distance matrix (U-matrix) [14]. In this representation, the Euclidean distance between the codebook vectors of neighbouring neurons is depicted with different colourings. These colourings are used to visualize the data in a high-dimensional space [15]. The SOM is coloured by the values of u-matrix elements. Usually a grey scale is employed for U-matrix representation, but in some more advance applications, multi colours like RGB are also used. The SOM toolbox compatible with Matlab provides a SOM U-Matrix representation with differencing colouring scales. This makes visual cluster analysis more informative. Again the size of the SOM is case dependent. For calculating the optimal number of map units in SOM, the following heuristics formula is used:

$$m = 5\sqrt{N} \quad (6.3)$$

Where m is the size of the map and N is total number of samples. Figure 6.5 below shows the SOM U-Matrix representation with non-standardized data, modelled and

implemented in the SOM-Toolbox. The U-matrix gives a visual representation of the distances between neurons in the input space. The scale adjacent to the maps gives the upper and lower ranges represented by red and blue respectively. An individual hexagon or an individual cluster in a variable can be compared with all other variables to recognize patterns. For example in figure 6.6 a blue cluster can be seen to represent very low values of variable 1. This cluster can be compared with all other variables to see how they behave when 1 variable has a very low value.

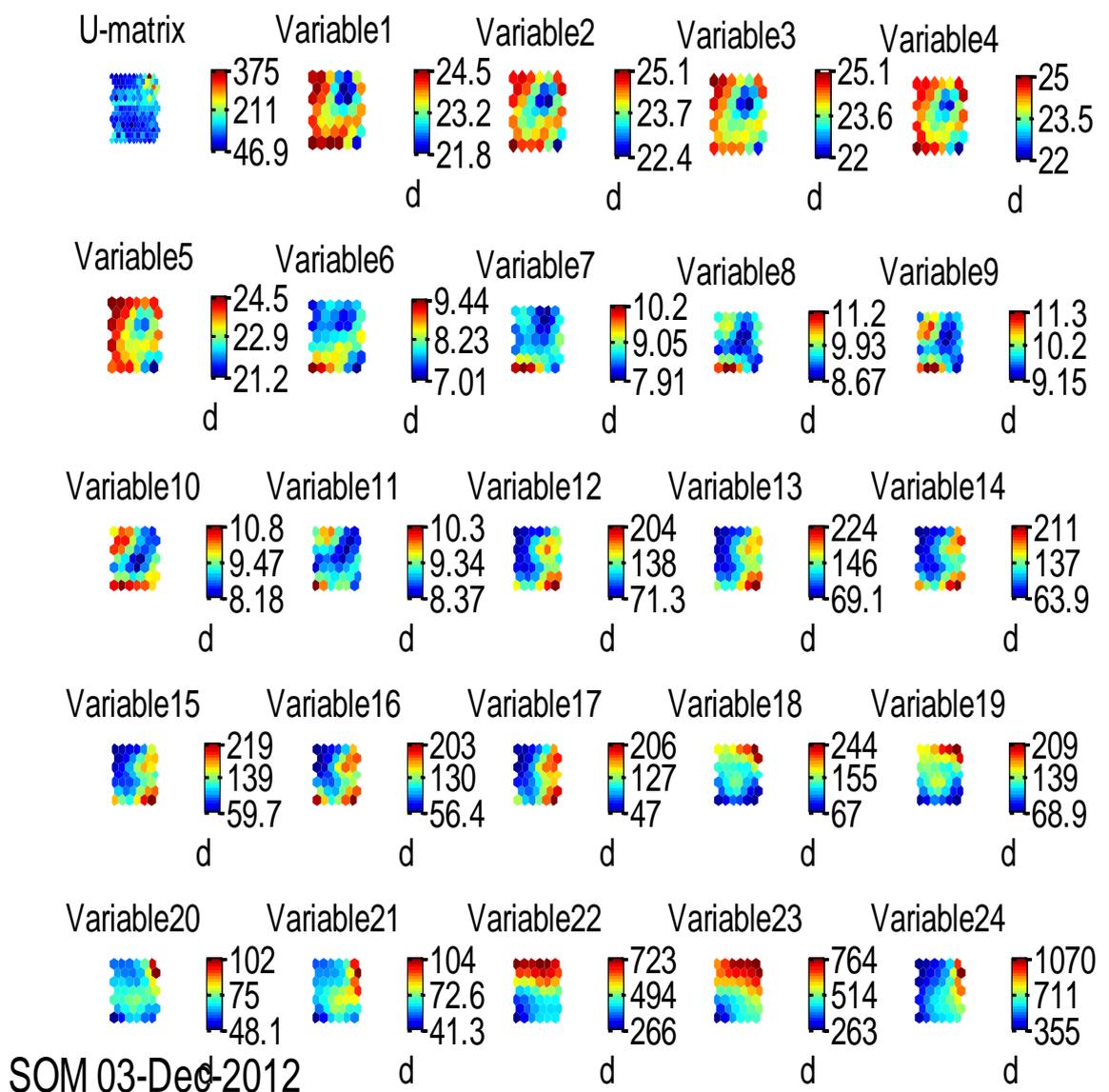


Figure 6.5: SOM U-matrix representations with non-standardized data

Figure 6.6 below shows an SOM U-Matrix representation with standardized data. Comparing the first two rows of variable 1 (having red colour), it can be concluded that when vertebral heights are very large, disc heights are moderate to large, disc signal intensities are very low, para spinal muscle and subcutaneous fat signals are high, psoas is moderate and CSF is on the lower side. In this way patterns are recognised by comparing the characteristics of a cluster in one variable with all others.

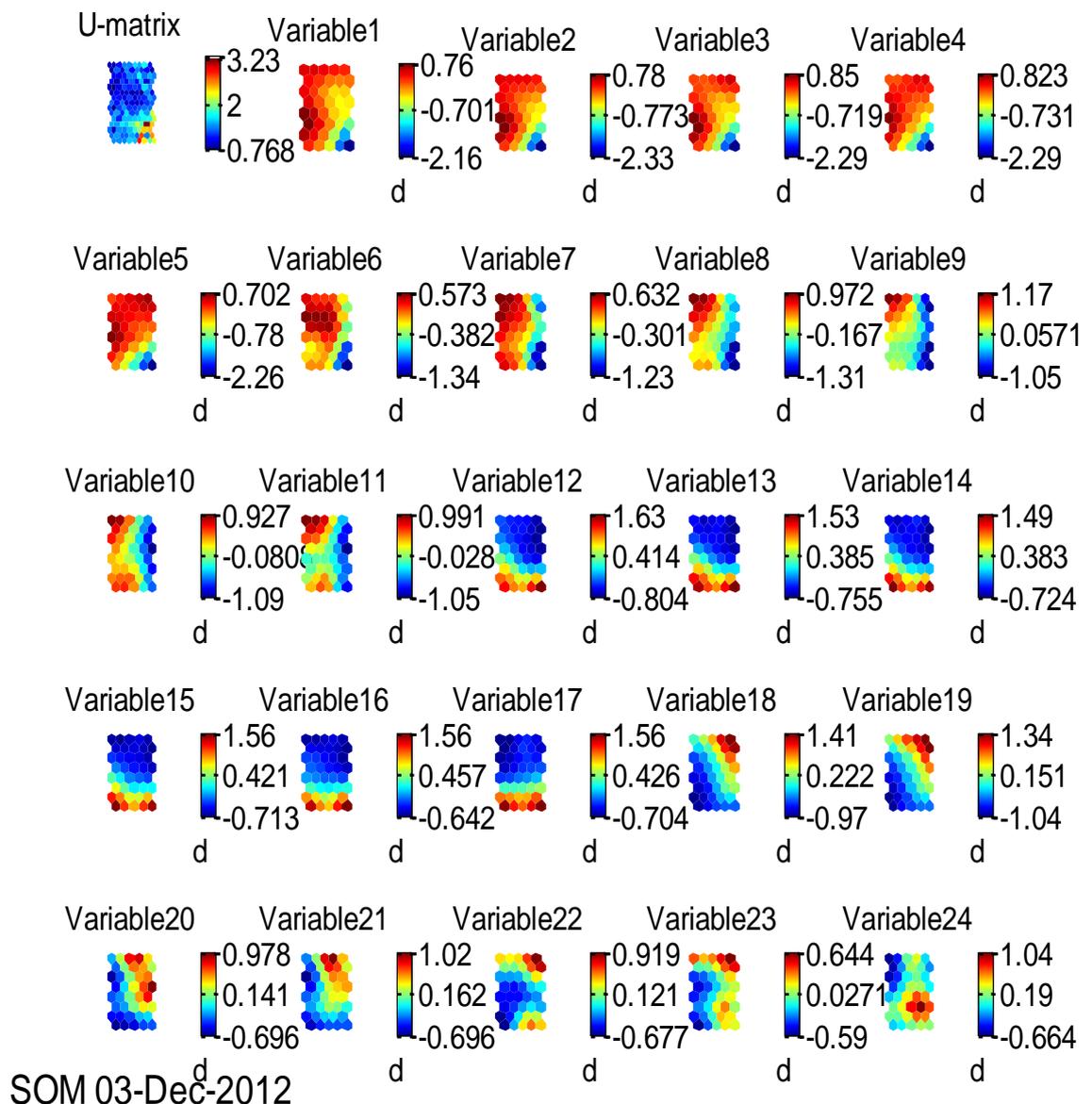


Figure 6.6: SOM U-matrix representations with standardized data

6.3.2 VISUAL CLUSTER ANALYSIS

Viscovery SOMine is a good desktop application for creating self-organizing maps and visual cluster analysis. Viscovery SOMine 6.0 was used for modeling, visualizing, classification, clustering and exploratory analysis of lumbar spine data. The measurements taken from the 61 lumbar MRIs were used to create a SOM model using Viscovery. The 24 features of each sample were fed to the model as inputs. Self-organizing maps of these 24 input features along with age and gender were created and the results are shown in figure. 6.7. The first two sub maps are of gender and age respectively. The other 24 maps are of inputs as labeled. These input features have different ranges so the inputs are standardized to get a better comparison.

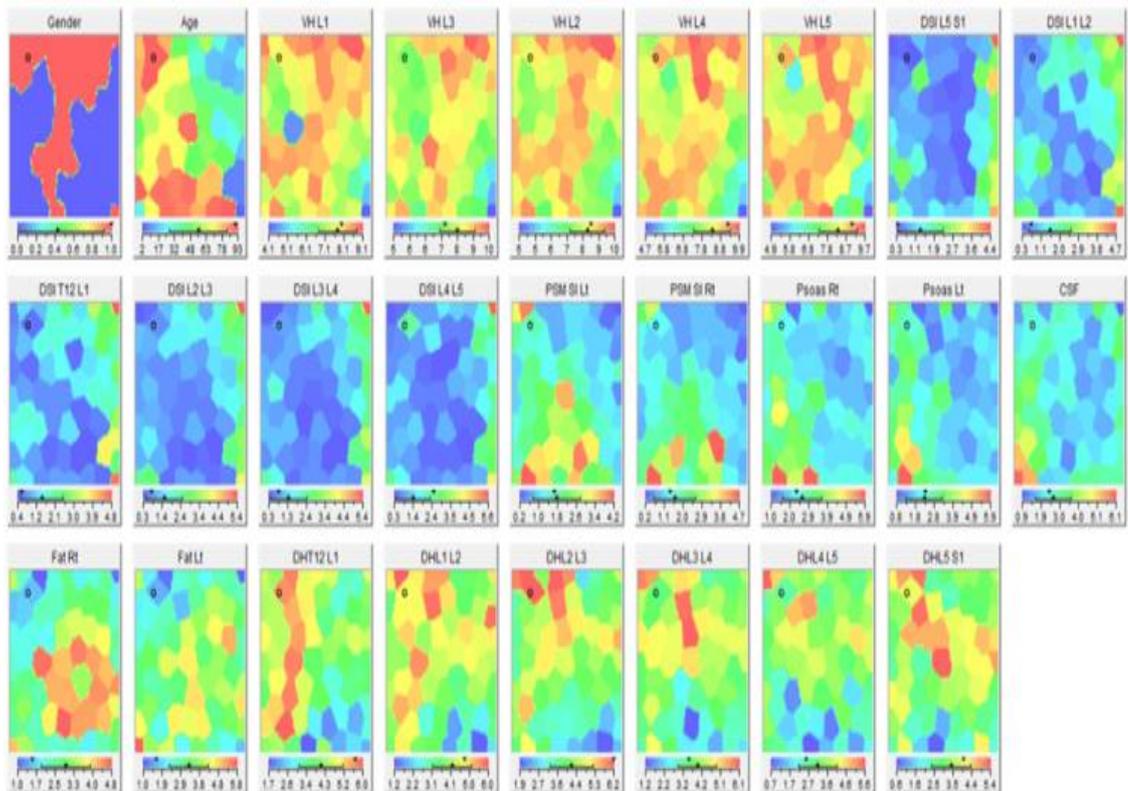


Fig. 6.7: SOM based on 24 input features along with age and gender

In the SOM map, each hexagonal cell represents an individual neuron. The neurons are drawn into distinct clusters during model training. Characteristics of the neurons are displayed by the intensity of the colors, with red color representing a high value and blue representing a low value. The SOM display is based on the red, green and blue (RGB) color model. A similar color shows that the neurons are close to one another or have similarity among them. In the component planes for individual variables, the color coding corresponds to actual numerical values for the input variables that are referenced in the scale bars adjacent to each plot. Here in figure 6.7, blue represents lower values, green represents intermediate values and red corresponds to higher values of corresponding input variables. By looking at this SOM map pairwise relationships among all input variables were studied. For example, those subjects who score very high on vertebral heights (red color) scores very low on disc signal intensities (blue color). In the first map (gender), red represents the male population and blue represents the female population.

6.4 CLUSTERING ANALYSIS FOR AGEING PATTERN OF SPINE

Cluster analysis creates groups, or form clusters of given data. Clusters are formed in such a way that the objects very similar to each other are put in a same group and distinct objects are placed in different groups. Measures of similarity depend on the nature of the application. The two clustering analysis techniques used in this chapter are given below.

6.4.1 WARD'S CLUSTERING METHODS

The Ward hierarchical clustering method has been widely used since its first description by Ward in a 1963 [16]. Unlike other clustering methods, this method uses an analysis of variance approach to evaluate the distances between clusters. This method involves an agglomerative clustering algorithm. Generally, this method produces very good results. In this method, cluster membership is evaluated by calculating the total sum of squared deviations from the mean of a cluster. A fusion criterion is to produce the smallest possible increase in the sum of squares error.

The hierarchical agglomerative cluster algorithm of Ward is characterized by the following steps: Starting with a clustering, where each single node forms a cluster by itself, in each step of the algorithm the two clusters with minimal distance (according to a specific distance measure technique) are merged. In Ward's method, the distance measure is based on the variance criterion. The goal is to have minimum variance within each cluster and large variance among the clusters. The two clusters are merged in each step according to the variance criterion mentioned above. This distance measure is called the Ward distance and is defined by equation 6.4 below:

$$d_{ab} = \frac{n_a \cdot n_b}{n_a + n_b} \cdot \|\bar{x}_a - \bar{x}_b\|^2 \quad (6.4)$$

where a and b denote two specific clusters, n_a and n_b denote the number of data points in the two clusters. \bar{x}_a and \bar{x}_b denote the cluster centroids and $\| \cdot \|$ is the Euclidean norm.

Beginning with the full distance matrix, a row and a column are stripped at every step. This continues until the matrix is completely cleared and a single cluster remains. The formulas shown in equation 6.5 and 6.6 are used to determine the centroid and the number of elements of the new cluster.

$$\bar{x}_a \text{ (new)} = \frac{1}{n_a + n_b} \cdot (n_a \cdot \bar{x}_a + n_b \cdot \bar{x}_b) \quad (6.5)$$

$$n_a \text{ (new)} = n_a + n_b \quad (6.6)$$

6.4.2 MODIFIED SOM-WARD CLUSTERING

SOM-Ward clustering is the modified form of the Ward method, which uses nodal characteristics of the map and a topological location of clusters. In this method, the distance matrix is initialized by taking into account the number of data records matching to the nodes of the map. The nodes that have more matching data records are weighted higher than nodes with fewer matching records. Here, the Ward distance measure has to be modified because it is likely that the SOM contains "empty" nodes in the map.

Algorithm

Let a and b be the two nodes for computing the distance. Let further be n_a and n_b be the number of data records that match the nodes a and b and \bar{x}_a and \bar{x}_b their node vectors. Then the distance d_{ab} is defined as follows

$$d_{ab} = \begin{cases} 0 & \text{if } n_a = n_b = 0, \\ \frac{n_a \cdot n_b}{n_a + n_b} \cdot \|\bar{x}_a - \bar{x}_b\|^2 & \text{otherwise.} \end{cases} \quad (6.7)$$

By this definition, it is ensured that in the first merge step, only nodes (and in the sequel clusters) with $n_a = 0$ ("empty clusters") are merged until only clusters with $n_a > 0$ remain. If there is at least one empty cluster, there exist many entries in the distance matrix with $d_{ab} = 0$, which are all candidates for the next merge step (since for all these the Ward distances are minimal). This implementation chooses those clusters among the rest that are Euclidean-nearest. For the SOM-Ward method, the distance measure is redefined as given in equation (6.8) below:

$$d'_{ab} = \begin{cases} d_{ab} & \text{if clusters } a, b \text{ are adjacent in SOM,} \\ \infty & \text{otherwise.} \end{cases} \quad (6.8)$$

In this way the locations of the clusters are observed by SOM-Ward distance. If the two clusters are not adjacent in SOM, they are never considered to be merged.

Cluster Indicator

An indicator is computed for each element of this hierarchical sequence of clustering that indicates the quality measure for each cluster count. This indicator is obtained by calculating the Ward distances for all possible numbers of clusters. Then the minimal distance between the clusters is normalized with an exponential function. The ratio of minimal distance between the two neighbored clusters makes up the cluster indicator. The cluster indicator helps in finding an initial clustering. A higher value of indicator points to a possibly good clustering. Let c be the number of non-empty nodes in SOM. The indicator $I(c)$ of c clusters can be calculated as:

$$I(c) = \max(0, I'(c)) \cdot 100 \quad (6.9)$$

where,

$$I'(c) = \frac{\mu(c)}{\mu(c+1)} - 1 \quad (6.10)$$

$\mu(c)$ and $\mu(c + 1)$ are the variances of cluster c and $c+1$ respectively.

$$\mu(c) = d(c).c^\beta \quad (6.11)$$

$d(c)$ is the Ward minimal distance between the clusters that was used to merge c clusters into $c - 1$ clusters, with $3 \leq c \leq C$. As stated above the $d(c)$ behaves like $c^{-\beta}$ where $-\beta$ is the linear regression coefficient for the data points: $[\ln(c), \ln(d(c))]$ (where $2 \leq c \leq C$).

$$-\beta = \frac{s_{\gamma\delta} - \bar{\gamma}\bar{\delta}}{s_{\gamma\gamma} - \bar{\gamma}^2}, \text{ where } \gamma = \ln(c) \text{ and } \delta = \ln d(c). \quad (6.12)$$

The first two cluster indicators are defined as $I(1) = 0$ and $I(2) = 0$ and for the SOM-Ward method, the cluster indicator is defined as $I(c) = 0$ for inversions at c clusters, i.e. when $d(c) < d(c + 1)$. For the Ward method, the distances $d(c)$ are monotonically decreasing with c increasing but that is not true for the SOM-Ward method. Also $I'(c)$ can be negative when $\mu(c) = 0$. In such cases, the displayed indicator is set to zero. When $d(c)$ is high, but $d(c + 1)$ is low, the c cluster is a good cluster because the next merge step (resulting in $c - 1$ clusters) would result in a high variance within the clusters.

6.4.3 CHARACTERISTICS OF THE CLUSTERS

Figure 6.8 below shows the SOM model built on the basis of 24 input variables after standardization and a map size of 5000. The size of the map is deliberately kept high to obtain a high resolution. By using Ward's method, three optimal clusters

were formed. These clusters were labeled as C1, C2 and C3 in figure 6.9 below. The cluster C1 is the largest cluster with mean age of samples of 52.76 years. The mean age of samples for cluster C2 is 81.75 years and for C3 it is 6.5 years. These clusters are formed by taking all 24 spine features into consideration with equal priority. It can be said at the age of 7, 53 and 82 lumbar spine features are very distinct in nature. Blue color indicates a lower value, green and yellow color represents moderate values, and red color shows high values. Scale below each sub map gives the corresponding range of that input variable. This scale also gives the average score of all the neurons in a sub map, which helped in comparing the individual score of specific sample within that sub map, and with all other sub maps.

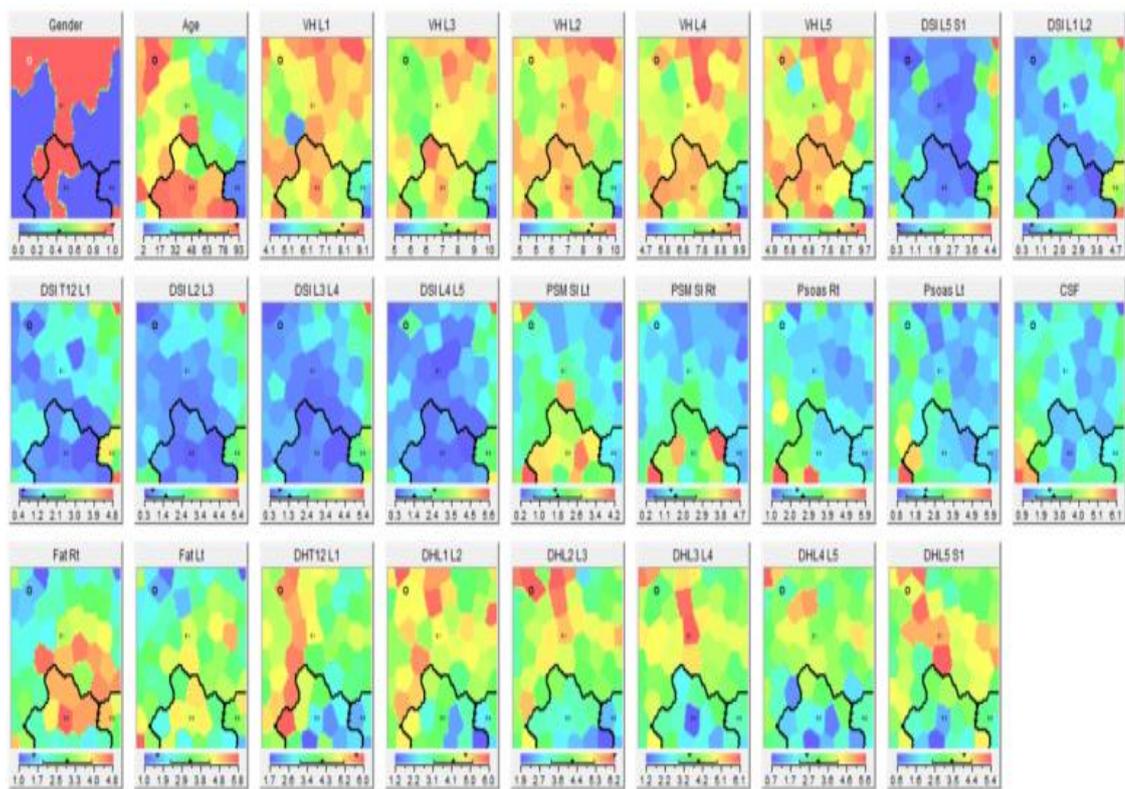


Figure 6.8: Ward clustering with standardized data

Another attempt was made to cluster data using the modified Ward's method. This modified method, termed here as SOM-Ward method apparently provides better

clustering. Figure 6.9 below shows the SOM-Ward map with clusters. Both of the methods produced three optimal clusters, but their composition was different. From SOM-Ward clustering, cluster C1 has a mean age of 50.41 years, C2 has a mean age of 74.5 years and C3 has a mean age of 6.5 years.

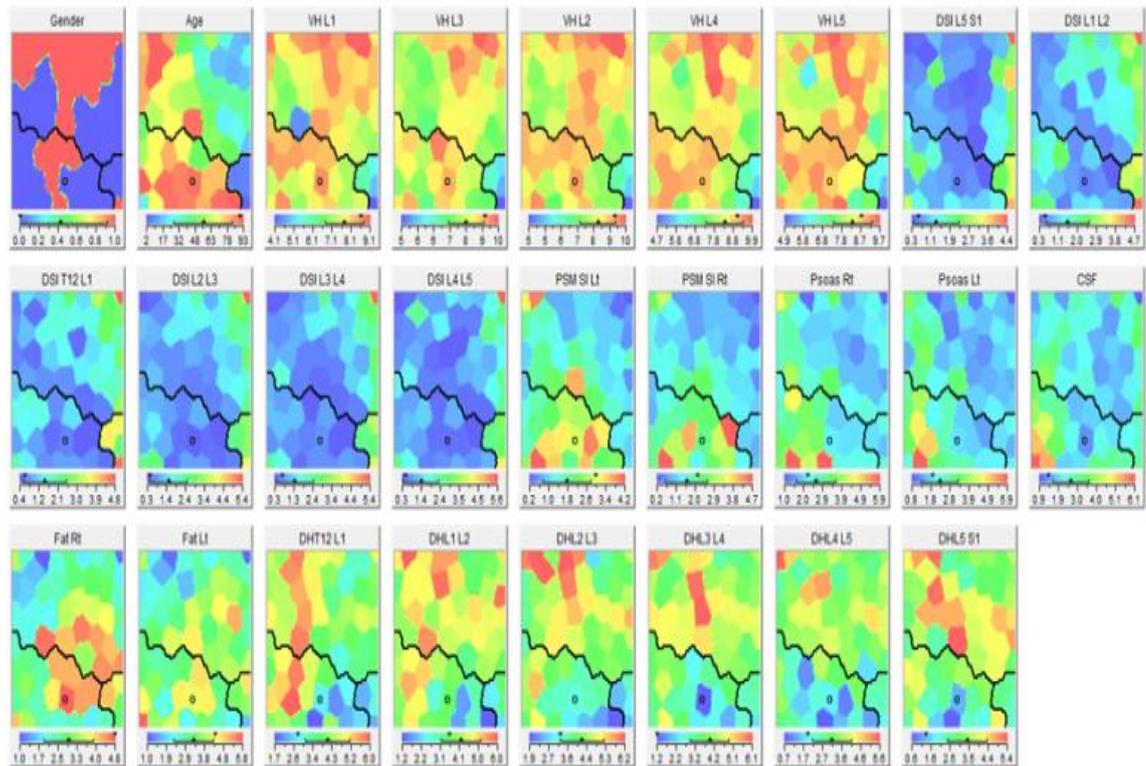


Figure 6.9: SOM-Ward clustering with standardized data

The relationships between variables are visualized by comparing the color patterns for individual maps. One of the benefits for using SOM is that the relationships between all of the variables in the model can be examined simultaneously or in pair-wise combinations. The comparison of clusters formed by both of the methods is shown below in figure 6.10. Both Ward and SOM-Ward produces three optimal clusters.

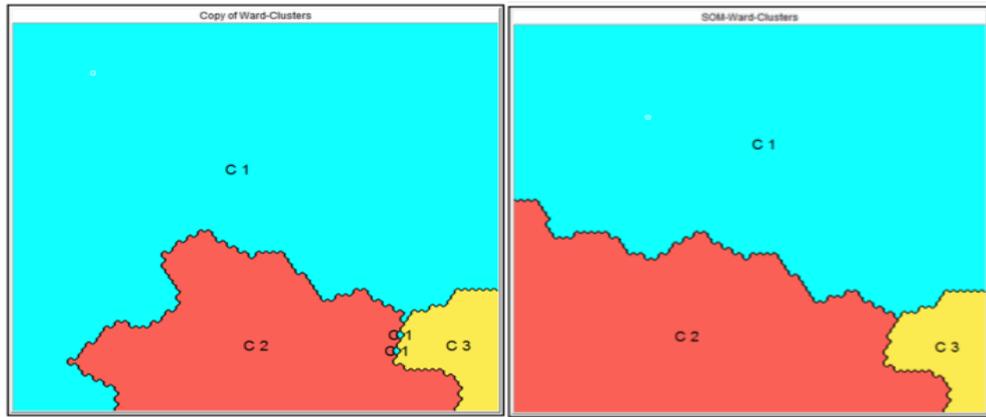


Figure 6.10: Clusters formed by using (a) Ward’s method (b) SOM Ward method

Cluster C3 represented by yellow appears to be the same in both methods whereas clusters C1 (represented by blue) and cluster C2 (represented by red) are different in two methods used. SOM-Ward shows a better clustering than the Ward method. In figure 6.10 (a) it can be seen the few samples belonging to cluster C1 are sandwiched between cluster C2 and C3, which is confusing and can be misleading. On the other hand the SOM-Ward method produces fairly smooth clusters as shown in figure 6.10 (b). A comparison of the characteristics of the three optimal clusters formed by Ward and SOM-ward is given in table 6.1 below.

Table 6.1: Characteristics of Ward and SOM-Ward clusters

Cluster	SOM-Ward Method			Ward Method		
	C 1	C 2	C 3	C 1	C 2	C 3
Abs. Profile Median	0.4014	0.6077	1.840	0.2756	0.9551	1.840
Frequency %	60.66	32.79	6.56	73.77	19.67	6.56
Gender	0.541	0.25	0.25	0.467	0.333	0.25
Age (Mean)	50.41	74.5	6.5	52.76	81.75	6.5
Vertebral Height L1	7.927	7.951	5.178	7.983	7.755	5.178
Vertebral Height L2	8.63	8.59	5.69	8.7	8.31	5.69
Vertebral Height L3	8.47	8.49	5.6	8.49	8.41	5.6

Vertebral Height L4	8.555	8.58	5.77	8.613	8.379	5.77
Vertebral Height L5	8.457	8.642	5.746	8.519	8.533	5.746
Disc Signal T12-L1	1.574	1.018	3.555	1.561	0.696	3.555
Disc Signal L1-L2	1.593	1.208	3.476	1.632	0.803	3.476
Disc Signal L2-L3	1.538	0.989	3.237	1.525	0.672	3.237
Disc Signal L3-L4	1.457	0.932	3.183	1.43	0.684	3.183
Disc Signal L4-L5	1.353	1.007	3.238	1.364	0.735	3.238
Disc Signal L5-S1	1.158	1.164	3.093	1.251	0.818	3.093
Paraspinal Muscle Left	1.408	2.614	1.15	1.526	2.977	1.15
Paraspinal Muscle Right	1.173	2.464	0.949	1.32	2.771	0.949
Psoas Left	1.967	2.903	1.894	2.245	2.483	1.894
Psoas Right	2.293	3.248	2.174	2.533	2.986	2.174
Cerebrospinal Fluid at L3	2.313	3.145	2.68	2.676	2.338	2.68
Subcutaneous Fat Left	3.086	3.587	3.426	3.199	3.497	3.426
Subcutaneous Fat Right	2.775	3.127	3.383	2.795	3.285	3.383
Disc Height T12-L1	4.364	3.938	2.66	4.417	3.454	2.66
Disc Height L1-L2	4.536	3.648	2.854	4.525	3.097	2.854
Disc Height L2-L3	4.666	3.712	2.882	4.632	3.204	2.882
Disc Height L3-L4	4.473	3.197	2.877	4.373	2.722	2.877
Disc Height L4-L5	3.64	2.342	2.45	3.475	2.097	2.45
Disc Height L5-S1	3.797	2.797	2.401	3.721	2.417	2.401

Some interesting patterns emerge from table 6.1 given above. By comparing the variations shown by features among all three clusters, it can be seen that vertebral heights L1,L2,L3,L4 and L5 increases with the age. This increase is notable between cluster C3 and C1 or in other words the vertebral height is increased by almost 50% from a mean age of 6.5 year to a mean age of 50.41 years. However, there is very little growth (0.5%) in vertebral height from a mean age of 50.41 years to 74.5 years (cluster C1 to Cluster C2). Similarly, disc signal intensities T12-L1, L1-L2, L2-L3, L3-L4,

L4-L5 and L5-S1 decrease with the age. A higher rate of descent is seen from cluster C3 to C2, compared to the rate of descent from cluster C1 to C2. However, disc heights T12-L1,L1-L2,L2-L3,L3-L4,L4-L5 and L5-S1 shows an increasing trend from cluster C3 to C1 and a decreasing trend from cluster C1 to C2.

Similarly, the growth and degeneration trends of para spinal muscles, psoas muscle, subcutaneous fat, and CSF with the age were evaluated. Using the SOM-Ward method, three optimal clusters were formed. These clusters were formed by giving all attributes equal priority. As many features were found to be correlated, it was interesting to study the behavior of one attribute vs. the others. For example after forming the self-organizing map, clustering was performed by prioritizing the gender attribute over the others. Figure 6.11 below shows the SOM-Ward clustering performed on the basis of gender.

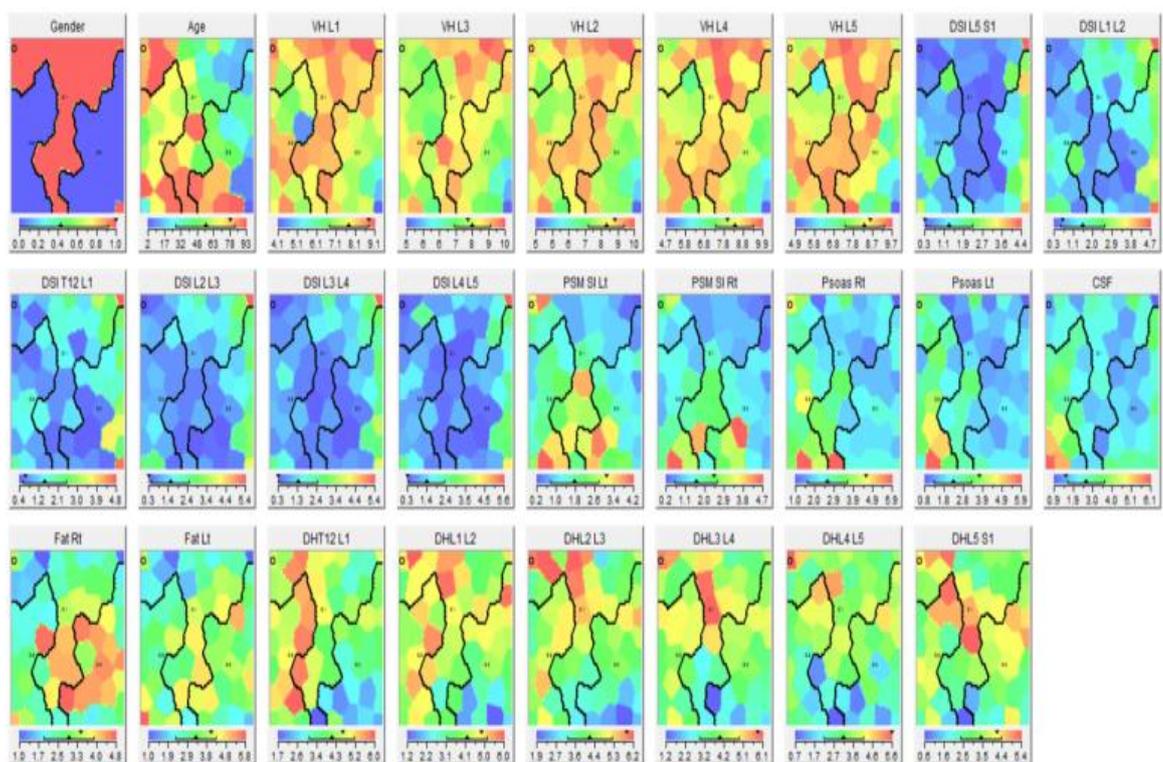


Figure 6.11: SOM-Ward clustering on the basis of gender

Both the Ward and SOM-Ward methods of clustering were employed on the basis of gender. The Ward's method of clustering produced two clusters, one male cluster (mean age 55.85 years) and one female cluster (mean age 55.11 years). Interestingly, the number of clusters formed by SOM-Ward clustering on the basis of gender was three. The first map in Figure 6.12 (a) below gives gender a representation by the SOM-Ward method. In this map, the blue cluster represents male samples whereas red and yellow clusters correspond to female samples. There were two clusters for female and one cluster for male. The mean age of samples in the male cluster (C1) was 58 years. The mean age for the female cluster (C2) was 43.71 years and for (C3) it was 67.53 years. This shows that the female spine at the age of 44 is somehow distant to the female spine at age 68. This is may be due to the fact that women reach menopause in their late 40's, which affects the anatomy of their spine.

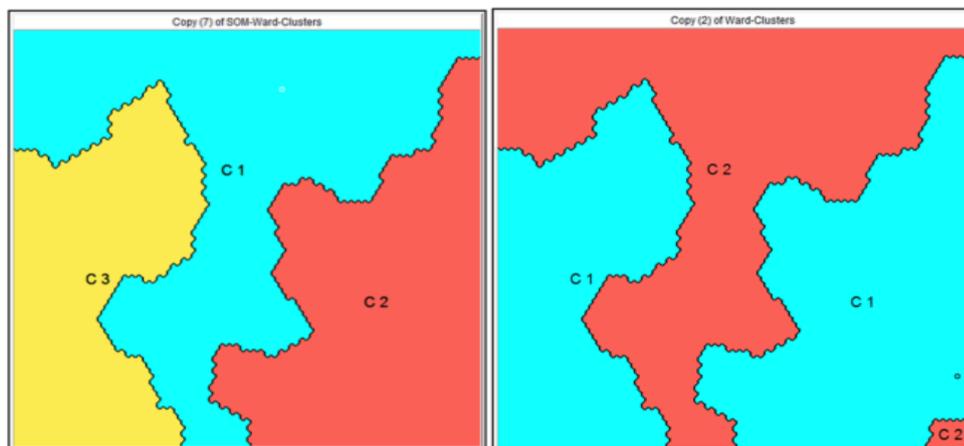


Figure 6.12: Clustering on the basis of gender (a) SOM-Ward (b) Ward's method

The characteristics of each cluster formed by the two methods were noticed. SOM-Ward clustering produced three clusters for the gender (C1, C2. and C3). The profile of cluster C1 is shown in figure 6.13 (a) below. This is the largest cluster having 25

samples with mean age 58 years. The key characteristics of this cluster were that it scored high on vertebral heights, and low on DSL5-S1 and CSF.

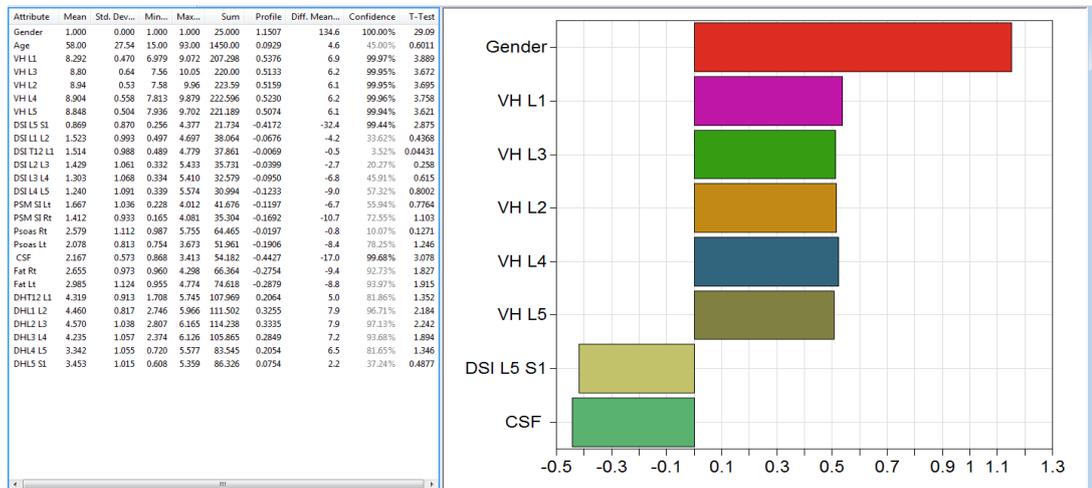


Figure 6.13: (a) Group profile of cluster C1 with SOM-Ward clustering on the basis of gender

The profile of cluster C2 is shown in figure 6.13 (b). This cluster consists of 21 samples with mean age of 43.71 years. The key characteristics of this cluster were that it scored very low on vertebral heights, disc heights, and psoas, and scored high on DSIL-S1 and fat right.

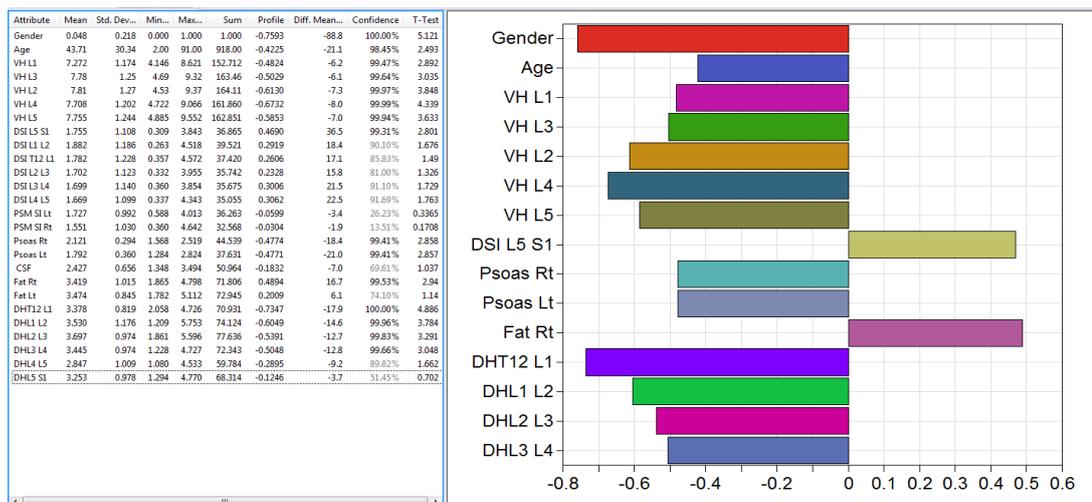


Figure 6.13: (b) Group profile of cluster C2 with SOM-Ward gender clustering

The profile of cluster C3 is shown in figure 6.13 (c). This cluster consists of 15 samples with mean age of 67.53 years. The samples of this cluster scored high on psoas, DHT12-L1, and CSF.

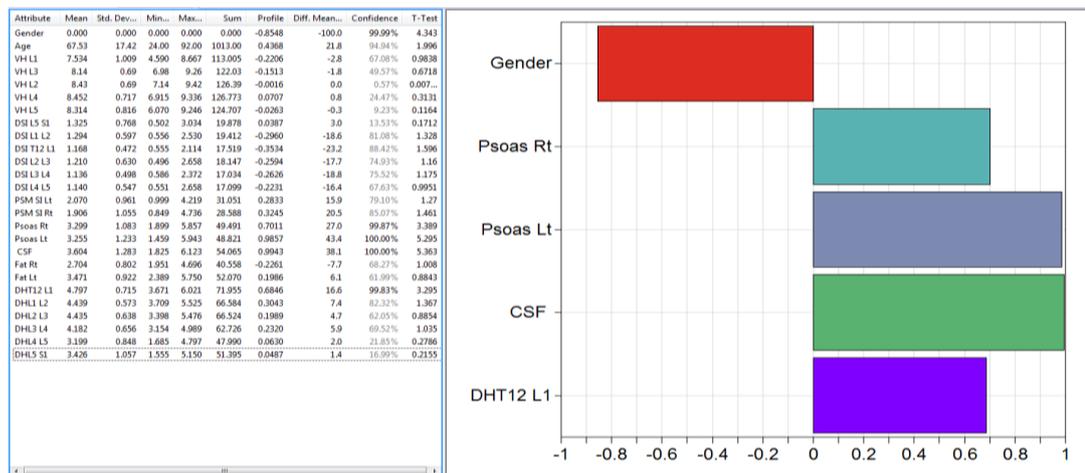


Figure 6.13: (c) Group profile of cluster C3 with SOM-Ward gender clustering

By considering age as a key attribute, nine clusters were formed by using the red-green-blue (RGB) color mode. These nine age clusters represent different phases of change in spinal features. In figure 6.14 (a) clusters are labeled with the samples' mean age. A black boundary separates clusters that are physically close to each other. The red line (of neurons) symbolizes a wall, meaning that clusters are not physically close to each other but they are adjacent for better representation in 2D. The change in the intensity of the colors depicts changes in the spinal features with age. For example, a gradual change of the blue color in left half of figure 6.14 (b) shows a gradual change in the lumbar spine from a mean age 15.83 years to 27.5 years. In right half of figure 6.14 (b) a steep change of color from yellow to red is seen, which shows that there is a steep change in the lumbar spine features from a mean age of 68.5 years to 80 years.

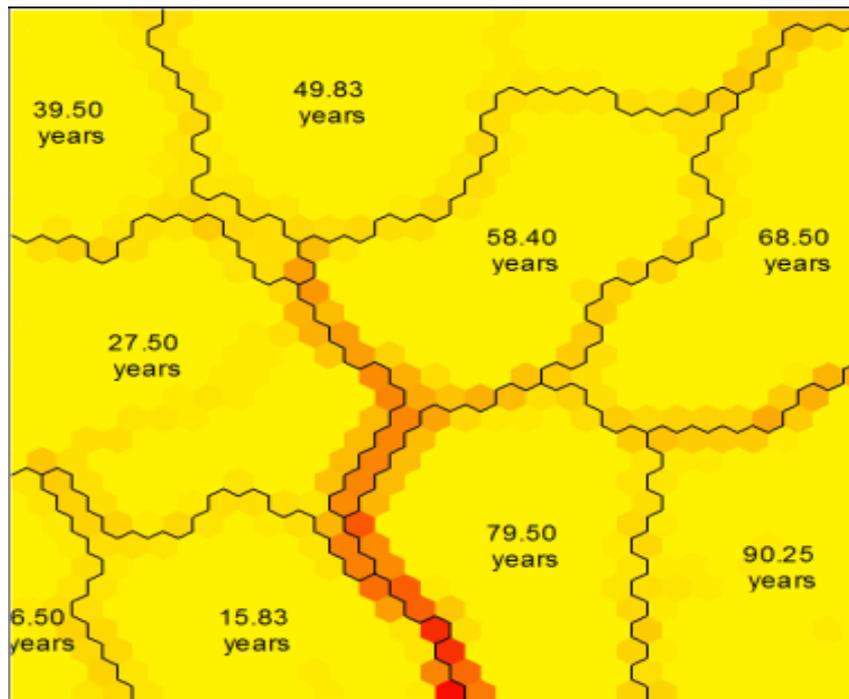


Figure 6.14: (a) Clustering (single color) on the basis of age considering all attributes

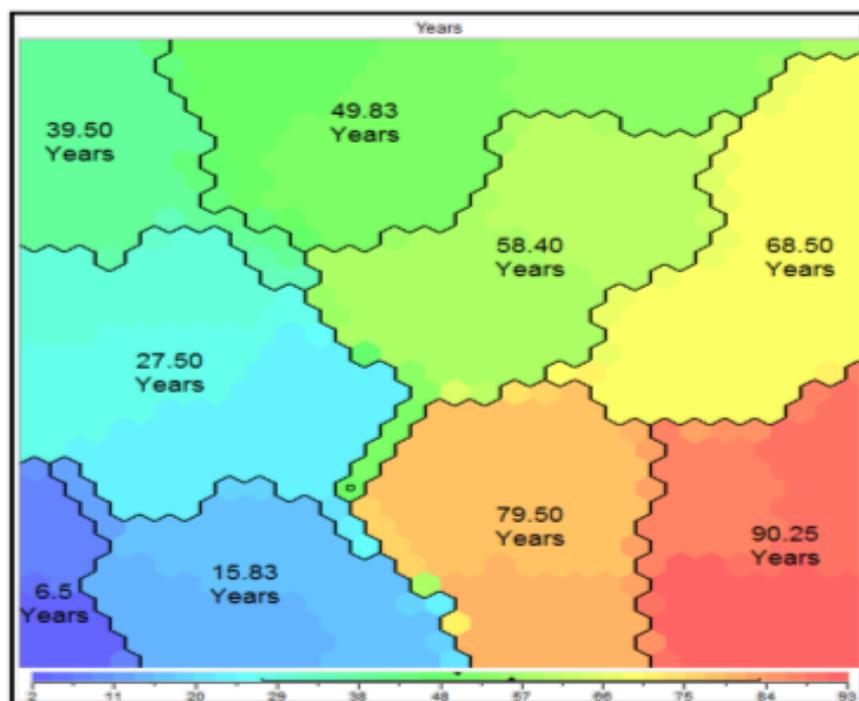


Figure 6.14: (b) Clustering (multicolor) on the basis of age considering all attributes

In this chapter, SOM was successfully applied for the visual analysis of age related variations seen in the lumbar spine. This 2D representation of multivariate data

proved easy to understand and gave meaningful information. A number of experiments were performed by using both Ward and SOM-Ward methods and varying the priorities of input variables and the number of clusters. For example, SOM-Ward clustering was performed on the basis of age with 3,4,5,6,7,8,9, and 10 clusters. Then the clusters were made by prioritizing age and vertebral heights only. In next attempt, clustering was made on the basis of age and CSF, then age and VH, age and PSM, VH and DH, DH and DSI, and so on. In each of these experiments, the numbers of cluster were varied to get a good approximation of patterns among the clusters. The profile of each cluster was observed to understand the characteristics of that cluster. In this way, relationship among the variables and their pair-wise combinations were understood.

From SOM analysis, we were able to study the behaviour shown by an individual sample and the groups. Whenever a new sample is added to the map, it is given a specific position in the map based on its MRI scores. For example, if a new male sample is having age in the 40s is assigned a place in the cluster with mean age in the 60s. It therefore shows that his MRI scores (features) matched more to the age 60s cluster. In other words, his spine starts to degenerate early for his age. In this way, healthy and problematic spines can be easily classified, which will eventually help clinicians to better diagnose and make decisions.

The two clustering methods used in addition to self-organizing maps give detailed characteristics of the potential three phases of spinal ageing. A comparison of features between mean ages 7, 50 and 75 gives a clear indication of which features increase with age, remain constant and decrease with age. The percentage of

change in features from one age cluster to another was also calculated. Knowing the amount of change in features with age will help us in setting the standards for age related changes.

6.5 SUMMARY

In this chapter SOM models were presented using different platforms. Self-organizing maps provided a good visual representation of the multivariate lumbar spine data, which helped in understanding the age related variations seen in the lumbar spines. SOM with colour coding particularly helped in formulating the relationships among age, gender and 24 lumbar spine characteristics. In addition to the SOM analysis, the results of Ward and modified Ward clustering techniques were presented in order to quantify the norms of different lumbar spine features among different groups.

REFERENCES

- [1] Alpaydin, Ethem. "Introduction to machine learning", Chapter 1: Introduction, MIT press, (2004): 10-11.
- [2] Carpenter, Gail A., and Stephen Grossberg. "The ART of adaptive pattern recognition by a self-organizing neural network." *Computer* 21, no. 3 (1988): 77-88.
- [3] Bhaskar, Harish, David C. Hoyle, and Sameer Singh. "Machine learning in bioinformatics: A brief survey and recommendations for practitioners." *Computers in biology and medicine* 36, no. 10 (2006): 1104-1125.
- [4] Saeys, Yvan, Iñaki Inza, and Pedro Larrañaga. "A review of feature selection techniques in bioinformatics." *bioinformatics* 23, no. 19 (2007): 2507-2517.
- [5] Kohonen, Teuvo. "Self-organizing maps," Springer Series in Information Sciences 3rd edition, Springer, 30, (2001).
- [6] Barlow, Horace B. "Unsupervised learning." *Neural computation* 1, no. 3 (1989): 295-311.
- [7] G.A. Carpenter, S. Grossberg, "Pattern recognition by self-organizing neural networks". MIT Press, (1991).
- [8] Himberg, Johan, Esa Alhoniemi, and Juha Parhankangas. SOM toolbox for Matlab 5. Helsinki, Finland: Helsinki University of Technology, (2000).

- [9] Vesanto, Juha, Johan Himberg, Esa Alhoniemi, and Juha Parhankangas. "Self-organizing map in Matlab: the SOM Toolbox." In Proceedings of the Matlab DSP conference 99, (1999): 16-17.
- [10] "Viscovery SOMine 6.0", Viscovery Software GmbH, Vienna, Austria.
<http://www.viscovery.net/>
- [11] Ultsch, Alfred, and H. Peter Siemon. "Kohonen's Self Organizing Feature Maps for Exploratory Data Analysis." (1990): 305-308.
- [12] Juha Vesanto, Johan Himberg, Esa Alhoniemi, and Juha Parhankangas, "SOM Toolbox for Matlab 5," Technical Report A57, ISBN 951-22-4951-0, University of Technology, Helsinki, Finland, April (2000).
- [13] Demuth, Howard, Mark Beale, and Martin Hagan. "Neural Network Toolbox™ 6." User Guide, (1992).
- [14] Ultsch, Alfred, "U*-matrix: a tool to visualize clusters in high dimensional data". Fachbereich Mathematik und Informatik, (2003).
- [15] Ultsch, Alfred. "Maps for the visualization of high-dimensional data spaces." In Proc. Workshop on Self organizing Maps, (2003): 225-230.
- [16] Ward Jr, Joe H. "Hierarchical grouping to optimize an objective function." Journal of the American statistical association 58, no. 301 (1963): 236-244.

7

CONCLUSION AND FUTURE WORK

7.1 OVERVIEW

7.2 RESULTS FROM PRINCIPAL COMPONENT AND FACTOR
ANALYSIS

7.3 COMPARISON OF RESULTS WITH STATISTICAL ANALYSIS

7.4 RESULTS FROM NEURAL NETWORK MODELLING

7.5 RESULTS FROM SOM AND CLUSTERING

7.6 FINAL PROTOTYPE OF THE SYSTEM

7.7 FUTURE RESEARCH DIRECTIONS

7.1 OVERVIEW

This chapter summarises the significant findings and concludes with a comparative analysis of results obtained throughout the thesis. The first two chapters introduced the problem and provided descriptions of several intelligent systems techniques and their application to complex medical datasets. The third chapter explained the nature of the dataset, data pre-processing techniques and provided information about the clinical features of the lumbar spine. The fourth chapter explored the existing correlation between the features and rank the spinal features on the basis of their clinical significance with respect to natural ageing. In chapter five a neural network model was presented. The model was shown to be capable of predicting spinal age and gender from lumbar spine characteristics. This chapter further extracted fuzzy rules to present the patterns in terms of linguistic rules. In chapter six, data samples were grouped into different clusters using self-organizing maps, which helped in setting the standards for lumbar spine features among different age groups. Most of the previous studies on lumbar spine ageing focused on 4-5 features at a time. In this study twenty four lumbar spine features were used, making it one of the most advanced studies to investigate changes in lumbar spine characteristics with natural ageing. A brief discussion on the results obtained from the different techniques is given in following sections.

7.2 RESULTS FROM PRINCIPAL COMPONENT AND FACTOR ANALYSIS

In this thesis, principal component analysis (PCA) was used for two reasons; first to reduce the dimensionality of the dataset, and second to visualize the multivariate complex dataset. Both these tasks were achieved successfully. Using PCA, redundancy was removed and the 24x61 dataset was transformed to a 3x61 dataset based on principal components. In reducing the dimensionality, only 11.5% of the information was lost. The first three (principal) components explained almost 90% of the variance in the dataset. Visualization of multivariable lumbar spine data produced informative relationships between the variables and features. Looking at the elements of the principal components and scores of samples against these components, it was possible to analyse specific age group performance against certain features. For example, by looking at the figure 7.1 (below) of the first vs. second principal component, it can be seen that as we move right to left on the x-axis (component 1), the age of the samples increase. As principal component 1 mainly constitutes disc signal intensities, it was concluded that disc signal intensities decrease as we age. From PCA it was found that males have higher vertebral heights than females.

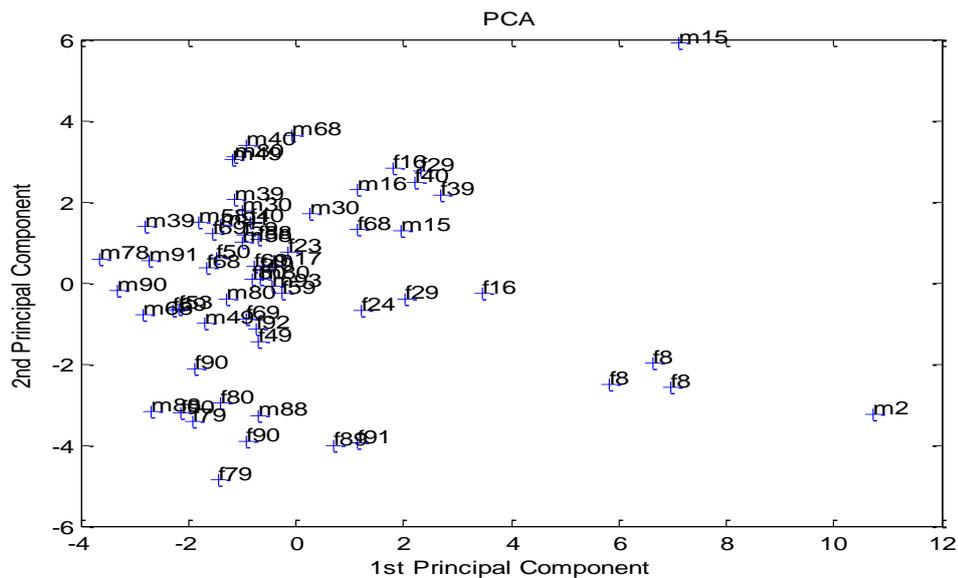


Figure 7.1: Results of principal component analysis

PCA also helped in anomaly detection, giving a visual representation of the data. By doing cluster analysis on PCA representation, samples with unusual readings were identified. For example, considering the two clusters formed on PCA representation as shown in figure 7.2 below, it can be seen that one sample (male age 15) is far away from the samples of his respective age group.

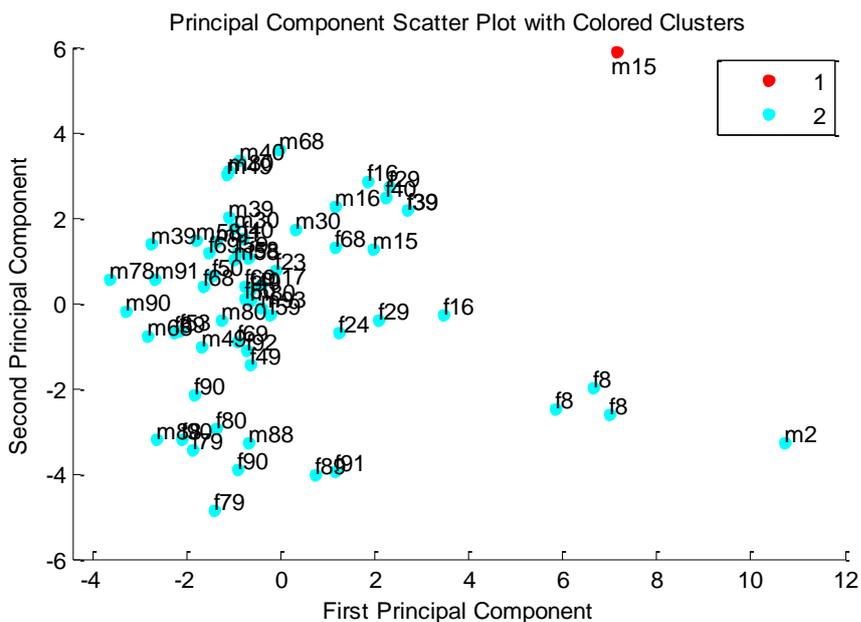


Figure 7.2: Anomaly detection with PCA

Closely examining the features of this specific sample shown in red, it was noticed that the disc signal intensity of this sample is almost twice that of the other samples in same age decade. Furthermore, one of the vertebral heights has an unusual reading. This unusually high reading of the disc signal intensity could be due the onset of some spinal disease. By using this PCA based clustering technique; unusual samples can be identified and eliminated. This model can be employed in the practical domain for classifying problematic spines, when a visual inspection of the lumbar spine MRI scan by medical specialists can easily overlook such changes.

One of the objectives of this research was to study the behaviour of different features and uncover the correlations between them. Factor analysis not only provided the significance of each input features but also provided the relationships among the lumbar spine features. Some interesting patterns were recognised from factor analysis. Figure 7.3 below shows the relationship between input features. Disc signal intensities were found to have a very strong correlation with natural ageing. Disc signal L2-L3 was the one most affected by the ageing. Disc heights and vertebral heights also showed a strong correlation with natural ageing. Vertebral height L3 and disc height L2-L3 were most prominent in their respective groups. Para-spinal muscles showed a moderate correlation with age; with left muscle scored slightly higher than right one. Psoas showed a very weak correlation whereas subcutaneous fat signal and cerebrospinal fluid (CSF) were almost non-correlated to age. Disc heights and vertebral heights were found to be correlated to each other. Similarly, psoas, para spinal muscle, and subcutaneous fat signals were also found to be correlated and had a negative correlation with disc signal

intensities. Disc and vertebral heights were negatively correlated with the cerebrospinal fluid.

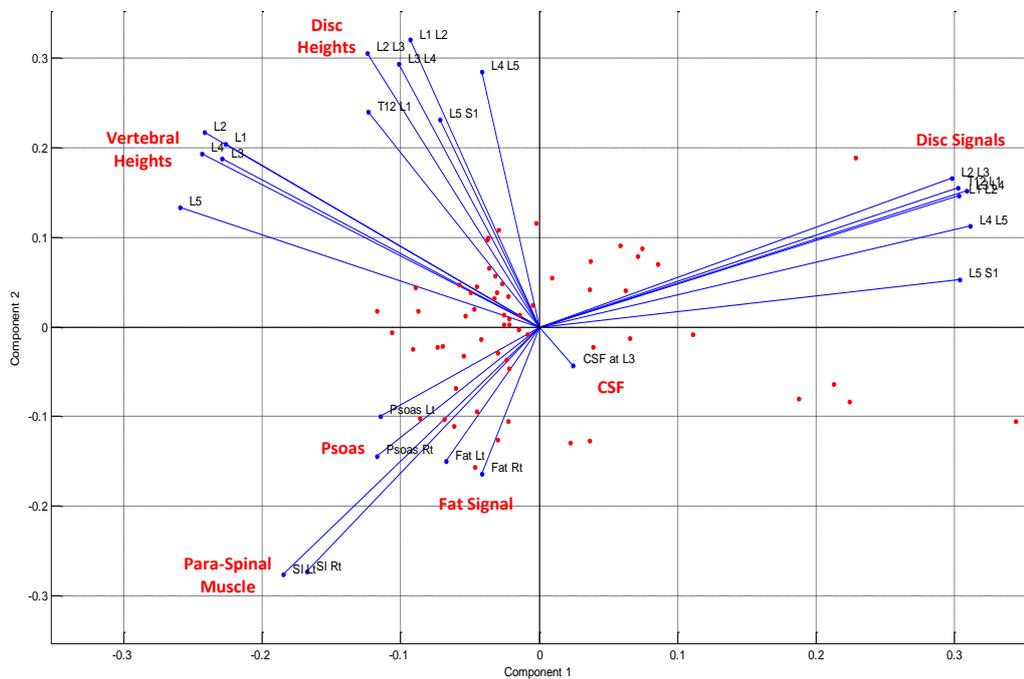


Figure 7.3: Results from factor analysis

7.3 COMPARISON OF RESULTS WITH STATISTICAL ANALYSIS

A statistical significance testing was performed to see the features that correlate with natural aging and that remain constant. The results are presented in table 7.1. Disc signals and disc heights were found to have a high correlation with age ($R^2 > 0.7$). Para spinal muscles, psoas muscles and vertebral heights were somewhat correlated with age ($0.3 < R^2 < 0.7$). Vertebral heights appear to have a low correlation with age overall; the L3 vertebra is found to be significant but the correlation is indeed low. Cerebrospinal fluid (CSF) and subcutaneous fat were non-correlated with age (age ($R^2 < 0.3$)).

Table 7.1: Correlation of spinal features with age

Variable		Correlation with age	p value	Significance
Disc Signal	T12 L1	-0.786	1.065×10^{-13}	**
	L1 L2	-0.818	1.412×10^{-15}	**
	L2 L3	-0.832	2.2×10^{-16}	**
	L3 L4	-0.811	3.844×10^{-15}	**
	L4 L5	-0.667	5.877×10^{-9}	**
	L5 S1	-0.613	1.896×10^{-7}	**
Disc Height	T12 L1	-0.783	9.174×10^{-14}	**
	L1 L2	-0.775	2.178×10^{-13}	**
	L2 L3	-0.765	7.482×10^{-13}	**
	L3 L4	-0.749	3.909×10^{-12}	**
	L4 L5	-0.639	3.037×10^{-8}	**
	L5 S1	-0.593	4.742×10^{-7}	**
Vertebral Height	L1	-0.215	0.1298	
	L2	-0.270	0.05522	
	L3	-0.369	0.007621	**
	L4	-0.264	0.06137	
	L5	-0.196	0.1688	
Para-Spinal Muscle	Right	0.588485	6.106×10^{-7}	**
	Left	0.6379052	3.213×10^{-8}	**
Psoas Muscle	Right	0.4446081	0.0003309	**
	Left	0.4171204	0.000825	**
Subcutaneous Fat	Right	-0.1699296	0.1904	
	Left	-0.1313895	0.3128	
CSF signal	at L3	0.03970446	0.7613	

These results are very similar to the results reported from a principal component analysis with exception of the vertebral heights. In PCA, vertebral heights were found to have a significant correlation with age. However, in this analysis vertebral heights were found to have a weak correlation with age. It was also found that there is a slight difference in the left vs. right signal intensities of the para spinal muscles, psoas muscles and subcutaneous fat. This finding was discussed with the

medical practitioners and they confirmed that the signal intensities of MRI are not usually consistent (on left and right sides) and the best approximate can be obtained by taking measurements from the mid line.

7.4 RESULTS FROM NEURAL NETWORK

MODELLING

In this thesis, a feed-forward artificial neural network (ANN) model was presented. The model is capable of estimating gender and age from spinal features. Gender was estimated with high accuracy. The results for spinal age estimation were not as good as for gender estimation but were acceptable. While interpreting the results and analysing the model accuracy, it was important to define the acceptable range for error.

The spinal age estimated by the ANN model was mapped onto age clusters. If the ANN output is mapped onto the correct cluster (having similar age samples), then estimation of ANN was assumed to be accurate. The number of clusters can be varied. With a lower number of clusters, the ANN model produced more accurate estimates. If we consider only three clusters (as discussed in the SOM analysis results) with mean ages 7, 50, and 75, the output from the ANN is mapped onto these three clusters. Based on the degree of belongingness, an output is assigned to one of these clusters. For example, a sample with actual age 60 is estimated by the ANN as having a spinal age 45; this sample will fall in the cluster of mean age 50. Although there is an estimation error of -15 years, the classification is correct. The ANN model will perform fairly accurately with 5-10 clusters but with more than 10

clusters it fails to achieve the required accuracy. This is due to the fact that a limited number of data samples were available for training and learning, not allowing the network to learn the patterns sufficiently well. Also, to improve the accuracy with 10 or more clusters, some other key features like height, weight, and body fat of the subjects need to be included in the model along with other spinal features.

7.5 RESULTS FROM SOM AND CLUSTERING

The self-organizing map (SOM) model proved versatile for the visual analysis of age related variations seen in the lumbar spine. A 2D representation of multivariate data through SOM proved easy to understand and provided meaningful information. SOM analysis made it possible to study the behaviour of samples individually as well as in groups. Data samples similar in characteristics were assigned a place close to each other in the map, thus forming the clusters. The two clustering methods used in addition to self-organizing maps provided detailed characteristics of different phases of spinal ageing. The formation of clusters helped in setting the standards for spinal features among different age groups. Figure 7.4 below shows the SOM-Ward representation with five clusters of mean ages 7, 30, 62, 70, and 87 years.

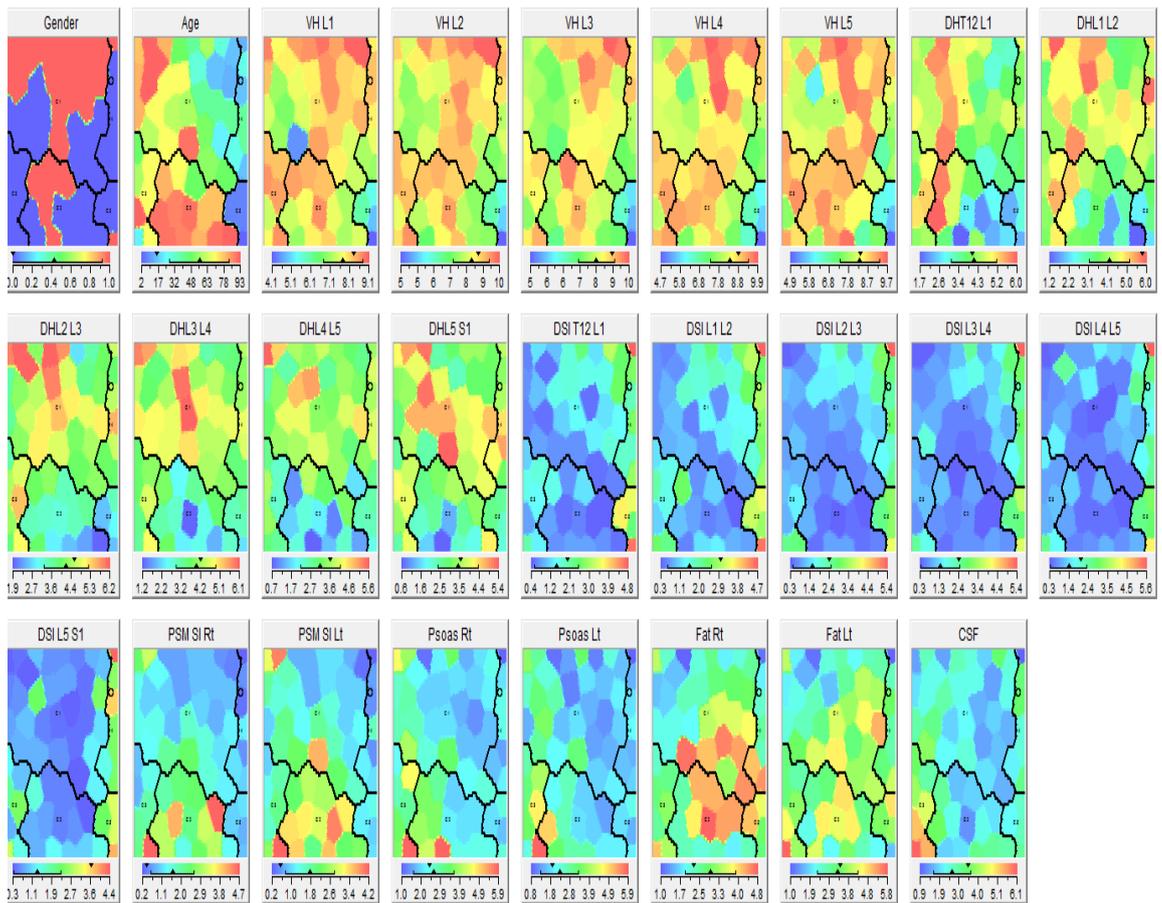


Figure 7.4: Results from SOM-Ward clustering with standardized data

A comparison of different spinal features between these five age groups gives an indication of which features increase with age, decrease with age, and remain constant with age. Figure 7.5 and 7.6 show variations in features between these 5 age groups.

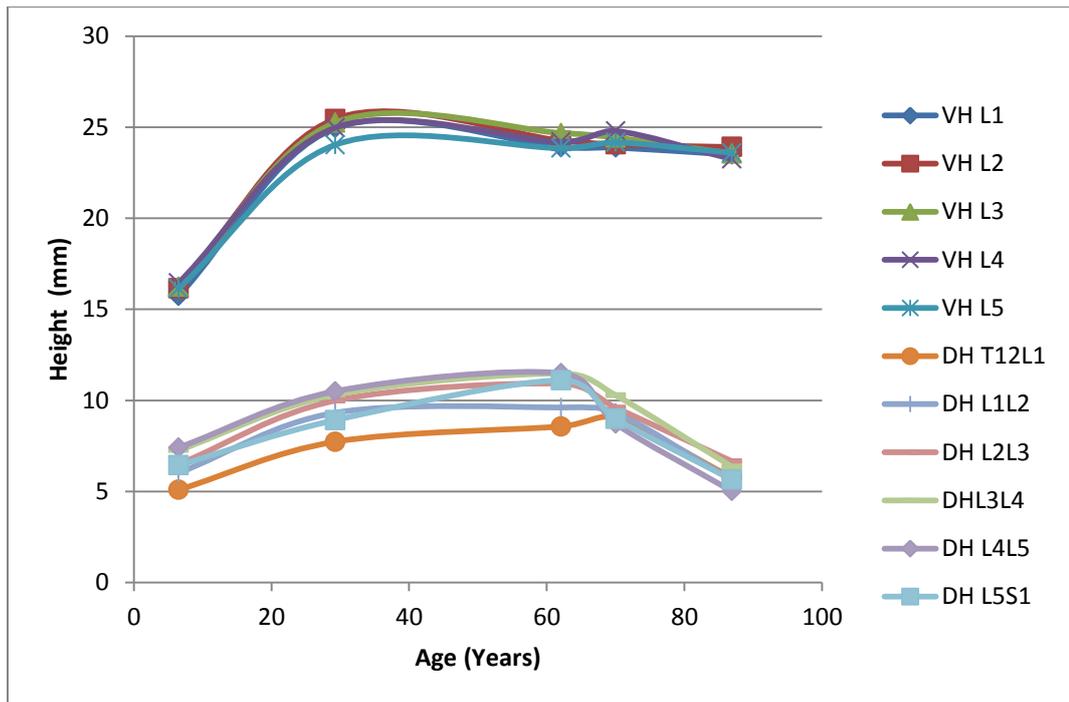


Figure 7.5: Changes in vertebral and disc heights among different age groups

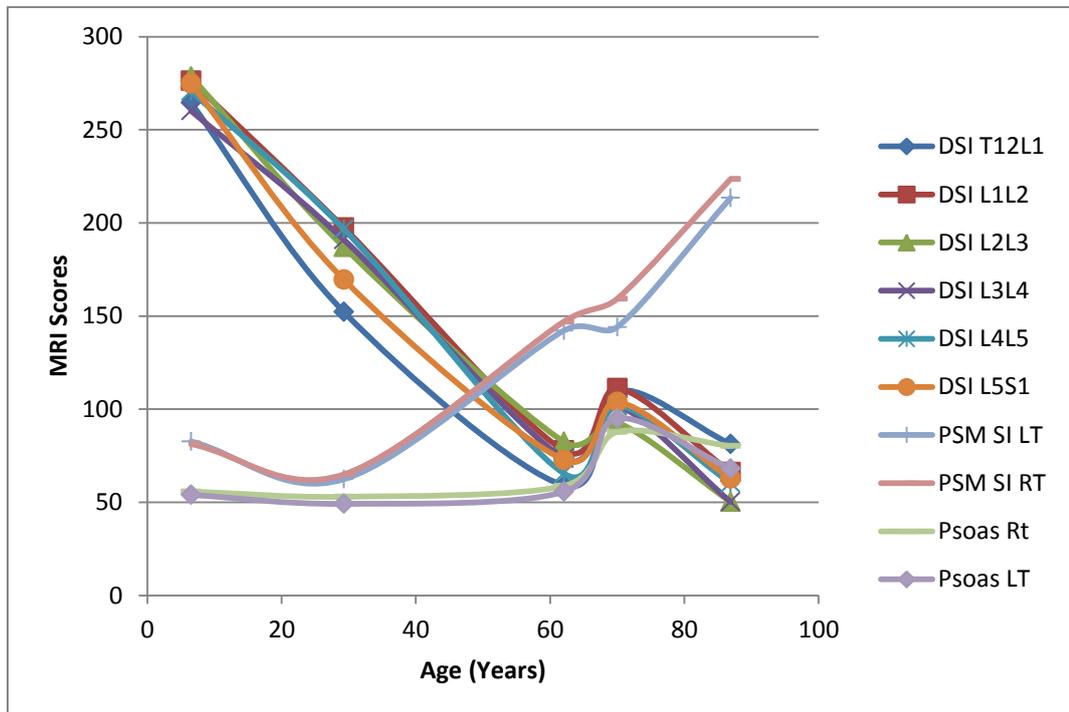


Figure 7.6: Changes in lumbar spine features among different groups

It can be seen from figure 7.5 above that the vertebral height increases gradually with age until the age of 30, then it remains the same until the age 60's and shows a slight degeneration after the age 60's. Similarly, the height of the lumbar disc increases with age until the age 30's, and begins to degenerate in the age 60's. From figure 7.6 above, it can be seen that disc signal intensities decrease consistently with natural ageing. The psoas muscle intensity increases in later ages whereas the para-spinal muscle signal intensities increase gradually with natural ageing. In this way, the effects of ageing on all the features were noted to set the reference values for spinal features in different age groups.

Whenever a new sample is added to the map, it is allocated a specific position in the map based on its MRI scores. For example, if a new male sample having age in the 40s is assigned a place in the cluster with mean age 70s. Therefore, it shows that his/her MRI scores (features) matched more to that of the age 60s cluster. In other word his/her spine exhibits degeneration earlier than expected age. In this way, healthy and problematic spines can be easily identified, eventually helping clinicians to better diagnose and make decisions.

7.6 FINAL PROTOTYPE OF THE SYSTEM

In this section, a prototype of the decision support systems is proposed. This decision support system is formed by the integration of all the data analysis techniques presented in this thesis. The block diagram of this decision support system is shown in figure 7.7 below. When a new MRI is performed, it is stored in the MRI database. Then this MRI is pre-processed, features are extracted and measured. These features are passed through a PCA/FA block to visualize the actual

location sample in the 2D plot. This will show how close the new sample lies to his/her age cluster. Then this information is propagated through a fuzzy inference system, neural network model, and self-organizing map model. The neural network estimates age and gender on the basis of spinal features. The fuzzy inference system estimates the spinal age by making decisions on the basis of if-then fuzzy rules.

Finally, the SOM allocates a specific place to the new sample in the map where characteristics of the input sample are matched with the characteristics of its respective cluster. Results from all these models are compared for any inconsistencies and are presented to medical specialist for drawing conclusion and making appropriate decisions.

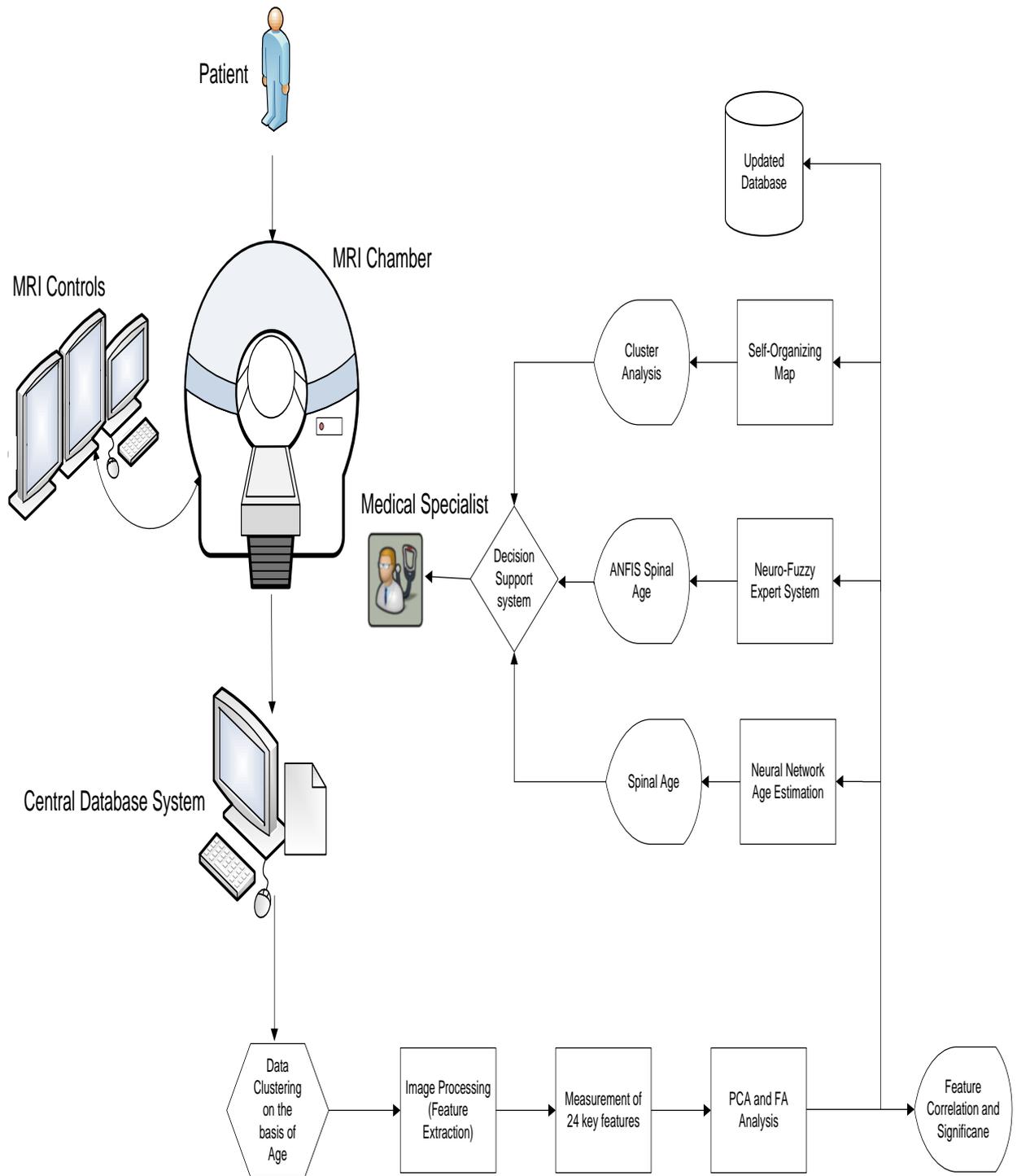


Figure 7.7: Final prototype of the system

Medical specialists often use lumbar spine MRI for clinical diagnosis. Due to complex nature of the spine and the amount of information presented in MRI, some small variations in spinal structures are likely to be overlooked by the naked eye. This proposed system can serve as an intelligent diagnostic assistant to counter verify the findings and decisions made by medical specialists. If there is any contradiction in the system output and the doctor's finding, this system will notify the doctor and request him to take a second look. The system also pinpoints the areas that have been overlooked. The results of the new sample are finally stored in the database. With each incoming sample, the system is matured (improved training and generalisation) and the database is update.

7.7 FUTURE RESEARCH DIRECTIONS

Since this was a pilot research study with a limited number of samples available to design, implement and test the models, more MRI samples are required to enhance the accuracy and reliability of the models and therefore the overall system proposed above. It will also be informative to study the demographic changes in the lumbar spine features. This can be achieved by collecting samples from different geographic locations and studying the ageing effect on the lumbar spine features, using the methods employed in this thesis.

This research was based on 24 lumbar spine features: vertebral heights (VHL1, VHL2, VHL3, VHL4 and VHL5), disc heights (DHT12-L1, DHL1-L2, DHL2-L3, DH-L4L5 and DHL5-S1), disc signal intensities (DSIT12-L1, DSIL1-L2, DSIL2-L3, DSIL4-L5 and DSIL5-S1), para-spinal muscles signal intensity (PSM SI left and PSM SI right), subcutaneous fat signal (Fat left, Fat right), psoas signal (Psoas left and Psoas right),

and cerebrospinal fluid (CSF). In future, we will be interested in adding more lumbar spine features to the model in addition to these 24 features. Some of the other notable features that can be extracted from MRIs are Schmorl's nodes, Modic changes, vertebral alignment, osteophytes, ligamentum flavum, and facet joints. Adding these features to the existing model will present a more comprehensive picture of age related changes in the lumbar spine.